Testing the Effects of GraphoGame Against a Computer-Assisted Math Intervention in Primary School

Julie Lassault, Liliane Sprenger-Charolles, Jean-Patrice Albrand, Edouard Alavoine, Ulla Richardson, Heikki Lyytinen, Johannes C Ziegler

Scientific Studies of Reading

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hssr20

Testing the Effects of GraphoGame Against a Computer-Assisted Math Intervention in Primary School

Julie Lassault, Liliane Sprenger-Charolles, Jean-Patrice Albrand, Edouard Alavoine, Ulla Richardson, Heikki Lyytinen & Johannes C. Ziegler

To cite this article: Julie Lassault, Liliane Sprenger-Charolles, Jean-Patrice Albrand, Edouard Alavoine, Ulla Richardson, Heikki Lyytinen & Johannes C. Ziegler (2022): Testing the Effects of GraphoGame Against a Computer-Assisted Math Intervention in Primary School, Scientific Studies of Reading, DOI: 10.1080/10888438.2022.2052884

To link to this article: https://doi.org/10.1080/10888438.2022.2052884

Published online: 16 Mar 2022.

SCIENTIFIC STUDIES OF READING
https://doi.org/10.1080/10888438.2022.2052884

Testing the Effects of GraphoGame Against a Computer-Assisted Math Intervention in Primary School

Julie Lassault[a], Liliane Sprenger-Charolles[a], Jean-Patrice Albrand[a], Edouard Alavoine[a], Ulla Richardson[b], Heikki Lyytinen[b,c], and Johannes C. Ziegler[a]

[a]Laboratoire de Psychologie Cognitive (UMR 7290), Aix-Marseille University, CNRS, Marseille, France; [b]Centre for Applied Language Studies, University of Jyväskylä, Jyväskylä, Finland; [c]UNESCO Chair on Inclusive Literacy Learning for All, University of Jyväskylä, Jyväskylä, Finland

ABSTRACT

Purpose: This study was designed to assess the efficiency of a French version of *GraphoGame* (GG) against an equally engaging math intervention (*Fiete Math*, FM) in a large school sample of children from socioeconomically disadvantaged neighborhoods in grade 1 (N = 921).

Method: The intervention was implemented in two different cohorts who used GG or FM for about four months four times a week for 30 minutes. Gains in reading and mathematics were assessed before and after intervention. Given the nested nature of the data, results were analyzed using hierarchical linear mixed effect models with intervention and initial pretest level as fixed effects and individuals and classes as random effects.

Results: We found positive intervention effects of GG on phoneme awareness (effect size, ES = 0.23), orthographic choice (ES = 0.27) and word reading fluency (ES = 0.18). FM had a significant effect on math achievement (ES = 0.28) but not number comparison. Correlations between intervention gains and game variables (overall accuracy, number of levels played) suggest that the effects of GG were specific.

Conclusions: Positive effects for focused digital reading and math interventions were found in a large school sample of children from socially disadvantaged neighborhoods.

Introduction

The accurate and rapid reading of isolated words is a fundamental cornerstone for the acquisition of literacy skills (Castles et al., 2018; Rayner et al., 2001). International reading assessments (PISA) continue to show that, on average across OECD countries, 23% of the 15-year old students fail to acquire the technical skills to read simple text for comprehension (see OECD, 2019). These students are unable to identify the main idea in a text of moderate length and unable to find information based on explicit criteria. In almost all countries that participated in PISA 2018, students who were socio-economically disadvantaged were less likely to attain the critical level for adequate functioning in modern society (i.e., level 2 in the PISA assessment). The strength of the relationship between a student's socio-economic status and his or her performance varies greatly across countries and was particularly strong in France, Hungary, Luxembourg, Peru, the Philippines, Romania and the Slovak Republic (OECD, 2019). This result is in line with previous epidemiological studies, which showed that French students from socially-disadvantaged neighborhoods were almost 10 times more likely to encounter reading difficulties than students from socially-advantaged neighborhoods (Fluss et al., 2009). In the results of PISA 2018 of France (but also Germany, Hungary, Israel, Peru and the

CONTACT Johannes C. Ziegler Johannes.Ziegler@univ-amu.fr Laboratoire de Psychologie Cognitive, Aix-Marseille University, 3, place Victor Hugo, Marseille 13003, France

Slovak Republic), the gap in reading performance between the 10% most socio-economically advantaged and the 10% most disadvantaged students was over 170 score points – the equivalent of well over four years of schooling.

There is an increasing consensus that many children are at risk for reading difficulties from their first day of school (Loeb & Bassok, 2007) and that early and focused interventions for at-risk children within the school setting is the way to go about reducing them (Castles et al., 2018; Foorman et al., 2003; Foorman & Torgesen, 2001; Hulme et al., 2020; Rayner et al., 2001; Ziegler et al., 2020). The recommendations of OECD (2019) also state that interventions should occur early with a special focus on socio-economically disadvantaged students and/or schools. There are indeed some examples of fairly successful early interventions (e.g., oral language training, phonological awareness, alphabetic knowledge) that produce sustainable effects especially when implementation quality is high (for review, see Hulme et al., 2020).

In the context of early focused intervention for reading within the school context, educational technology, and in particular computer-assisted instruction (CAI), held the promise to support the learning process in unpreceded ways because of its potential to provide individualized and adaptive support (Cheung & Slavin, 2013). However, one of the first studies that systematically evaluated five first-grade reading software products in 43 schools (2600 students) in the US found no significant effect of CAI on reading outcomes (Dynarski et al., 2007). In a follow-up study, the effect was even negative for low-achieving students (Campuzano et al., 2009). In their review of 14 studies, Slavin et al. (2011) found only very small effects (Effect Size, ES = 0.09) of educational technology programs for struggling elementary readers. More recently, a meta-analysis by Cheung and Slavin (2013) on the effects of 20 CAI studies on the reading skills of struggling readers in comparison to "business as usual" methods showed somewhat larger effects (ES = 0.14). Thus, by and large, there is still little evidence that educational technology programs have the expected effect of improving reading skills for at-risk children in primary school. One of the reasons might have to do with the students' motivation. Indeed, in a meta-analysis of 31 studies, Wouters et al. (2013) found that, contrary to common beliefs, serious games were not more motivating than conventional instruction methods.

GraphoGame intervention

One of the reading intervention games, which had not been tested in any of the above-mentioned meta-analyses, is GraphoGame (GG, Richardson & Lyytinen, 2014). GG was initially developed in Finland in the context of dyslexia prevention (Lyytinen et al., 2015), but its use was rapidly extended to promoting reading acquisition in normally-developing readers and assessing children's response to intervention for diagnostic purposes (Lyytinen et al., 2009). Over the past years, GG has been implemented and tested in many languages, such as German (Brem et al., 2010; Huemer et al., 2008), English (Kyle et al., 2013), Portuguese (Carvalhais et al., 2018) or Spanish (Rosas et al., 2017). The core idea of GG is in line with major theories of reading acquisition, which are all based on the idea that the initial steps of reading acquisition are all about learning the mapping between visual symbols (e.g., letters, graphemes) and their equivalent units in spoken language (Castles et al., 2018; Ehri, 1992; Goswami, 2002; Share, 1995; Sprenger-Charolles et al., 2003; Ziegler & Goswami, 2005; Ziegler et al., 2014). According to the phonological decoding and self-teaching theory (Share, 1995), this mapping provides an extremely parsimonious and straightforward way to retrieve the spoken form and therefore the meaning of the thousands of words children have stored in their phonological lexicon prior to reading. Importantly, the mapping needs to be automatized to make fluent reading possible. In a natural reading situation, this happens in a non-supervised way through "self-teaching" (Share, 1995; Ziegler et al., 2020). That is, children start to decode words autonomously when reading short sentences and they create orthographic representations for successfully decoded words that fit the context. Yet, some children struggle to enter this unsupervised self-teaching loop because they might lack sufficient basic decoding skills, vocabulary or reading opportunities (Perry et al., 2019).

This is where GG provides children with opportunities to automatize initial decoding and word recognition skills in a highly systematic and supervised fashion through massive repetitions and explicit error correction.

In line with these theories, the key idea behind GG was to systematically introduce all grapheme-phoneme correspondences (starting with a small and consistent set) and, most importantly, find a way to intensively train and automatize these correspondences (see also, Potier Watkins et al., 2020). This is done in GG by presenting auditory stimuli and have children select or manipulate their orthographic equivalents in various games that are designed to be motivating to the child (for examples, see Figure 1). While the original Finnish version was a basic letter-sound game, which might be sufficient in a highly transparent writing system where literally all words can be correctly decoded on the basis of 20 letter-sound correspondences (see Landerl et al., 2013; Ziegler et al., 2010), it is necessary for the acquisition of less transparent writing systems (such as English and to a lesser extent French) to introduce and train these mappings at various grain sizes including graphemes, syllables, rimes, whole words and even sentences (Kyle et al., 2013; Lassault & Ziegler, 2018).

The first validation study of GG was conducted in Finland with beginning readers who were at risk of reading difficulties (Saine et al., 2010, 2011). The authors initially screened 166 children and then selected 50 of them (the lowest performing 30%) for the intervention study. All children received a regular phonics-based remedial training that was spread out over a 28-week period, with four weekly 45-minute sessions. Each session contained prereading activities, word segmentation, decoding and spelling activities and vocabulary training. For half of the children (N = 25), which were selected randomly, the prereading activities of the regular phonics-based remedial training were replaced by a 15-minute GG session. The results showed that the additional GG intervention produced significant improvements over and above the regular phonics-based remedial training not only in letter knowledge, decoding, and accuracy, but also in fluency and spelling (the effect sizes varied between 0.22 and 2.08).

In contrast to the impressive results of Saine et al. (2010, 2011), a recent meta-analysis by McTigue et al. (2020), which included 15 studies that used the GG intervention in 7 different languages (English, Polish, Spanish, German, Finnish, Portuguese, Kiswahili), came to a somewhat less optimistic

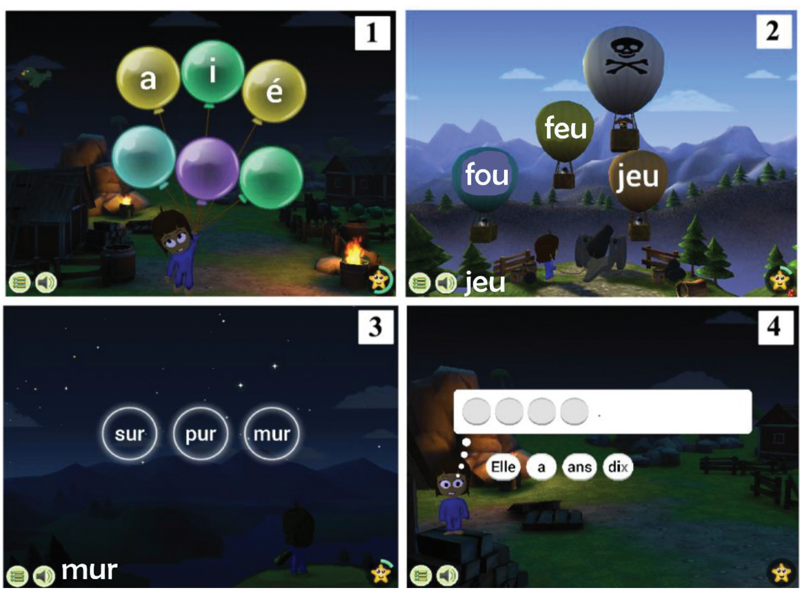

Figure 1. An illustration of four of the levels used in the French version of GraphoGame: phoneme level (1.), Syllable level (2.), Word level (3.), Sentence level (4.).

conclusion. Across the 15 studies (19 comparisons), the authors reported a slightly negative mean effect size (g[1]= −0.02, 95% CI [−0.14, 0.09]) that was not significantly different from zero (p = .70). Note that this meta-analysis focused only on word reading skills as outcome measures and not on reading-related components, such as letter knowledge or phoneme awareness. The authors came to the conclusion that "as a word-reading intervention, there is no evidence that GG produced growth in students' word reading . . . although students often learned from GG, their learning did not typically surpass that of control groups" (McTigue et al., 2020, p. 60). However, the authors found a wide range of effect sizes (g = −1.07 to 1.55), which was taken to suggest that additional, contextual variables might moderate the intervention effects. They investigated the effects of four moderator variables, which were orthographic transparency of the writing system (shallow, moderate, deep), duration of the intervention (short, medium, long), type of control group (untreated control, math game, reading intervention) and the level of adult interaction (low, high). They found that only the level of adult interaction significantly moderated the effects. Studies with a high level of adult interaction showed significant effect sizes (g = .48).

Goal of the present study

The goal of the present study was four-fold. First, we wanted to test a French version of GG that was developed to cover the entire 1st grade reading program based on a theoretically optimal progression (Dehaene, 2011; Sprenger-Charolles, 2017), which took into account the frequency and consistency of grapheme-phoneme correspondences (GPCs) and the frequency of words (Peereman & Sprenger-Charolles, 2018). That is, it was structured in a way that the most consistent and frequent grapheme-phoneme correspondences were introduced progressively and they were then trained in the context of syllables, words and sentences that were made up of previously trained GPCs (Lassault & Ziegler, 2018). Thus, the French version of GG went well beyond teaching basic letter-sound correspondences.[2] Second, we wanted to test the GG intervention in a large school sample of beginning readers (grade 1) from socioeconomically-disadvantaged neighborhoods because the need for early within-school intervention is particularly strong in this part of the population and this is particularly true in France (Fluss et al., 2008, 2009). Third, we wanted to assess the GG intervention against an equally engaging math intervention while keeping teacher and school effects constant. It is well known that teacher effects are the dominant factor affecting student academic gain (Sanders et al., 1997). Thus, the design of the study was quasi-experimental because a first cohort of grade 1 children received the GG intervention in the first year, while a new cohort received the math intervention in the second year with the same teachers in the same schools. Finally, we wanted to investigate whether intervention gains were modulated by the initial level of the child on the various outcome measures at pretest.

Intervention gains in reading were assessed through four reading related outcome measures: phoneme awareness, pseudoword reading fluency, orthographic choice, and word reading fluency. We expected positive effects of GG intervention on phoneme awareness because it is well established that the systematic teaching of grapheme-phoneme correspondences boosts phoneme awareness in a reciprocal way (Perfetti et al., 1987). We expected relatively small effects on pseudoword reading fluency because the French version of GG does not explicitly train pseudoword reading as almost all grapheme-phoneme correspondences are learnt in the context of frequent words. This is a notable departure of the original Finnish version of GG, which was heavily based on training individual letter-sound correspondences. We expected strong effects on orthographic choice, because this task is conceptually very close to the intervention program implemented in GG (i.e., a child heard a word or saw a picture of a word and had to select the correctly spelled word amongst incorrect alteratives, which included pseudohomophones and orthographic neighbors). Finally, we expected moderate effects for word reading fluency (number of words read per minute in lists of words or short passages of text) because GG does not explicitly train word reading fluency (i.e., children never read aloud printed words in GG). Thus, a positive effect on word reading fluency is a rather strong generalization test for the GG intervention.

With respect to the math intervention, we assessed basic mathematical operations (additions and subtractions) with the standardized TEDI math evaluation battery (Van Nieuwenhoven et al., 2001). We expected positive effects on this outcome variable because our math intervention explicitly trained additions and subtractions (see below). We also included a symbolic number comparison tasks as a second outcome measure because previous studies showed that performance on this task was related to math achievement (Holloway & Ansari, 2008; Sasanguie et al., 2013).

Methods

Participants

A total of 921 children participated in the study. They attended grade 1 in nine public schools (37 classes) that were located in socially disadvantaged neighborhoods (priority education areas, PEA) in the city of Marseille, south of France. The schools and classes were selected by the local school authorities on the basis of the following criteria: they had to have at least four classes in that grade, they had to be accessible by public transportation and they should not have been involved in other intervention programs. All participants and their legal guardians gave their informed consent and children gave their assent prior to their inclusion in the study. The present study conforms to recognized standards of the World Medical Association Declaration of Helsinki and was approved by the Institutional Review Board of Aix-Marseille University.

In the first year of our study, 451 children (239 boys and 212 girls) took part in the GG intervention. Their average age was 6;4 years (range 5;10–8;1) at the time of pretests (T1), which was in November (~2 months after the beginning of the school year). In the second year, 470 (244 boys and 226 girls) took part in the math intervention. Their mean age was 6;4 years (range 5;10–8;1) at T1. The children came from the same 9 schools (36 classes). Most of the teachers participated in the study for two years and hence delivered the GG intervention in the first year and the math intervention in the second year. This is possible in France because teachers change classes every year (i.e., after teaching a grade 1 class in one year, a teacher will take on a new grade 1 class the next year).

Interventions

Reading intervention using GG

GG is an audio-visual reading training software based on the simultaneous presentation of auditory stimuli and spelling choices embedded in a series of games (Richardson & Lyytinen, 2014). The children typically hear an auditory input corresponding to a phoneme, syllable word or sentence and, at the same time, several orthographic choices are presented on the screen (for examples, see Figure 1). In most games, children choose the one that matches the spoken input by clicking on one of the displayed response alternatives. The number of alternatives increases adaptively as a function of a child's performance. Errors are indicated immediately, and children have to correct incorrect responses. To make sure that children find the correct answer and learn from the error, only the correct option is shown on the screen immediately after an error occurred. If children made more than 25% errors in a given level, they cannot move on to the next level. However, after five repetitions of the same level, they are allowed to move on to the next level.

The content of the French version of GG was specifically developed for this school-based intervention study on the basis of the theoretically optimal progression initially described by Dehaene (2011) and further developed by Sprenger-Charolles (2017), which took into account the frequency and consistency of grapheme-phoneme correspondences (Peereman & Sprenger-Charolles, 2018) as they occurred in a grade-level lexical database of 54 elementary school French textbooks (MANULEX; Lété et al., 2004),

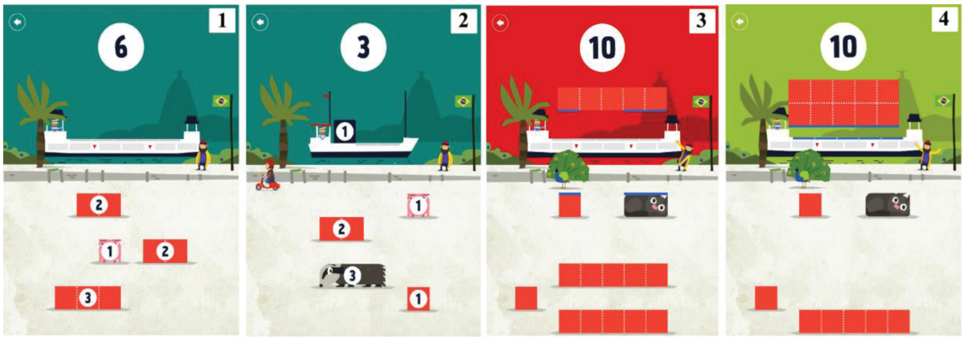

Figure 2. Illustrations of four different levels in Fiete Math: Load an empty ship (1.); Load a partially loaded ship (2.); Feedback when the shipment is not accurate (red) (3.); Feedback when the shipment is correct (green) (4.).

which contained approximately 2,000,000 non-lemmatized forms. The progression is built such that the most frequent and consistent GPCs are introduced first and more complex or context-sensitive correspondences are introduced later. Apart from a small number of highly frequent and potentially inconsistent sight words, all newly introduced words can be decoded on basis of previously learnt GPCs.

The theoretical progression was organized into 67 learning lessons, which were organized around a specific content, such as consistent vowels (a-i-o-u-e-eu-ou) and consistent consonants (j-f-l-r) in sequence 1. Each lesson contained about 10 levels, which trained the specific content of the lesson through various games. Each game lasted about 2–3 minutes. Thus, each lesson would last about 15 to 30 minutes depending on the speed and accuracy of the child. A description of the 67 lessons can be found in the Appendix. Note that the correspondences were rarely trained in isolation but rather embedded in syllables, rimes and frequent words. Some lessons are repetitions of earlier ones in order to consolidate learning. At the beginning and end of each lesson, we added a pretest and posttest level that contained some typical items of each sequence to test the within-game progression for each sequence.

Math intervention

We used a commercial application for training mathematical skills, *Fiete Math* (Ahoiii Entertainment), which was based on the principles of *Number Catcher* (Dotan et al., 2011) and its follow-up called *Number Race* (Wilson et al., 2006). This commercial application was preferred over Number Catcher because it contained many more levels (>1000). Fiete Math explores the basic concepts of number and quantity with simple number blocks that children combine and divide. Levels progress through increasingly more complex number concepts, starting with basic counting and combining and ending with addition and subtraction. The game environment takes the form of a harbor, where kids must load cargo ships with the correct number of blocks to send the ships off to sea (see Figure 2). They are asked to combine or divide blocks to make them represent bigger or smaller numbers. There are six “harbors” in six countries that represent different number manipulations skills. Concepts include simple counting and combining, working with groups of five, and addition and subtraction. In order to keep the children motivated, they can unlock new countries and harbors, win medals and larger ships.

Outcome measures

Phoneme awareness

Two tasks from the EVALEC battery (Sprenger-Charolles et al., 2005) were used to assess phoneme awareness at T1 and T2. In the first task, the examiner pronounced a consonant-vowel-consonant pseudoword and the students had to mentally delete the first phoneme and sound out the pseudoword without the first phoneme (12 items). In the second task, the examiner pronounced a consonant-

consonant-vowel pseudoword and students had to delete the first phoneme (12 items). The final score was the total of correct responses. Asymmetry and kurtosis values are presented in Table 1. Cronbach's alpha was 0.83.

Pseudoword reading fluency

Pseudoword reading fluency was assessed with a 1-minute reading test, in which children had to read a list of pseudowords (max 30) in one minute (Gentaz et al., 2013). The pseudowords were getting increasingly more complex. The score was the number of pseudowords read correctly in one minute. Asymmetry and kurtosis are presented in Table 1.

Orthographic choice

We used the first two tasks of an untimed standardized reading test, the *Timé 2* (Ecalle, 2003). In the first task (12 items), the examiner pronounces a target word and the student had to find the target word among five written alternatives presented on a response sheet (i.e., the correct word, a peudohomophone, a visually close pseudoword, a spelling neighbor, and an unrelated word). In the second task (12 items), the student is presented with a picture and has to identify the corresponding word amongst five written alternatives on a response sheet (same as in task 1). Cronbach's alpha was 0.75.

Word reading fluency

Two tasks were combined to assess timed word reading (i.e., word reading fluency) at T1 and T2. The first task was again a 1-Minute Reading test (Gentaz et al., 2013), in which participants had to read aloud a list of frequent words (max 35). The score was the number of words read correctly in one minute. The second task was the classic standardized French reading test *Alouette* (Lefavrais, 2005), in which children have to read aloud a short text. The text is meaningless such that children cannot use the context to guess words. The score is the number of correctly read words in 3 minutes. The two word reading fluency measures were combined into a single score (correctly read words per minute).

Math achievement (additions and subtractions)

Three tasks were taken from the TEDI math evaluation battery (Van Nieuwenhoven et al., 2001). In the first task (6 items), children had to make additions and subtraction with an image support. For example, the child was presented with a picture of two red and three blue balloons and the experimenter said "there are two red balloons and three blue balloons. How many balloons are there in all?." In the second task, children had to make eight additions or subtractions in the standard format (e.g., 5 + 4 = ?). In the third task, children were presented with incomplete additions or subtractions and had to find the missing operator (5 – ? = 2).

Table 1. Means and psychometric properties of all outcome variables at pretest (pre) and posttest (post).

Variables	Mean	SD	Range	Asymmetry	Kurtosis
Phoneme Awareness (Pre)	10.4	7.2	0–24	0.12	−1.12
Phoneme Awareness (Post)	15.5	6.4	0–25	−0.65	−0.50
Pseudoword Fluency (Pre)	9.2	6.1	0–38	1.16	1.97
Pseudoword Fluency (Post)	28.1	16.0	1–104	0.79	0.75
Orthographic Choice (Pre)	6.5	3.4	0–24	1.08	1.78
Orthographic Choice (Pre)	13.0	4.4	0–24	−0.27	−0.27
Word Fluency (Pre)	10.3	7.5	0–40	1.14	1.37
Word Fluency (Post)	53.3	31.9	1–179	0.60	0.08
Math Achievement (Pre)	11.5	4.6	0–18	−0.55	−0.39
Math Achievement (Post)	14.7	3.4	2–18	−1.29	1.19
Number Comparison (Pre)	1707.9	483.8	556–3500	1.38	2.80
Number Comparison (Post)	1534.5	401.5	676–3500	1.45	3.21

Symbolic number comparison

In the symbolic number comparison task, 36 pairs of one-digit numbers (from 1 to 9) were presented on a tablet screen, one number to the right and one to the left of the central fixation point. The child was asked to indicate as quickly as possible the largest number of the pair by pressing a left or right response key. Each distance (e.g., 2 as in 2–4, 5–7, 3–1, 9–7) was presented four times leading to a total of 36 trials. Response side and order of magnitude was counter-balanced. The critical outcome measure response latency for correct responses.

Procedure

All pretests were administered to the children of the two cohorts (year 1, year 2) in November of Grade 1 (T1). The posttests took place in June of the following year (T2). At T1, each child was tested on the four outcome measures for reading (phoneme awareness, pseudoword reading fluency, orthographic choice and word reading fluency) and two outcome measures for mathematics (math achievement and number comparison). We also obtained other measures (non-verbal IQ, receptive vocabulary, listening and reading comprehension), which are not presented in the present article. The pretests were carried out in two separate sessions and lasted about 1 hour. At T2, the same tests were used again. Both assessments took place in a quiet room in the schools. The children were tested individually, except for the orthographic choice test, which was done in the classroom.

The interventions took place between T1 and T2 (i.e., 5 months). Children were typically trained in groups of six, which in most cases corresponded to half of the class.[3] Each child was attributed an individual Android tablet, which remained at the school. The tablet only contained the GG application (year 1) or the math application (year 2). Prior to the intervention, teachers received a training on the content of GG and Fiete Math and they were trained on how to optimally use the software in the classroom. Teachers signed an agreement to implement the GG and math application approximately 20 minutes per day, 4 times per week, for a total of 16 weeks. Implementation fidelity was ensured through regular school visits from the research team. According to the game log data, the mean duration of the GG intervention was 14.73 h (± 3.51 h).

Statistical analyses

The general analysis strategy was the following: Given that each individual *i* is nested in a class *j*, we used hierarchical linear mixed (HLM) models to take class into account as a random effect using the IBM SPSS MIXED procedure (see Heck, Thomas & Tabata, 2009). For each outcome variable (i.e., phoneme awareness, pseudoword reading fluency, orthographic choice, word reading fluency, math achievement, number comparison), we predicted the posttest scores by the pretest scores using intervention as a fixed factor. This allows one to assess to which extent individual gains in the outcome variable are affected by the intervention. For each outcome variable, we computed three models and used the one that showed the best log-likelihood fit.

In the first model, we simply predicted the posttest outcome score ($POST_{ij}$) as a function of the pretest score (PRE_{ij}) with intervention as a fixed effect (see Equation (1)). β_0 is the intercept; u_{0j} is random error at level 2 (class); e_{ij} is random error at level 1 (individual); all other β are slope coefficients.

$$POST_{ij}=\beta_0+\beta_1 PRE_{ij}+\beta_2 INTERVENTION_j+u_{0j}+\varepsilon_{ij} \quad (1)$$

In the second model, we allowed for the possibility that the slopes vary across classes, which takes into account that some teachers/classes might have used the interventions more efficiently as a function of the child's initial level. Thus, we simply added a random slope to the previous equation (see Equation (2)).

$$POST_{ij} = \beta_0 + \beta_{1j} PRE_{ij} + \beta_2 INTERVENTION_j + u_{0j} + \varepsilon_{ij} \quad (2)$$

In the third model, we allowed for the possibility that the initial level of a child on the pretest outcome variables interacted with the intervention (i.e., the intervention worked more or less well depending on the initial level of the child). Thus, we added an interaction term to the previous equation (see Equation (3)).

$$POST_{ij} = \beta_0 + \beta_{1j} PRE_{ij} * \beta_2 INTERVENTION_j + u_{0j} + \varepsilon_{ij} \quad (3)$$

Given that we used random slopes and interaction terms in Equations (2) and (3), all outcome variables were standardized prior to the analysis (i.e., M = 0; SD = 1). This also has the advantage that it standardizes all slope coefficients, which makes them appropriate effect size measures in mixed effect models (see Lorah, 2018).

Results

We first examined all distributions of the outcome variables for normality. This analysis showed that the kurtosis values for pseudoword and word reading fluency at T1 were too high (7.4 and 15.2, respectively) because of a long tail in the distribution (i.e., a few very good readers). Placing a cutoff at 40 words per minutes normalized the distribution for pseudoword (kurtosis = 1.9, N = 913) and word fluency (kurtosis = 1.4; N = 890). Means and psychometric properties for all outcome variables at pretest and posttest are presented in Table 1.

It should be noted that the Math group had significantly higher scores than the GG group at the first assessment point (T1) on the reading outcome variables (phoneme awareness: 11.2 vs. 9.5; pseudoword reading fluency: 11.0 vs. 7.9; orthographic choice: 6.9 vs. 6.0; word reading fluency: 14.5 vs. 9.6, all *p*s < .001). The same was true for number comparison (1655 vs.1762, $p < .001$), but not for math achievement (11.6 vs. 11.3, $p > .19$). These differences at T1 were most likely due to the fact that the math group was trained one year after the GG group with the same teachers. Because the initial level of a child at T1 was entered as a fixed effect in all equations of the HLM analyses (see above), between-group differences at T1 do not affect the estimates of the intervention effects. The results from the HLM analyses are presented below for each outcome variable.

Phoneme awareness

The best fitting HLM model included random slopes and an interaction term between the effects of intervention and initial level of PA (Equation (3)). All standardized parameter estimates are presented in Table 2. As can be seen in Table 2, intervention was significant ($p < .001$) with larger gains for the GG group than for the math group. The effect size of the intervention on PA was 0.23. The interaction between PA_{pre} and Intervention was significant reflecting the fact that children with initially lower PA scores benefitted to a greater extent from the GG intervention that children with initially better PA scores. The fitted results are plotted in Figure 3(a).

Pseudoword reading fluency

The best fitting HLM model included random slopes and an interaction term between the effects of intervention and initial level of pseudoword reading (Equation (3)). As can be seen in Table 2, Intervention failed to reach significance ($p = .081$; ES = 0.11), but there was a significant interaction between the initial level of pseudoword reading fluency (PWR_{pre}) and Intervention reflecting the fact that the GG intervention was more beneficial for students who had initially better decoding skills than for students who had initially weaker decoding skills. The effects are plotted in Figure 3(b).

Table 2. Standardized parameter estimates for the best-fitting HLM model predicting performance on the dependent variables at posttest as a function of intervention (GG vs. Math), initial level on that variable at pretest (pre), and the interaction between the two (when appropriate). Between classes and between students are random effects. *SE*, Standard Error.

Parameters	estimates	*SE*	*t*	significance
Phonological Awareness (PA)				
Between classes	.030	.014	2.048†	.041
Between students	−.119	.048	−2.482	.016
Intervention	.233	.068	3.412	.001
PA_{pre}	.621	.045	13.926	<.0001
PA_{pre} * Intervention	−.125	.058	−2.173	.030
Pseudoword Reading (PWR)				
Between classes	.023	.012	1.902†	.057
Between students	−.027	.044	−.610	.544
Intervention	.112	.063	1.769	.081
PWR_{pre}	.654	.040	16.433	<.0001
PWR_{pre} * Intervention	.136	.062	2.198	.028
Orthographic Choice (OC)				
Between classes	.012	.012	.964†	.335
Between students	−.132	.045	−2.965	.004
Intervention	.272	.064	4.267	<.0001
OC_{pre}	.478	.029	16.214	<.0001
Word Reading Fluency (WRF)				
Between classes	.025	.012	2.188†	.029
Between students	.000	.043	−.007	.995
Intervention	.178	.061	2.906	.005
WRF_{pre}	1.068	.042	25.456	<.0001
Math achievement (Math)				
Between classes	.041	.015	2.708†	.007
Between students	.138	.049	2.835	.006
Intervention	−.281	.069	−4.059	<.0001
$Math_{pre}$	.634	.034	18.724	<.0001
Number Comparison (NC)				
Between classes	.035	.017	2.136†	.033
Between students	−.016	.042	−.387	.699
Intervention	.019	.060	.310	.757
NC_{pre}	.470	.039	12.119	<.0001

† *z*-value (Wald)

Orthographic choice

The best fitting HLM model did not include random slopes or interactions (Equation (1)). As can be seen in Table 2, the effect of Intervention was highly significant with larger gains for the GG group than the math group ($p < .0001$). The effect size of the main intervention effect was 0.27. The results are presented in Figure 3(c).

Word reading fluency

The best fitting HLM model included random slopes but no interaction term (Equation (2)). As can be seen in Table 2, on the measure of word reading fluency, the main effect of intervention was significant ($p < .05$) with larger gains for the GG group than for the math group. The effect size was 0.18. The results are plotted in Figure 3(d).

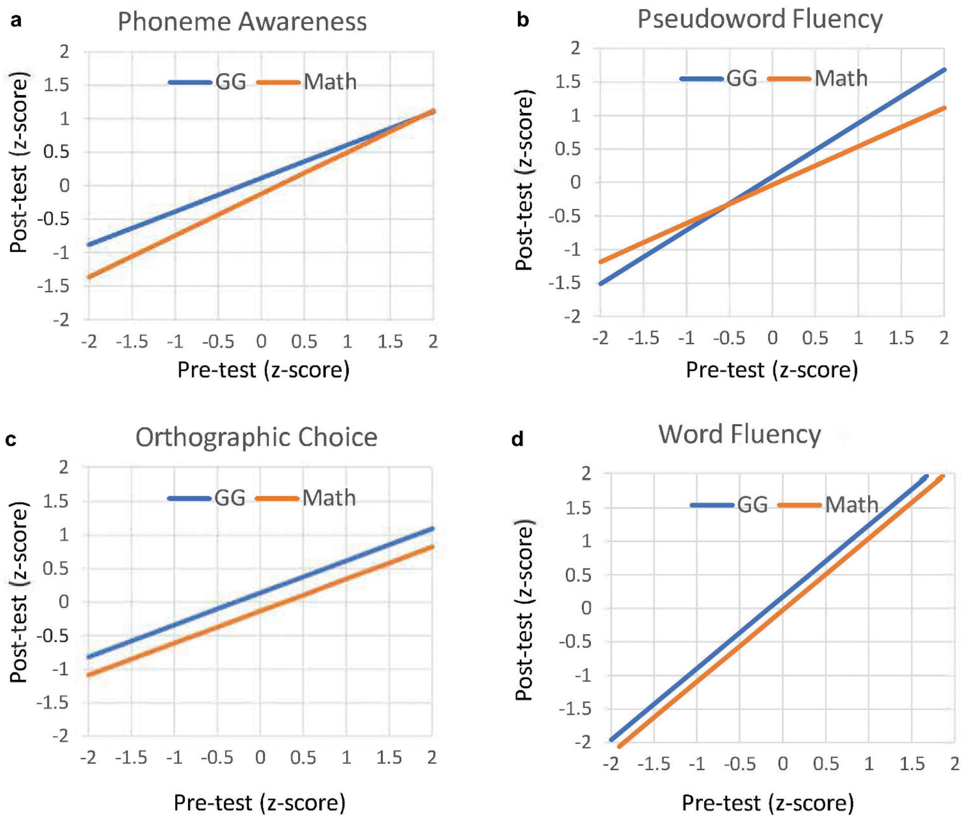

Figure 3. Model-fitted posttest scores on the four reading outcome measures as a function of intervention and pretest level: phoneme awareness (a), pseudoword reading fluency (b), orthographic choice (c) and word reading fluency (d). All variables are z-scored.

Math achievement

The best fitting HLM model included random slopes but no interaction term (Equation (2)). The results are presented in Table 2 and Figure 4(a). The main effect of intervention was significant ($p < .0001$) with larger gains for the Math group than for the GG group. The effect size was 0.28. Although initial math achievement predicted future math achievement, there was no interaction between the effects of intervention and initial math achievement.

Number comparison

The results of the number comparison task are presented in Figure 4(b) and the parameters estimates are presented in Table 2. The best fitting HLM model included random slopes but no interaction term (Equation (2)). The analysis showed no effect of intervention and no interaction between the initial level of number comparison ability and the effects of intervention.

Correlations with GG game variables

One way to address the crucial question as to whether the gains in the four reading outcome measures were specifically related to what happened during GG is to look at the correlations between the reading outcome measures and some of the game variables, such as the number of levels played, the mean game accuracy or the total time played. If the gains in reading were due to the GG intervention, we would expect a positive

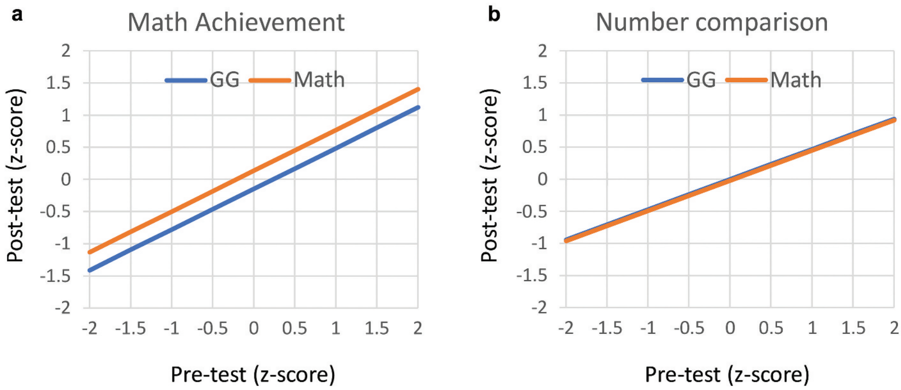

Figure 4. Model-fitted posttest scores on math achievement (a) and number comparison (b) as a function of intervention and pretest level. All variables are z-scored.

Table 3. Correlations between three GG game variables (number of levels played, accuracy [% correct trials] and total time played) and the gains (difference between posttest and pretest scores) in five outcome measures.

Gains	Number of Levels	Accuracy	Total Time
Phoneme Awareness	−.13**	−.04	−.10*
Pseudoword Fluency	.49***	.49***	.08
Orthographic choice	.41***	.46***	.14**
Word Fluency	.59***	.57***	.11*
Math achievement	−.02	.03	−.06

*$p < .05$, ** $p < .001$, *** $p < .0001$

correlation between the game variables and the reading outcome measures. Moreover, if the effects of the GG intervention were specific to reading outcomes, we would predict a correlation only with reading gains but not with gains in math achievement (i.e., gains in math achievement should not depend upon how long or how well a person used GG). The correlations between three GG game variables and the five outcome measures are presented in Table 3. As can be seen in Table 3, for three of the four outcome variables (pseudoword reading fluency, orthographic choice and word reading fluency), we observed fairly strong and positive correlations with the number of levels played and overall game accuracy ($.41 < r < .59$). Interestingly, the total time spent in the game yielded much weaker correlations than the number of levels played or the overall accuracy, which suggests that what matters is game performance or engagement rather than pure exposure. Finally, none of the GG game variables correlated with gains in math achievement, which shows that the effects of GG intervention are specific. To illustrate this point, Figure 5 presents the correlations between number of levels played for gains in word reading fluency versus gains in math achievement.

Discussion

The goal of the present study was four-fold. First, we wanted to test a French version of GG that was developed to cover the entire 1st grade reading program based on a theoretically optimal progression (Dehaene, 2011; Potier Watkins et al., 2020; Sprenger-Charolles, 2017). Second, we wanted to test the GG intervention in a large school sample of beginning readers (1 grade) from socioeconomically-disadvantaged neighborhoods because the need for early within-school intervention is particularly strong for these children (Fluss et al., 2008, 2009). Third, we wanted to assess the GG intervention against an equally engaging math intervention (active control group) while keeping teacher and school effects constant. Finally, we wanted to analyze the results using hierarchical linear mixed models, which consider interindividual differences (i.e., initial pretest level of the child) and class/teacher effects.

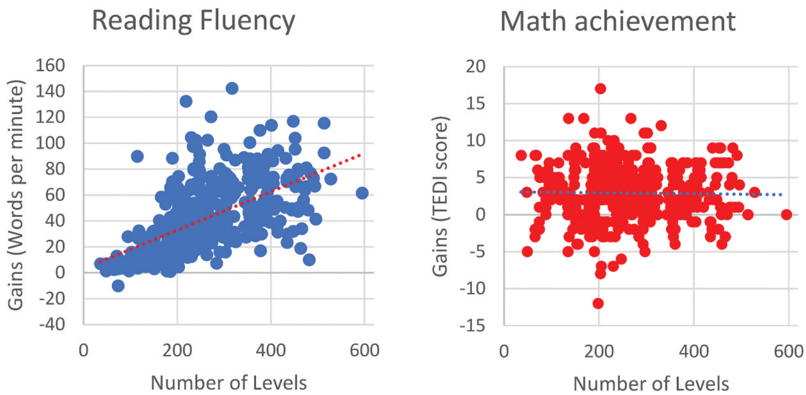

Figure 5. Scatter plots between the number of levels played in GG and gains in word reading fluency and math achievement . Gain scores are the differences between the posttest scores minus the pretest scores.

We tested the effects of GG intervention on four reading outcome variables (phoneme awareness, pseudoword reading fluency, orthographic choice and word reading fluency), for which we expected different effect sizes. By and large, the results confirmed our predictions. First, we found strong and robust effects on phoneme awareness (ES = 0.23), which was expected because the systematic teaching of grapheme-phoneme correspondences boosts phoneme awareness in a reciprocal way (Hulme et al., 2012; Perfetti et al., 1987). Interestingly, we found that children with initially weaker phoneme awareness skills benefitted more from GG intervention than children with initially stronger phoneme awareness skills (see Figure 3(a)). This result suggests that GG intervention is particularly useful for children who fail to master the initial stages of learning to read, such as understanding the alphabetic principle and the mapping between graphemes and phonemes.

Second, we found the strongest and most robust effect of GG intervention in the orthographic choice task (ES = 0.27), which is an untimed word reading task that measures word-specific orthographic knowledge. This task is conceptually the closest to the GG intervention because the child hears a word or sees a picture of a word and has to choose the correct word amongst incorrect alternatives, which included sound-alike pseudowords (pseudohomophones) and orthographically similar neighbors. There was no interaction with initial level of the child, which suggests that GG intervention can help all children to boost word-specific orthographic knowledge.

Third, the effects of GG intervention on timed reading measures (fluency) were weaker for word reading fluency (ES = .18) and failed to reach significance for pseudoword reading fluency (ES = .11). For pseudoword reading fluency, we observed an interaction with the initial level of decoding, which suggested that GG intervention had a beneficial effect on pseudoword reading fluency only for children with initially better decoding skills. How could we interpret the finding that intervention effects were weaker for the two fluency measures than for the untimed reading measure? It could be argued that the effects on untimed word reading materialize before the effects on word reading fluency because basic decoding skills (cipher skills) are a necessary prerequisite for storing word specific knowledge which in time will lead to more fluent reading. In addition, it should be noticed that the reading fluency measures are based on reading aloud, which is never directly trained in GG. Moreover, reading aloud requires the mapping from orthography to phonology (O- > P), whereas GG trains the mappings from phonology to orthography (P -> O). That is, in GG, a child is presented with a spoken phoneme, syllable or word and has to choose between orthographic alternatives. Thus, taken together, it seems logical that GG is more efficient in training orthographic access and the quality of orthographic representations (as measured in the orthographic choice task) rather than word reading fluency.

Finally, how could one explain the relatively small and nonsignificant effects on pseudoword reading fluency (ES = 0.11)? First of all, the French version of GG does not explicitly train pseudoword reading and almost all grapheme-phoneme correspondences are learnt in the context of frequent words. Also note that the decoding measure was also a fluency measure (i.e., pseudoword reading fluency). As explained above, GG does not explicitly train fluency. It is therefore not surprising that children with initially weak decoding skills did not show much improvement in pseudoword reading fluency as opposed to children with initially better decoding skills. This finding is in line with the phonological decoding self-teaching theory according to which a minimal number of grapheme-phoneme mappings must be mastered to get the decoding network of the ground (Ziegler et al., 2014).

One way to investigate whether the gains on the outcome measures are specifically related to the intervention is to calculate the correlation between the gains on the outcome measures and the time that children spent using the game (dosage), the overall performance (success rate) or the number of levels played (progress). As predicted, the correlations between the GG variables and reading outcomes were significant and positive for three of the four outcome variables (pseudoword reading fluency, orthographic choice and word reading fluency). The correlations with phoneme awareness were much smaller and negative. However, one should keep in mind that gain scores (difference between posttest and pretest performance) are linear and do not take into account the initial level of a child. Yet, our results on phoneme awareness showed that the GG intervention was only efficient for children with initially weaker phoneme awareness skills.

Importantly, the within-game variables in GG (i.e., dosage, success rate, progress) correlated with reading outcome variables but not with math achievement, which suggests that the reading gains are specifically related to the GG intervention. Indeed, one might argue that the good and motivated students might not only play GG more seriously but would also become better readers, which would produce a positive correlation without a necessarily causal link between the GG intervention and their reading progress. However, if this explanation were correct, such good and motivated students should have also progressed more strongly in math skills, which was not the case. In general, the positive correlations are encouraging and seem to suggest that one needs to consider game performance (dosage, overall success rate and progress) but also motivation and engagement as critical moderator variables for the success of an intervention (Ronimus et al., 2014; Ronimus & Lyytinen, 2015). A previous study showed that within-game variables are useful to predict intervention outcomes (Thomson et al., 2020). Indeed, it seems almost trivial that we should not expect positive intervention outcomes for children who play little, who are not motivated or for whom the game was too difficult (poor performance, few levels). Yet, motivation, engagement and game performance have not been used as moderator variables to assess the efficiency of an intervention. This point has also been made by McTigue et al. (2020) who stated that "unfortunately, the intervention research in computer games has been dominated by those providing simple exposure to technology but often ignoring teacher and learner variables" (p. 59).

One moderator variable that has shown a strong effect on intervention outcomes is the level of adult involvement. In their meta-analysis, McTigue et al. (2020) coded interventions as either low or high adult interaction. In low-interaction interventions, students worked individually within a large group (often in a computer lab) or at home. In contrast, in high-interaction interventions, GG was played with an adult (1:1 or 1:2 adult /child ratio) or integrated into small-group teacher-led lessons ((two to six students). In the meta-analysis, the level of adult involvement proved to be a statistically significant moderator, with high interactions associated with better word reading (g = 0.47) and low adult interaction associated with a small negative effect (g = −0.07). Our study falls into the category of high-interaction interventions because children played in small groups of a maximum of six students and teachers provided active support.

Although the assessment of the efficiency of the math intervention was not the main goal of the study (i.e., active control group), we nevertheless investigated whether the math intervention produced a significant increase in mathematical achievement over and-above the effects of the reading intervention. Indeed, on the math achievement posttest that evaluated basic operations (additions and subtractions), we obtained significant math intervention effects and the effect size was comparable to that of the GG intervention (i.e., ES = 0.28). We did not find a significant effect on symbolic number comparison. This could be due to the fact that number comparison was not directly trained in the *Fiete* intervention program.

The present study has a number of limitations. The most obvious one is that we did not choose a randomized controlled trial, in which interventions were randomly attributed to classes or schools. We rather used a quasi-experimental design, in which the same teachers administered both interventions successively (GG the first year, Fiete Math the second year). This was done to reduce teacher effects and biases due to social comparisons between teachers who believed to be in the "good" treatment group (i.e., the GG intervention) as opposed to the "bad" control group '(i.e., the math intervention). We have previously seen that teachers who believed to be in the control group (math intervention) increased the intensity of the reading instruction because they felt that their students were disadvantaged and that their own teaching skills were being evaluated (Lassault, 2021). The down-side of our quasi-experimental design was that it turned out that the math group started out at a significantly higher level at T1 than the GG group on all reading measures. However, although we acknowledge the fact that this is not optimal, it does not undermine the positive effects of GG intervention, because, if anything, the differences between the groups at T1 should have weakened the positive effects of GG intervention (i.e., children in the GG group had to first catch up and eventually bypass the math group). In turn, the fact that the math group started higher on all reading measures than the GG group seems to suggest that the teachers might have changed the way they taught reading and possibly became more efficient, which again would reduce the true effect size of the GG intervention.

The second limitation is that we did not test the students one or two years later to see whether the intense GG training durably improved their reading performance. This longitudinal follow-up was initially planned and actually carried out for one of the groups but because of the corona-related lockdown of France in spring 2020, we could not obtain the follow-up measures for the math group. Thus, we could only compare the reading age of the GG-intervention group one year later to the norms on a standardized reading test that was administered at T1, T2 and T2 + 1 year. While the reading age of the children in the GG group was 3 months below their chronological age at T1, their reading age was identical to their chronological age at T2 and this was also the case at the 1-year follow-up (see, Lassault, 2021).

A final limitation is that GG does not train reading aloud, which partially explains the weaker effects on the reading fluency measures. Also, recent data suggest that children with poor phonemic discrimination skills need to use lip reading more than their peers without such difficulties (Piquard-Kipffer et al., 2021). The GG training program does not allow the use of lip reading, which could explain why some of the intervention effects are weaker than expected.

The present work opens some interesting research avenues. The most important one is to use the within-game data and outcome measures of the several hundreds of students to make the training content and the progression more adaptive using unsupervised learning algorithms that maximize the gains as a function of initial skills and progress through the game using within-game variables (Thomson et al., 2020). Other perspectives include the development of games that train reading comprehension (e.g., Javourey-Drevet et al., 2022) and the use of automatic speech recognition software to train reading aloud (O -> P mappings), which are two dimensions that are currently not being trained.

Notes

1. Hedges's g was calculated as the difference in gain (pretest to posttest) between the GG and control groups divided by the pooled standard deviation.
2. The French version of GG is freely available in France and French-speaking territories on google play and apple store. It can be downloaded in all other countries for a small fee.

3. In grade 1 and 2, the size of a class in PEAs in France has been reduced to 12 children per class in 2018.

Acknowledgments

We thank all participants and teachers for their participation and the school authorities for their support. Thanks are extended to Jacques Ginestié, Karine Télouk, and Jean-Louis Leydet who helped at various stages of the project, to Pascal Bressoux and Laurent Lima for their advice on statistical analyses, to Iivo Kapanen and Mika Halttunen for their help with the implementation of GG, and to Jesper Ryynänen and the GraphoGame Group Ldt for their support. We finally thank three anonymous reviewers who provided very helpful suggestions.

Ethics approval statement

All participants and their legal guardians gave their informed consent and children gave their assent prior to their inclusion in the study. The present study conforms to recognized standards of the World Medical Association Declaration of Helsinki and was approved by the Institutional Review Board of Aix-Marseille University.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research has been supported by the French National Agency for Research (ANR) through the GraphoGame project (ANR-13-APPR-0003), the eFRAN Lemon project, the Institute of Convergence ILCB (ANR-16-CONV-0002) and the Excellence Initiative of Aix-Marseille University A*MIDEX (ANR-11-IDEX-0001-02). The work that led to this publication has also been supported by AMPIRIC, a center of excellence on research in education and teacher training.

ORCID

Johannes C. Ziegler http://orcid.org/0000-0002-2061-5729

References

Brem, S., Bach, S., Kucian, K., Guttorm, T. K., Martin, E., Lyytinen, H., . . . Richardson, U. (2010). Brain sensitivity to print emerges when children learn letter-speech sound correspondences. *Proceedings of the National Academy of Sciences*, *107*(17), 7939–7944. https://doi.org/10.1073/pnas.0904402107

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts. NCEE 2009-4041. National Center for Education Evaluation and Regional Assistance.

Carvalhais, L., Richardson, U., & Castro, S. L. (2018). Computer-assisted reading and spelling intervention with Graphogame fluent Portuguese. In Á. Rocha, H. Adeli, L. Reis, & S. Costanzo (Eds.), *Trends and advances in information systems and technologies. WorldCIST'18 2018. Advances in intelligent systems and computing* (Vol. 746). Springer. pp. 1452–1460.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. Educational Research Review, 9(1), 88–113. https://doi.org/https://doi.org/10.1016/j.edurev.2013.01.001

Dehaene, S. (2011). *Apprendre à lire: Des sciences cognitives à la salle de classe*. Odile Jacob.

Dotan, D., Dehaene, S., Huron, C., & Piazza, M. (2011). *The number catcher*. http://www.thenumbercatcher.com/nc/home.php

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., ,andSussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort*. Research Report - Report Number NCEE 2007-4005. Washington, DC: U.S. Department of Education. https://telearn.archives-ouvertes.fr/hal-00190019

Ecalle, J. (2003). *Timé-2: Test d'identification de mots écrits pour enfants de 6 à 8 ans*. Edition du Centre de Psychologie Appliquée.

Ehri, L. C. (1992). Reconceptualizing the development of sight word reading and its relationship to recoding. In P. B. Gough, L. E. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 105–143). LEA.
Fluss, J., Ziegler, J., Ecalle, J., Magnan, A., Warszawski, J., Ducot, B., . . . Billard, C. (2008). Prevalence of reading disabilities in early elementary school: Impact of socioeconomic environment on reading development in 3 different educational zones. *Archives de Pédiatrie*, *15*(6), 1049–1057. https://doi.org/10.1016/j.arcped.2008.02.012
Fluss, J., Ziegler, J. C., Warszawski, J., Ducot, B., Richard, G., & Billard, C. (2009). Poor reading in French elementary school: The interplay of cognitive, behavioral, and socioeconomic factors. *Journal of Developmental and Behavioral Pediatrics*, *30*(3), 206–216. https://doi.org/10.1097/DBP.0b013e3181a7ed6c
Foorman, B. R., Breier, J. I., & Fletcher, J. M. (2003). Interventions aimed at improving reading success: An evidence-based approach. *Developmental Neuropsychology*, *24*(2–3), 613–639. https://doi.org/10.1207/S15326942dn242&3_06
Foorman, B. R., & Torgesen, J. (2001). Critical elements of classroom and small-group instruction promote reading success in all children. *Learning Disabilities Research & Practice*, *16*(4), 203–212. https://doi.org/10.1111/0938-8982.00020
Gentaz, E., Sprenger-Charolles, L., Theurel, A., & Colé, P. (2013). Reading comprehension in a large cohort of French first graders from low socio-economic status families: A 7-month longitudinal study. *PLoS ONE*, *8*(11), e78608. https://doi.org/10.1371/journal.pone.0078608
Goswami, U. (2002). Phonology, reading development and dyslexia: A cross-linguistic perspective. *Annals of Dyslexia*, *52*(1), 141–163. https://doi.org/10.1007/s11881-002-0010-0
Heck, R., Thomas, S., & Tabata, L. (2009). An introduction to multilevel modeling techniques (Quantitative methodology series). New York: Routledge.
Holloway, I. D., & Ansari, D. (2008). Domain-specific and domain-general changes in children's development of number comparison. *Developmental Science*, *11*(5), 644–649. https://doi.org/10.1111/j.1467-7687.2008.00712.x
Huemer, S., Landerl, K., Aro, M., & Lyytinen, H. (2008). Training reading fluency among poor readers of German: Many ways to the goal. *Annals of Dyslexia*, *58*(2), 115–137. https://doi.org/10.1007/s11881-008-0017-2
Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The causal role of phoneme awareness and letter-sound knowledge in learning to read: Combining intervention studies with mediation analyses. *Psychological Science*, *23*(6), 572–577. https://doi.org/10.1177/0956797611435921
Hulme, C., Snowling, M. J., West, G., Lervåg, A., & Melby-Lervåg, M. (2020). Children's language skills can be improved: Lessons from psychological science for educational policy. *Current Directions in Psychological Science*, *29*(4), 372–377. https://doi.org/10.1177/0963721420923684
Javourey-Drevet, L., Dufau, S., François, T., Gala, N., Ginestié, J., & Ziegler, J. (2022). Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of French. *Applied Psycholinguistics*, *43*(2), 485–512. https://doi.org/10.1017/S014271642100062X
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. https://doi.org/10.1002/rrq.038
Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppanen, P. H., Lohvansuu, K., . . . Schulte-Korne, G. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *54*(6), 686–694. https://doi.org/10.1111/jcpp.12029
Lassault, J. (2021). *Le numérique comme outils pédagogique pour l'apprentissage de la lecture et des mathématiques: Du laboratoire à la salle de classe* [Digital tools to improve reading and math: From the laboratory to the classroom]. Aix-Marseille University.
Lassault, J., & Ziegler, J. C. (2018). Les outils numériques d'aide à l'apprentissage de la lecture. *Langue Francaise*, *119*, 111–121. https://doi.org/10.3917/lf.199.0111 .
Lefavrais, J. (2005). *Test de l'Alouette (revised version)*. ECPA.
Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments and Computers*, *36*(1), 156–166. https://doi.org/10.3758/BF03195560
Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. Economics of Education Review, 26(1), 52–66. https://doi.org/https://doi.org/10.1016/j.econedurev.2005.11.005
Lorah, J. (2018). Effect size measures for multilevel models: definition, interpretation, and TIMSS example. Large-scale Assessments in Education, 6(1), 8. https://doi.org/10.1186/s40536-018-0061-2
Lyytinen, H., Erskine, J., Hamalainen, J., Torppa, M., & Ronimus, M. (2015). Dyslexia-early identification and prevention: Highlights from the Jyvaskyla longitudinal study of dyslexia. *Current Developmental Disorders Reports*, *2*(4), 330–338. https://doi.org/10.1007/s40474-015-0067-1
Lyytinen, H., Erskine, J., Kujala, J., Ojanen, E., & Richardson, U. (2009). In search of a science-based application: A learning tool for reading acquisition. *Scandinavian Journal of Psychology*, *50*(6), 668–675. https://doi.org/10.1111/j.1467-9450.2009.00791.x

McTigue, E. M., Solheim, O. J., Zimmer, W. K., & Uppstad, P. H. (2020). Critically reviewing GraphoGame across the world: Recommendations and cautions for research and implementation of computer-assisted instruction for word-reading acquisition. *Reading Research Quarterly*, *55*(1), 45–73. https://doi.org/10.1002/rrq.256

OECD. (2019) . *PISA 2018 results (volume I): What students know and can do.*

Peereman, R., & Sprenger-Charolles, L. (2018). Manulex-MorphO, une base de données sur l'orthographe du français intégrant les morpho-phonogrammes. *Langue Française*, *199*(3), 99–109. https://doi.org/10.3917/lf.199.0099

Perfetti, C. A., Beck, I., Bell, L. C., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal - a longitudinal-study of 1st-grade children. *Merrill-Palmer Quarterly-Journal of Developmental Psychology*, *33*(3), 283–319.

Perry, C., Zorzi, M., & Ziegler, J. C. (2019). Understanding dyslexia through personalized large-scale computational models. *Psychological Science*, *30*(3), 386–395. https://doi.org/10.1177/0956797618823540

Piquard-Kipffer, A., Cavadini, T., Sprenger-Charolles, L., & Gentaz, É. (2021). Impact of lip-reading on speech perception in French-speaking children at risk for reading failure assessed from age 5 to 7. *Année Psychologique/ Topics in Cognitive Psychology*, *121*(2), 3–18. https://doi.org/10.3917/anpsy1.212.0003

Potier Watkins, C., Caporal, J., Merville, C., Kouider, S., & Dehaene, S. (2020). Accelerating reading acquisition and boosting comprehension with a cognitive science-based tablet training. *Journal of Computers in Education*, *7*(2), 183–212. https://doi.org/10.1007/s40692-019-00152-6

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science*, *2*(2 Suppl), 31–74. https://doi.org/10.1111/1529-1006.00004

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*, *10*(1), 39–60. https://doi.org/10.17011/ht/urn.201405281859

Ronimus, M., Kujala, J., Tolvanen, A., & Lyytinen, H. (2014). Children's engagement during digital game-based learning of reading: The effects of time, rewards, and challenge. *Computers & Education*, *71*(2), 237–246. https://doi.org/10.1016/j.compedu.2013.10.008

Ronimus, M., & Lyytinen, H. (2015). Is school a better environment than home for digital game-based learning? The case of graphogame. *Human Technology: An Interdisciplinary Journal on Humans in ICT Environments*, *11*(2), 123–147. https://doi.org/10.17011/ht/urn.201511113637

Rosas, R., Escobar, J.-P., Ramírez, M.-P., Meneses, A., & Guajardo, A. (2017). Impact of a computer-based intervention in Chilean children at risk of manifesting reading difficulties/Impacto de una intervención basada en ordenador en niños chilenos con riesgo de manifestar dificultades lectoras. *Journal for the Study of Education and Development*, *40*(1), 158–188. https://doi.org/10.1080/02103702.2016.1263451

Saine, N. L., Lerkkanen, M.-K., Ahonen, T., Tolvanen, A., & Lyytinen, H. (2010). Predicting word-level reading fluency outcomes in three contrastive groups: Remedial and computer-assisted remedial reading intervention, and mainstream instruction. *Learning and Individual Differences*, *20*(5), 402–414. https://doi.org/10.1016/j.lindif.2010.06.004

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. https://doi.org/10.1111/j.1467-8624.2011.01580.x

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57–67. https://doi.org/10.1023/A:1007999204543

Sasanguie, D., Göbel, S. M., Moll, K., Smets, K., & Reynvoet, B. (2013). Approximate number sense, symbolic number processing, or number–space mappings: What underlies mathematics achievement? *Journal of Experimental Child Psychology*, *114*(3), 418–431. https://doi.org/10.1016/j.jecp.2012.10.012

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

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*(1), 1–26. https://doi.org/10.1016/j.edurev.2010.07.002

Sprenger-Charolles, L. (2017). Une progression pédagogique construite à partir de statistiques sur l'orthographe du français (d'après Manulex-Morpho): Pour les lecteurs débutants et atypiques. *Approche Neuropsychologique des Apprentissages chez l'Enfant (ANAE)*, *148*(5), 247–256.

Sprenger-Charolles, L., Colé, P., Bechennec, D., & Kipffer-Piquard, A. (2005). French normative data on reading and related skills from EVALEC, a new computerized battery of tests. *Revue Européenne de Psychologie Appliquée*, *55*(3), 157–186. https://doi.org/10.1016/j.erap.2004.11.002

Sprenger-Charolles, L., Siegel, L. S., Béchennec, D., & Serniclaes, W. (2003). Development of phonological and orthographic processing in reading aloud, in silent reading and in spelling: A four year longitudinal study. *Journal of Experimental Child Psychology*, *84*(3), 194–217. https://doi.org/10.1016/S0022-0965(03)00024-9

Thomson, J. M., Foldnes, N., Uppstad, P. H., Njå, M., Solheim, O. J., & Lundetræ, K. (2020). Can children's instructional gameplay activity be used as a predictive indicator of reading skills? *Learning and Instruction*, *68*, 101348. https://doi.org/10.1016/j.learninstruc.2020.101348

Van Nieuwenhoven, C., Noël, M.-P., & Grégoire, J. (2001). *Tedi-math: Test diagnostique des compétences de base en mathématiques*. ECPA.
Wilson, A. J., Dehaene, S., Pinel, P., Revkin, S. K., Cohen, L., & Cohen, D. (2006). Principles underlying the design of "the number race," an adaptive computer game for remediation of dyscalculia. *Behavioral and Brain Functions*, *2*(1), 19. https://doi.org/10.1186/1744-9081-2-19
Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, *105*(2), 249–265. https://doi.org/10.1037/a0031311
Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., . . . Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, *21*(4), 551–559. https://doi.org/10.1177/0956797610363406
Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*(1), 3–29. https://doi.org/10.1037/0033-2909.131.1.3
Ziegler, J. C., Perry, C., & Zorzi, M. (2014). Modelling reading development through phonological decoding and self-teaching: Implications for dyslexia. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120397. https://doi.org/10.1098/rstb.2012.0397
Ziegler, J. C., Perry, C., & Zorzi, M. (2020). Learning to read and dyslexia: From theory to intervention through personalized computational models. *Current Directions in Psychological Science*, *29*(3), 293–300. https://doi.org/10.1177/0963721420915873

Appendix

Sequence	Content
1.1	Vowel (a-i-o-u-e-eu-ou) Consonant (j-f-l-r)
1.2	Repetition of sequence 1
1.3	Addition of "s" at the beginning
1.4	Repetition of sequence 1.3
1.5	Words Vowel-Consonant and Consonant-Vowel-Consonant
1.6	Silent letter and frequent words
2.1	Consonants (p-m-d-v)
2.2	Consonants (b-n-t-ch)
2.3	Repetition of sequences 2.1 and 2.2
2.4	Words Consonant-Consonant-Vowel
2.5	Repetition of sequence 2.4
2.6	Assessment of sequence 2
3.1	Visual discriminations (u-n, . . .)
3.2	Phonemic discrimination (t-d, . . .)
3.3	Mixed discrimination (p-b, . . .)
4.1	Nasal vowels (an-on)
4.2	Nasal vowels (in, un)
5.1	Summary of Module 1 – (Part 1)
5.2	Summary of Module 1 – (Part 2)
6.1	Vowel (oi) and nasal vowel (oin)
6.2	Vowel (ui)
6.3	Vowel (ai)
6.4	Vowel (ei)
6.5	Vowels (au-eau)
6.6	Silent consonant (h)
6.7	Vowels (e-è-ê)
7.1	Visual discriminations (u-n, . . .)
7.2	Phonemic discrimination (t-d, . . .)
7.3	Mixed discrimination (p-b, . . .)
8	Consonants (qu-k)
9.1	Contextual consonant (c – sound:/k/)
9.2	Contextual consonant (c – sound:/s/)
10.1	Consonant (z)
10.2	Contextual consonant (s – sound: /z/)
10.3	Contextual consonant (s, ss – sound: /s/)
11	Vowels (es-ez-er-and)
12.1	Contextual consonant (g – sound: /g/)
12.2	Contextual consonant (g – sound: /Z/)

(*Continued*)

Sequence	Content
13	Consonant (gn)
14.1	Nasal vowels (ant-and-ent)
14.2	Nasal vowels (om-im-am-am-em)
14.3	Nasal vowels (ain-ein-aim)
15.1	Vowel (e) in front of Double Consonants (ell-ett-err-eff-ess)
15.2	Vowel (e) in the middle of a word (ec-es-er)
16	Consonant (ph)
17.1	Vowel (i) in front of a Vowel (sound: /j/)
17.2	Vowel (i) in front of one or two l (sound: /j/)
18.1	Vowel (y) between two Consonants (sound: /i/)
18.2	Vowel (y) in front of a Vowel
19	Exceptions (ill – sound /il/)
20.1	Silent letter supporting the derivation
20.2	Final consonant(s), plural mark
21	Consonant (x)
22.1	Consonants (ç, sc)
22.2	Consonant (t – sound: /s/)
23.1	Exceptions: consonant (t) pronounced at the end of a word
23.2	Exceptions: consonants (s and c) pronounced at the end of a word
24	Consonant (ch – sound /k/)
25	Vowel (oeu)
26	Exceptions
27.1	Present verbs from the first group
27.2	Presents verbs from the first group (sentences)
27.3	Present tense of verb “to be”
27.4	Personal pronoun subject and verb “to be”
27.5	Present of the verb “to have”
27.6	Personal pronoun subject and verb “to have”
28	Assessment: gender, number and verb