Dynamic assessment of the effectiveness of digital game-based literacy training in beginning readers: A cluster randomised controlled trial

Toivo Glatz[1,2,3], Wim Tops[1,7,8], Elisabeth Borleffs[1], Ulla Richardson[4], Natasha Maurits[2,5], Annemie Desoete[6] & Ben Maassen[1,2]

[1] Centre for Language and Cognition, Faculty of Arts, University of Groningen, Groningen, Netherlands

[2] University of Groningen, University Medical Center Groningen (UMCG), Behaviour and Cognitive Neuroscience (BCN), Groningen, the Netherlands

[3] Institute of Public Health, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[4] Centre for Applied Language Studies, University of Jyväskylä, Jyväskylä, Finland

[5] Department of Neurology, University Medical Center Groningen, Groningen, Netherlands

[6] Department of Developmental Disorders, Ghent University, Ghent, Belgium

[7] School of Educational Studies, Hasselt University, Hasselt, Belgium

[8] Rehabilitation Research Centre (REVAL), Hasselt University, Hasselt, Belgium

Corresponding author: Toivo Glatz[1,2,3]

Email address: toivo.glatz@gmail.com

1. Abstract

In this paper, we report on a study evaluating the effectiveness of a digital game-based learning (DGBL) tool for beginning readers of Dutch, employing active (math game) and passive (no game) control conditions. This classroom-level randomised controlled trial included 247 first graders from 16 classrooms in the Netherlands and the Dutch-speaking part of Belgium. The intervention consisted of 10 to 15 minutes of daily playing during school time for a period of up to 7 weeks. Our outcome measures included reading fluency, phonological skills, as well as purpose built in-game proficiency levels to measure written lexical decision and letter-sound association. After an average of 28 playing sessions, the literacy game improved letter knowledge at a scale generalizable for all children in the classroom compared to the two control conditions. In addition to a small classroom-wide benefit in terms of reading fluency, we furthermore discovered that children who played extensively and scored high on phonological awareness prior to training were more fluent readers than could be expected. This study is among the first to exploit game generated data for the evaluation of DGBL for literacy interventions.

2. Introduction

Adequate early literacy instruction and well-developed literacy skills are indispensable for a child's academic success and future career. It is therefore important to know how we can improve teaching methods and accurately monitor reading progress. Digital game-based learning shows potential in that it has a range of benefits over traditional (offline) teaching methods as it offers a multimodal learning environment to improve the engagement and learning of students (Potocki, Ecalle, & Magnan, 2013). Such games also provide immediate feedback for improved learning, can adapt to individual learners depending on their responses, and are highly motivating for the players (e.g., Desoete, Praet, Van de Velde, De Craene, & Hantson, 2016). In addition, they allow researchers to monitor the players' individual development of accuracy and response times over time in task-relevant contexts and to acquire valuable longitudinal playing data of large groups of participants (e.g., Praet & Desoete, 2014; Puolakanaho & Latvala, 2017). All this makes digital game-based learning a natural choice when investigating potential early recognition and remediation of reading difficulties in children.

2.1. Reading impairment

Developmental dyslexia or specific learning disorder in reading, henceforth dyslexia, is a developmental disorder characterised by persistent difficulties in word recognition (reading) and/or spelling, according to the DSM-5 (American Psychiatric Association, 2013). These difficulties are not caused by a general cognitive delay or by a hearing or vision impairment. Depending on a narrow or wider definition of poor reading proficiency, this developmental disorder affects around 4 to 12% of children across languages (e.g., Schulte-Körne, Deimel, Bartling, & Remschmidt, 1998; Schumacher, Hoffmann, Schmal, Schulte-Korne, & Nothen, 2007). Language and orthography both play an important role in reading (Borleffs, Maassen, Lyytinen, & Zwarts, 2017), with the prevalence of dyslexia differing across languages depending on their characteristics (Bergmann & Wimmer, 2008; Ziegler & Goswami, 2005). Because of differences in the mapping of grapheme-phoneme correspondences, the developmental trajectory and nature of the reading problems may also differ between languages with regular and less regular orthographies (Seymour, Aro, & Erskine, 2003; Bergmann & Wimmer, 2008; Ziegler & Goswami, 2005; Vaessen et al., 2010).

Dyslexia has been shown to be a disorder with a multifactorial aetiology in that it is associated with a range of genetic, environmental, and cognitive risk factors rather than with a single cause (Pennington, 2006). If one parent or sibling has dyslexia, the incidence rate rises to around 45%, indicating a familial risk (for a review, see Snowling & Melby-Lervag, 2016). However, genetic risks do not operate in isolation. Reading is also influenced by environmental factors such as those related to parental socioeconomic status and their interaction with genetic factors (Mascheretti et al., 2013). Moreover, reading problems often seem associated with below average performance on specific cognitive and behavioural factors. The most prominent ones are letter knowledge, phonological awareness, and rapid automatised naming (of letters, digits, and familiar objects). Early performance on these skills has been found to predict both reading accuracy and fluency (van der Leij, Bergen, Zuijen, Jong, Maurits, & Maassen, 2013; Lyon, Shaywitz, & Shaywitz, 2003; Lyytinen et al., 2004; Lyytinen, Erskine, Kujala, Ojanen, & Richardson, 2009; Moll et al., 2014). However, certain cross-linguistic variability exists with respect to the relative weight of each of the cognitive and behavioural predictor of reading acquisition (Landerl et al., 2013; Ziegler et al., 2010). Letter knowledge is most predictive in Finnish with its extreme letter-sound consistency (Lyytinen et al., 2009), rapid automatised naming is the best long-term predictor in German (Brem et al., 2013), and letter knowledge, rapid automatised naming and phonological awareness are important indicators in Dutch (van Bergen, Jong, Plakas, Maassen, & van der Leij, 2012).

For letter knowledge, we can further distinguish between letter-name knowledge and letter-sound knowledge, which in English orthography rarely coincide (e.g., ‘a’ in ‘bark’, ‘w’ in ‘wrong’). The assessment of letter-sound knowledge is the more suitable predictor of word reading in young children at the (very) beginning of reading instruction. However, letter-sound knowledge quickly reaches ceiling and therefore has limited use for long-term predictions. By adding time pressure to the letter-sound knowledge task, a more challenging task yielding a more sensitive measure can be created. Instead of measuring only the availability of letter-sound associations, timed letter-sound knowledge measures the fluency by which letter-sound associations can be retrieved from long term memory which is a proxy for the more general reading-related multimodal audio-visual information processing skill (Blomert, 2011; Hahn, Foxe, & Molholm, 2014).

Given the multifactorial aetiology of dyslexia and early predictors such as poor letter-sound knowledge, phonological awareness and rapid automatised naming, the question arises whether timely training of these skills might help remediate or even prevent reading difficulties. Rapid automatised naming seems more an individual characteristic than a trainable skill (Brem et al., 2013; Landerl et al., 2013, Landerl & Wimmer, 2008; de Jong & Vrielink, 2004; Wolff, 2014). In a position paper, van der

Leij (2013) summarised that an early intervention targeting reading precursors gives a head start but that training effects do not transfer to reading at the end of first grade. Therefore, effective interventions should not only start early but also be adapted to long-lasting educational needs.

2.2. The GraphoGame framework

During the past decade, it has been shown that one promising way to deliver such an extended training is by computerised gaming (Chambers et al., 2008; Richardson and Lyytinen, 2014). GraphoGame is such an adaptive computerised game targeting the training of reading-related skills, originally designed for Finnish with its very transparent orthography (Lyytinen, Erskine, Kujala, Ojanen, & Richardson, 2009). In transparent languages, letter-sound correspondences, and the speed by which these correspondences can be processed, were found to be the most consistent predictors for proficient reading (Lyytinen et al., 2009). The first version of the game therefore aimed to boost grapheme-phoneme correspondence knowledge in beginning readers by establishing solid and reliable connections between graphemes and phonemes (for a review, see Richardson & Lyytinen, 2014). Considering the empirical evidence for long-term development of word reading fluency, GraphoGame proved to work well for Finnish (Saine, Lerkkanen, Ahonen, Tolvanen, & Lyytinen, 2010; 2011). To make GraphoGame usable and effective in other, less transparent languages, more recent versions of the game do not only train grapheme-phoneme correspondences and phonological awareness, but also syllable and word reading fluency and spelling (Richardson & Lyytinen, 2014). So far, experimental studies using GraphoGame have been conducted in over 20 different languages, using different methodology and yielding mixed results.

A recent meta-analysis revealed that although the average gain in word reading fluency across 19 GraphoGame studies was close to zero, a few of the larger studies did show positive effects on reading fluency especially for at-risk readers, and many studies showed benefits for reading-related skills (McTigue, Solheim, Zimmer, Uppstad, 2019). Interpretation is complicated due to large differences between studies in various methodological aspects, including selection criteria for participants (poor performers below a certain cut-off score, teacher recommendation, genetic risk for reading impairment, entire classrooms), age of participants (5 to 10 years), the number and types of control groups (none, active and/or passive controls), sample sizes per group (N = 10 to 185), time on task (1 to 8 hours), training implementation (during school or at home, with or without adult engagement), training period (1 to 28 weeks), and type of language (ranging from transparent orthographies like Finnish to opaque orthographies like English).

Regarding reading fluency, for example, in a study by Saine et al. (2010) first graders at cognitive risk playing GraphoGame in Finnish caught up with children not at cognitive risk with respect to reading fluency in second grade. For reading-related skills, increased performance was observed in the domains of letter-sound knowledge in Dutch, Finnish, second language learners of English and French (Blomert & Willems, 2010; Lovio et al., 2012; Patel et al., 2018; Ruiz et al., 2017), phonological processing in Finnish (Lovio et al., 2012) and first and second language learners of English (Patel et al., 2018; Kyle et al., 2013), and sublexical skills in the form of syllable reading in versions made for German in Austria (Huemer, Landerl, Aro, & Lyytinen, 2008) and Finnish (Heikkilä et al., 2013). This confirms the promising traits of the GraphoGame framework, but the presence of large methodological variations makes it challenging to grasp the characteristics that make up a successful GraphoGame intervention.

2.3. The current study

The aim of the current study is to evaluate the effectiveness of a newly created version of GraphoGame for the semi-transparent Dutch orthography as compared to the earlier results described for a highly transparent language like Finnish (e.g., Saine et al., 2010; 2011) and more opaque languages like French (Ruiz et al., 2017) and English (Kyle et al., 2013; Worth et al., 2018). Our secondary aims are to investigate whether characteristics of participants and the intervention itself modulate the response to GraphoGame-NL intervention and what impact different forms of assessment have in the evaluation of intervention effects.

1. Our first and main research question is as follows: Do children playing GraphoGame-NL for up to seven weeks at the beginning of first grade show a larger response to intervention in word reading fluency, phonological awareness and/or letter-sound knowledge compared to children playing a control game and children not playing at all, while all groups follow the conventional classroom curriculum? As the participating children are mostly pre-readers at the onset of literacy, we hypothesise that GraphoGame-NL improves reading-related skills like letter-sound knowledge and phonological awareness to a bigger extent than word reading accuracy and speed itself.
2. With respect to participants characteristics, we ask the following questions: Are there certain subgroups of children who benefit more from GraphoGame-NL exposure than others? Apart from familial or cognitive risk for dyslexia, previous studies have not investigated effects relating to participant characteristics yet. We expect that children with a particular risk factor, such as familial risk for dyslexia, a young age (compared to the other participants), speaking a foreign language at home, or children who perform below average at pre-test on non-verbal intelligence and/or specific reading related cognitive skills, benefit more of the GraphoGame-NL than children without any such risk factor.
3. Regarding intervention characteristics, we ask: Are in-game metrics acquired from the training phase (played sessions and hours, highest game level that was reached, etc.) relevant predictors for the response to GraphoGame-NL intervention? Previous studies show that for robust training effects, GraphoGame should be played intensively and for a long enough period of time, depending on age group and training goals (Richardson & Lyytinen, 2014). We therefore hypothesise that there is a positive relationship between exposure as measured by in-game metrics and response to intervention, and that best intervention effects are achieved by children who strictly adhere to playing GraphoGame-NL fifteen minutes per school day for a period of seven weeks.
4. Does an assessment of children's literacy skills by means of a dynamic assessment that is fully integrated into the game, allow us to identify the response to intervention more reliably than

traditional pen-and-paper tests for word reading fluency or letter-sound knowledge? The hypothesis is that game-based assessment levels that provide online measures for response times and accuracy at the item level are more sensitive to change than traditional pen-and-paper tests as they can also capture automatisation of literacy skills.

3. Materials & Methods

For the methods and results sections, we follow the CONSORT guidelines for reporting of randomised controlled trials (Schulz, Altman, & Moher, 2010).

3.1. Trial Design

This study employed a multicentre cluster-randomised controlled superiority trial (16 clusters across eight sites).

3.2. Participants

Mainstream primary schools in the northern region (Groningen area) of the Netherlands and the western region (Ghent area) of the Dutch-speaking part of Belgium were contacted by phone or letter and invited to join the study. Initial requirements for participation were (a) schools had enough computers with headphones for students to play the GraphoGame-NL intervention on a daily basis, (b) classroom teachers allowed students to play the game for 10 to 15 minutes per day for at least seven weeks, (c) schools allowed trained clinicians to administer behavioural tests on site at school before and after the intervention during regular school hours, and (d) teachers agreed to adhere to their allocated gaming condition. Eight schools were willing and eligible to participate (three in the Netherlands, five in Belgium) with 16 classrooms (four in the Netherlands, 12 in Belgium). All 312 children attending these classrooms were eligible to participate in the study, of which 107 lived and went to school in the Netherlands, and 205 in Belgium. Subsequently, all parents of children from the selected classrooms were informed about the study and asked for their written informed consent for the gaming and additional behavioural tests. Children were asked for oral assent prior to assessment.

To enable as many children as possible to play the game, there were no initial eligibility or exclusion criteria, and parents were also given the option to consent to participation without additional behavioural assessments. Parents (or caregivers) of participating children were asked to complete a questionnaire about their child's handedness, language(s)/dialect(s) spoken at home, family history of reading problems, neurological problems, and medication.

3.3. Interventions

Our research group created a Dutch version of GraphoGame (GraphoGame-NL), specifically for the present study. Within the existing computerised gaming framework of GraphoGame, which was developed at the University of Jyväskylä (Richardson & Lyytinen, 2014), we added reading content (from letters to simple words), selected from Veilig leren lezen (VLL; 'Learning to read safely'; Mommers, Verhoeven, & van der Linden, 1990) a widely used literacy teaching method in both the Netherlands and the Dutch speaking part of Belgium, and a vocabulary achievement list for six-year-olds (Schaerlaekens, Kohnstamm, Lajaegere, & Vries, 1999). GraphoGame-NL included 650 items, ranging from simple and complex graphemes (e.g., ⟨n⟩, ⟨r⟩, ⟨ui⟩), over CV/VC syllables either representing separate words or occurring as parts of existing words (e.g., *vi* / *is*), to monosyllabic words with CVC structure (e.g., *vis*, 'fish') or targets with CCVC, CVCC or CCVCC consonant clusters (e.g., *prijs*, 'price'; *zwart*, 'black'). For a detailed description of the tasks and materials used within the game, the reader is referred to Appendix 1. We excluded a few infrequent complex graphemes (⟨ch⟩, ⟨sch⟩, ⟨aai⟩, ⟨auw⟩, ⟨eeuw⟩, ⟨ieuw⟩, ⟨oei⟩, ⟨ouw⟩) that are not typically taught at the beginning of the first grade. We also created a limited number of phonotactically legal pseudowords as minimal pairs using a pseudoword creator (Wuggy; Keuleers & Brysbaert, 2010).

Five female students of linguistics and speech-language pathology at the University of Groningen spoke the auditory stimuli. Native speakers of Dutch subsequently evaluated all items with respect to their prototypicality and only the most prototypical items were then included in the game, yielding one to four different spoken realisations per target. While there are some systematic differences in pronunciation between the Dutch spoken in the Netherlands and the northern part of Belgium (for phonetic distances between Dutch dialects, see Nerbonne, Heeringa, Van den Hout, Van der Kooi, Otten, & Van de Vis, 1996), this should not be a big problem as they are all exposed to standard Dutch through multimedia (movies, series, games, etc.).

A mathematics game specifically designed for this research was used as an active control condition. Its framework was identical to that of the reading game, featuring a range of similar reactive/interactive mini-games with varying graphics and task demands with the levels containing number/digit knowledge, counting, comparison of numbers and amounts, sorting of adjacent or nonadjacent numbers in ascending or descending order, as well as simple addition and subtraction. The range of numbers within the training goes from zero up to 20, thus mirroring the classroom content of the first half of first grade. Despite the reading and math interventions being based on the same gaming framework, there are inherent differences in cognitive load for these tasks. The math game generally features fewer distractors and shorter levels compared to the reading game, meaning that at identical exposure in terms of played sessions and hours, children playing the math game will see more levels, give more responses, but see fewer distractors on screen.

A third group of children formed a passive control group by following the conventional classroom curriculum without any additional computerised training. Notably, this group did take part in the in-game assessment sessions which the active intervention groups played in their first and last gaming session of the intervention (see details on assessments below).

The children in the two experimental groups played the respective games (reading or math) for 10 to 15 minutes every day during school hours, resulting in five to ten minutes of effective playing time on task. Children played individually on a computer or laptop wearing headphones. The supervision of the training sessions was carried out by the teachers and differed among schools depending on the numbers of computers, the curriculum, and other local circumstances. At some schools, all the children in a classroom played the respective games at the same time in a computer room, whereas at other locations children took turns using five to ten computers during classes. To ensure that the children understood the tasks and enjoyed playing the games, the teachers and student assistants at least once a week asked them about their progress and encouraged them to give the game another try when the content became more difficult.

3.4. Outcomes

Pre-tests (T1) commenced in September 2015, three to six weeks after the start of the new school term, followed by a seven-week playing phase in October and November 2015. Post-testing (T2) being conducted in November and December 2015. The behavioural assessments at both time points took place on site at schools during school hours. Testing took up to one hour per session and was administered by undergraduate and graduate students in speech-language pathology or linguistics under the supervision of trained clinicians. Whilst the intervention groups completed the game-based assessments during the first and last playing session, the passive controls did so at some point during September or October (T1) and November or December (T2). The following outcomes were evaluated in a response to intervention paradigm with respect to the intervention efficacy of GraphoGame-NL from baseline to eight weeks follow up: reading fluency, phonological awareness, and letter-sound association.

3.4.1. Reading fluency

In the pen-and-paper word reading fluency assessment at T2 students read out two custom lists of 45 words with a time limit of one minute per list. List A contained potentially familiar or trained items (words that occurred in the game), and list B contained untrained items (words that did not occur in the game nor in any other assessment). Words for both lists were selected from a vocabulary achievement list of six-year-olds (Schaerlaekens et al., 1999) and consisted of monosyllabic words ranging from two to five letters (mean and median length of 3.5 and four letters respectively) with a frequency range of 0.3 to 36608 per million (mean and median frequency of 1612 and 51 per million, respectively). Based on results of 272 children, both lists correlated strongly at $r = 0.93$, split-half reliability with Spearman-Brown correction was also very high at 0.96, as was Cronbach's α at 0.96. Because children's performance between the lists of potentially trained and untrained items did not statistically differ in preliminary analyses, we took the average of both lists per child and *z*-transformed the result.

3.4.2. Written lexical decision

As an in-game reading task, we implemented a written lexical decision assessment at T2 where children saw a word or pseudoword on screen and had to either accept it as a real word or reject it as a pseudoword by clicking on a green checkbox or a red cross. This task contained 16 words and 16 pseudowords and was split~~-up~~ into two levels of 16 items, each with a three-minute time limit. For data analyses, we used single-~~-~~trial measures in that we considered accuracy and response times for each word and pseudoword. ~~Similar to~~Like the reading fluency task, monosyllabic words with two to four characters (mean and median length 3.1 and three letters respectively) and a frequency range of four to 24266 per million (mean and median frequencies of 2546 and 124 per million, respectively) were used. The pseudowords were created based on those 16 words using the psycholinguistic pseudoword generating software Wuggy (Keuleers, & Brysbaert, 2010), in most cases, with an edit distance of one grapheme (e.g., *jas-jal* or *tijd-toed*). The pseudowords were therefore balanced in length and featured a high neighbourhood density of real words at an edit distance of one grapheme, ranging from 8 to 36 neighbours (with a mean and median of 21 neighbours). Based on results of 199 children, internal consistency was questionable as indicated by Cronbach α (0.7 and 0.68) as well as split-half reliability with Spearman-Brown correction (0.7 and 0.65) for level and item analyses, respectively.

3.4.3. Phonological awareness

Phonological awareness was assessed with two different pen-and-paper tests at T1 and T2. First, the phonological awareness subtest of the CELF-4-NL (Kort, Schittekatte, & Compaan, 2008) test battery was administered, including blending phonemes into words, identification of final and middle phonemes in words, sentence segmentation (by clapping words), final syllable deletion, word segmentation (by clapping syllables), syllable deletion of bi- and trisyllabic words, and initial phoneme substitution. Reliability of this test, as measured by stratified α, is very high at 0.91 (van den Bos & Lutje Spelberg, 2010), as is internal consistency measured by Cronbach's α at 0.94 (D'hondt et al., 2008). For analyses, we used both z-transformed raw scores ~~as well as~~and norm scores, acting ~~both~~ as both dependent and independent variables. Second, the Proef Fonologisch Bewustzijn ('Test Phonological Awareness', Elen, 2006) was presented, including rhyming, word segmentation (with the number of syllables being indicated by clapping), blending of phonemes, syllables, or lexemes into a word, and pseudoword repetition. No reliability measures are provided for this test by the author.

3.4.4. Timed letter-sound identification

Timed letter-sound identification was implemented into the game and assessed at both T1 and T2. Children heard a phoneme and had to select the corresponding grapheme with a computer mouse on the screen as fast as they could. Simple and complex graphemes were presented one by one with five to ten distractors per trial. We tested 32 different graphemes in 42 trials distributed across four levels, each with a time limit of one minute. The time limit meant that only the fastest children saw all 42 targets, while slower children were only able to see a fraction of that. Based on results of 270 children, internal consistency was high as indicated by Cronbach *α* (0.87 and 0.93) ~~as well as~~and split-half reliability with Spearman-Brown correction (0.87 and 0.91) for level and item analyses, respectively. For data analyses, we considered both, single-~~-~~trial accuracy (binary correct/incorrect) and response times as dependent variables, as well as the absolute number of correctly named letters within four minutes (letter knowledge) as a covariate. Due to their highly skewed nature, response times were always box-cox transformed (Sakia, 1992). For an in-depth description of the in-game assessments, see Appendix 1.

3.5. Additional covariates

To investigate sample characteristics, the following independent variables were available from parental questionnaires: age in years, binary sex (female, male), handedness (left, right, mixed), familial risk for dyslexia (yes, no), and language spoken at home (Dutch or a foreign language). To evaluate the influence of exposure to the game, we extracted six covariates related to the individual progress children made within the game: number of sessions played, hours played, levels played, number of items seen on the screen, given responses, and maximum level achieved at the end of the intervention. For the analysis of in-game assessments, we also extracted properties of the gameplay that are not relevant for our research questions (e.g., sequential trial number to adjust for autocorrelation of observations).

While there should be no unmeasured confounding in a randomised trial when randomisation is successful, adding prognostic covariates can increase power and yield more precise estimates (Kahan, Jairath, Doré, & Morris, 2014). We therefore also intended to measure and adjust for abstract reasoning and rapid automatised naming in our analyses, which were not part of our research questions but are known predictors for test performance in phonological awareness (van den Bos, 1998) and reading fluency (Moll et al., 2014; Vaessen et al., 2010). Abstract reasoning was measured with the analogies and categories subtests of the SON-R 6-40 (Tellegen & Laros, 2014) as an estimate of nonverbal fluid intelligence. Reliability of this test is generally high ranging from 0.87 to 0.95. Because norm scores are only available for children aged six and older and we had a substantial number of children under the age of six in our sample, the raw scores of both subtests were averaged and *z*-transformed. Rapid automatised naming was assessed for objects and colours (van den Bos, 2003). The test requires participants to name out loud 50 depicted objects and colours in five rows of 10 items as accurately and as quickly as possible. We noted the time (in seconds) it took to name the entire list of 50 items. The reliability of these subtests as indicated by stratified α is in the range of 0.89 to 0.91 (van den Bos & Lutje Spelberg, 2010).

3.6. Sample size

To detect medium sized group differences and/or intervention effects (Cohen's $d = 0.5$) with a power of 80% and an alpha level of 0.05 in a two-sided test, each group should contain at least 63 participants. Prior GraphoGame studies rarely exceeded group sizes of 30-50 participants. Our aim was to include 100 participants per group.

3.7. Randomisation

Since it was deemed too difficult and logistically challenging for teachers to ensure that every child played according to an individually randomised gaming condition, we used clustered randomisation to assign the gaming condition by classroom. Therefore, 16 clusters, each containing 10 to 33 children, were semi-randomly assigned to either play the reading version of GraphoGame-NL, a math version of GraphoGame (active controls), or attend the normal school curriculum (passive controls). Where possible, a within-school design was set up: three schools participated with three classrooms, so each of the three gaming conditions was randomly assigned to one of the three classrooms. Another two schools joined with two classrooms, where the reading and math game were randomly assigned to each classroom. The final three schools joined with one classroom and each of the three gaming conditions was again assigned to one of the classrooms. See Table 1 for an overview of the number of eligible children within each cluster and their assigned gaming conditions.

*** Table 1 near here ***

3.8. Blinding

Children and teachers had to be aware of the gaming condition they were assigned to and could not be blinded in our design. During the statistical analysis, those assessing the outcomes were also not blinded to group allocation.

3.9. Statistical methods

Statistical analyses were ~~carried out~~conducted in R (Version 4.1, R Core Team, 2021). Differences in baseline measures between the gaming conditions (reading, math, passive control) and countries

(Netherlands, Belgium) were tested using two-way ANOVAs. Significant main effects or interactions were then followed up with *t*-tests or Tukey HSD tests. The evaluation of intervention effects was conducted with linear mixed effects regression (Bates, Maechler, Bolker, & Walker, 2015). One mixed regression model was fit for each of the seven outcome variables at T2: word reading fluency, written lexical decision (accuracy and response time), phonological awareness (CELF and Proef), and timed letter-sound identification (accuracy and response time). To facilitate interpretation, the outcomes were centred and *z*-transformed where possible, so that the model coefficient β is identical to the effect size Cohen's *d*.

Due to the large number of potential covariates and the explorative nature of the research questions, it was not feasible to set up hypothesis driven models as these do not converge with the given sample size. We therefore opted to use a data driven approach and identify the best model based on Akaike's Information Criterion (AIC; Akaike, 1974). We tested the stepwise forward inclusion of main effects and interactions of all covariates mentioned above (fitted with maximum likelihood), as well as random intercepts and slopes of subjects, classrooms, and schools for the random effects structure (fitted with restricted maximum likelihood estimation). A potential covariate was only kept if it reduced AIC by at least two, thus indicating a better model fit while penalising the increase in complexity of the model. Where available, we always tested for inclusion of raw and percentile/norm scores as predictor (e.g., CELF phonological awareness yielded both raw and norm scores, we tested the inclusion of both, and if they reduced AIC by at least two, we picked the one with the lower AIC).

To ensure that presented effects are not carried by outliers, for each resulting model, we trimmed observations based on residuals beyond ±2 standard deviations of the model prediction and refitted the model. Each model then underwent model criticism to ensure that reported models fulfil regression assumptions of independence of observations as well as a normal and homoscedastic distribution of residuals. Usually, model fit is evaluated by the squared correlation between the observed and the fitted values (R^2). For mixed-effects models, this method can only estimate the residual variance and thus ignores the random effects present in the model. Following the approach proposed by Nakagawa and Schielzeth (2013), marginal and conditional R^2 were calculated, instead. The former is an estimation of the fixed-effects structure alone, while the latter incorporates both fixed and random effects.

As several of the fitted models did not show intervention effects, the R~~r~~esults section will focus on those models that show effects related to the gaming conditions and our research aims (see Table 2 for specifications of all the fitted models for each outcome and the supplementary R markdown for all results).

*** Table 2 near here ***

3.10. Changes to statistical methods because of baseline analyses

Baseline comparisons showed main effects of country (see 4.3. below), whereby the Dutch children consistently outperformed the Belgian children in timed letter-sound knowledge, phonological awareness, and rapid automatised naming at T1. Therefore, contrary to the initial analysis plan, the Belgian and the Dutch samples were used independently to evaluate the research questions. This doubled the number of planned analyses from 7 to 14 models and reduced statistical power due to smaller group sizes.

In addition, the described approach did not yield answers for RQ4 in that none of the game exposure measures ended up as relevant predictors (see 4.6. below). In an explorative approach, we therefore combined children from both countries to increase statistical power, merged all exposure measures by means of a principal component analysis and fitted two additional models for the outcome of reading fluency using non-linear mixed effects regression (Generalized additive model; Wood, 2006).

3.11. Research ethics

This research was approved by the ethical committee of the Faculty of Arts of the University of Groningen, and the Faculty of Psychology of the University of Ghent (2015/25).

4. Results

4.1. Participant flow

Out of 312 children in the selected classrooms, parents of 26 children did not give consent to participate. Eight children were lost to follow up as they did not attend the second assessment. For the final analysis, we also excluded data of children that were allowed to play but did not consent to the behavioural assessments ($N = 4$), were repeating the first grade ($N = 2$), were one year older than their peers without repeating first grade ($N = 1$), and those who were diagnosed with a neurodevelopmental disorder ($N = 5$). To conduct a response to intervention analysis, we further excluded children who failed to play at least 20 sessions (corresponding to four weeks of daily playing) or alternatively failed to accumulate at least 2.5 hours of game exposure ($N = 19$). The details of participant flow are specified in the CONSORT flow diagram (Figure 1).

*** Figure 1 near here ***

4.2. Data losses and exclusions

Due to cases of missing observations and trimming of outliers after model fitting, most analyses were ~~carried out~~conducted on smaller subsets of the data. Exact sample sizes are reported at the corresponding positions of the results section. Most notably, because of data retrieval problems, the results of the in-game assessments at T2 were lost for 66 children putting the available sample size at 60, 71, and 49 participants for the reading, math, and passive group, respectively.

4.3. Baseline data

At T1, the Dutch children significantly outperformed their Belgian peers in terms of abstract reasoning ($F_{1,242} = 15.74$, $p < .001$), letter knowledge ($F_{1,241} = 288.84$, $p < .001$), both phonological awareness tests (CELF: $F_{1,238} = 59.68$, $p < .001$; PROEF: $F_{1,242} = 37.81$, $p < .001$), and both rapid automatised naming measures (colours: $F_{1,242} = 31.40$, $p < .001$; objects: $F_{1,241} = 16.58$, $p < .001$; see Table 3, Session T1). For this reason, separate analyses were carried out for the Dutch and the Belgian children, as referred to also in the Statistical Methods section, above. For the Belgian children, there was a main effect of condition for rapid automatised naming of colours at T1 ($F_{2,158} = 3.06$, $p = .050$), where the math group was significantly faster than the reading group (post-hoc Tukey HSD test: $p = .039$). There was also a main effect of condition for letter knowledge ($F_{2,157} = 3.22$, $p < .043$), where the passive group knew more letters than the reading group (post-hoc Tukey HSD test: $p = .035$). For the Dutch sample, the math group had significantly more children who did not speak Dutch at home than the reading group (Fisher's exact test: $p = .018$).

*** Table 3 near here ***

4.4. First research question: response to intervention outcomes

For the main aim, we evaluated the game's effectiveness by assessing word reading fluency and phonological awareness with pen-and-paper tests. We also used two in-game tests to assess both response times and accuracy in written lexical decision and timed letter-sound identification. Main effects of gaming condition would indicate differences in response to GraphoGame-NL intervention.

4.4.1. Word reading fluency

Word reading fluency was assessed with two, one-minute reading lists at T2 (see Table 3, Session T2). While we did not find any effects associated with gaming condition in the Dutch sample, but there were effects in the Belgian group, as shown in Figure 2. At T2, neither the reading ($\beta = 0.27$, $t = 1.60$, $p = .09$) nor the passive ($\beta = -0.27$, $t = -1.63$, $p = .106$) group differed from the math group, but the reading group outperformed the passive group ($\beta = 0.55$, $t = 3.18$, $p = .002$). In terms of effect sizes, these differences were small ($d = 0.27$) for the passive and reading group compared to the math group, and medium-sized ($d = 0.55$) when comparing the reading group to the passive group. This best model was based on 150 children ($N_{\text{Passive}} = 48$, $N_{\text{Math}} = 52$, $N_{\text{Read}} = 50$, trimmed seven observations or 4.3% of data), controlled for the covariates of letter knowledge at T1, CELF phonological awareness at T1, log transformed rapid automatised naming of colours time at T1, and included random intercepts per school. The model had a conditional R^2 of 0.39 and a marginal R^2 of 0.30.

*** Figure 2 near here ***

4.4.2. Written lexical decision

The written lexical decision assessment as an additional measure of reading abilities was embedded into the last gaming session of the intervention at T2 and did not reveal any differences between groups in terms of accuracy or response times.

4.4.3. Phonological awareness

Phonological awareness at T2 was measured using the nine subtests of the CELF-IV and four subtests of the Proef. ~~We a~~Again, we did not find any effects related to gaming condition in the more advanced Dutch sample, but found ~~such~~ an effect for the CELF-IV in the Belgian children (see Figure 3). ~~Hereby,~~ ~~t~~The math group outperformed the passive group (β = 0.31, t = 2.36, p = .020) but did not differ from the reading group (β = 0.10, t = 0.83, p = .407), nor did the reading group differ from the passive ~~one~~ group (β = 0.21, t = 1.55, p = .123). This best model was based on 152 children ($N_{Passive}$ = 48, N_{Math} = 52, N_{Read} = 52, trimmed five observations or 3.1% of data), controlled for the covariates of abstract reasoning at T1, CELF phonological awareness at T1, Proef phonological awareness at T1 and included random intercepts per school. The model had a conditional R^2 of 0.64 and a marginal R^2 of 0.54.

*** Figure 3 near here ***

4.4.4. Timed letter-sound identification

The timed letter-sound identification assessment was embedded into the game itself, taking place in the ~~first (~~T1~~)~~ and ~~the last (~~T2~~)~~ gaming sessions of the intervention. ~~We found that f~~From T1 to T2, the reading game boosted accuracy in this task for the Belgian children and we saw a trend towards faster response speed in the Dutch sample. For the Belgian sample, the best model predicting single-trial accuracy at both testing sessions, for 6756 trials of 101 children ($N_{Passive}$ = 47, N_{Math} = 21, N_{Read} = 33), revealed an interaction of time × condition (see Figure 4). At T1, the passive control group knew significantly more letters than the math and reading groups (math: β = 0.53, z = 2.37, p = .018, reading: β = 0.51, z = 2.60, p = .009). While we did find a main effect of time for the math group (β = 1.78, z = 12.69, p < .001), this was significantly smaller for the passive group (β = -0.49, z = -3.02, p = .003) and somewhat bigger for the reading group (β = 0.34, z = 1.88, p = .060). The gain in accuracy of the reading group far exceeded that of the passive group (β = 0.83, z = 5.79, p < .001). This best fitting model controlled for game level, CELF phonological awareness at T1 and response time of the current trial. The random effect structure consisted of random intercepts per subject and target, and random slopes for previous trial response time and CELF phonological awareness at T1 by subject. The model had a conditional R^2 of 0.47 and a marginal R^2 of 0.20.

*** Figure 4 near here ***

For the Dutch sample, the best model, which predicted box-cox transformed single-~~-~~trial response times based on 3646 trials of 75 children (N_{Math} = 48, N_{Read} = 27, trimmed 212 trials or 4.9% of data), also revealed a time × condition interaction (see Figure 5). At T1, the reading and the math groups did not differ ~~from one another~~ (β = 0.01, t = 0.59, p = .597), and we found a significant main effect of time (β = 0.04, t = 7.76, p < .001). In addition, a marginally significant interaction of group and time (β = 0.01, t = 1.98, p = .052) indicates that the ~~speed up~~increased speed from T1 to T2 was bigger for the reading than the math group. This best model controlled for PROEF phonological awareness at T1, age at T1, trial number and previous trial response time. The random~~-~~effects structure consisted of intercepts per subject, class, target, and distractor order on screen, as well as random slopes for time by subject and random slopes for time by target.

*** Figure 5 near here ***

4.5. Participant characteristics

To answer the second research question (i.e., whether there are certain subgroups of children who benefit more from GraphoGame-NL exposure than others), we looked for possible interaction effects of gaming condition with pre-test scores, as well as age, binary sex, familial risk for dyslexia, abstract reasoning, and home language environment, on the above-presented response-to-intervention variables. In almost all analyses, participant characteristics explained unique variance as covariates and thus helped to describe more robust and generalisable intervention effects. However, we did not find ~~any~~ statistically significant interaction effects between participant characteristics and gaming condition, indicating that there ~~are~~ were no participant characteristics that modulated response specifically to GraphoGame-NL intervention.

Familial risk for dyslexia was assessed by parental questionnaires inquiring about the occurrence of reading difficulties in first degree relatives. According to the stepwise model building, familial risk was a relevant predictor for PROEF phonological awareness scores at T2 in the Belgian sample, reflected in slightly lower scores for children with a familial risk for dyslexia across all gaming conditions (β = -0.23, t = -1.34, p = .183) with a small effect size (d = 0.23). Otherwise, we found no evidence that familial risk for dyslexia ~~had an e~~affected ~~on~~ response to intervention.

Age was a relevant covariate in analyses of in-game response times of timed letter-sound identification (β = 0.02, t = 2.22, p = .030) at both testing points in the Dutch sample and of response times in the written lexical decision tasks (β = -0.20, t = -3.17, p = .002) at T2 in the Belgian sample. In both cases, on average, younger children took ~~on average~~ longer to respond than older children.

Sex was a relevant covariate for letter-sound identification accuracy in the Dutch sample at both testing points (β = -1.62, t = -4.59, p < .001) and for word reading fluency at T2 in the combined sample (F = 8.775, p < .001). In both cases, girls outperformed their male peers by a significant margin.

Nonverbal intelligence as measured by the SON-R 6-40 was a relevant covariate for the CELF phonological awareness at T2 in the Belgian sample ($\beta = 0.17$, $t = 2.65$, $p = .009$), with higher nonverbal intelligence at T1 ~~resulting in slightly~~associating with higher phonological awareness scores at T2 (*d = 0.17).*

Home language environment and handedness were never ~~came up as~~ relevant predictors.

4.6. Intervention properties

To answer the third research question, whether in-game metrics would be relevant predictors for training outcomes, we looked for possible associations between game exposure and response to intervention. Exposure measures potentially reflecting learning opportunity were the number of played sessions, played hours, played levels, seen items, given responses, and the maximum game level ~~that was~~ reached by the end of the intervention. The separate inclusion of these six measures of game exposure and progress was tested in all models fitted for the seven outcomes mentioned above and in no case ~~did they end up being a~~were they relevant predictor for intervention outcomes (see 4.4.1 - 4.4.4).

As an additional explorative analysis, we focussed on the word reading fluency outcome, re-included children who played less than 20 sessions ($N = 15$), and merged children from both countries to increase statistical power ($N = 210$). For this purpose, one additional model was fitted, in which the word reading fluency score was modelled nonlinearly as a function of an interaction of continuous independent variables (principal component of the six game exposure variables, phonological awareness, rapid automatised naming, letter-sound knowledge). The best model contained two nonlinear interaction surfaces of CELF phonological awareness at T1 and the first principal component of the six exposure variables. For the reading group, this nonlinear interaction was significant ($F = 2.99$, $p = .009$), while, for the math group, it was not ($F = 0.20$, $p = .653$). As shown in Figure 6, within the math group (subplot A), there is an almost linear relation between phonological awareness skills at T1 and reading fluency as indicated by the vertical and equidistant topographic lines, whereas, for the reading group (subplot B), there is a nonlinear interaction of these two variables. The effect of phonological awareness on the reading fluency outcome is positive when exposure is above average (i.e., a *z*-score around 1), and absent when exposure is below average (i.e., a *z*-score around -1).

*** Figure 6 near here ***

4.7. Assessment tools

Our fourth research question asked whether using in-game assessments of literacy skills allows us to identify the response to intervention more reliably than traditional pen-and-paper tests. For the assessment of reading fluency, we found intervention effects and group differences with a conventional one-minute reading, pen-and-paper test (see 4.4.1), which the in-game assessment did not capture (see 4.4.2). In contrast, we were able to show the advantage of obtaining in-game data of letter-sound knowledge at the item level. Analysing single-trial data showed that children who played the literacy game made more pronounced progress than their peers who played the math game or who did not play any game (see 4.4.4), whereas analysing aggregated data of the same task (i.e., by creating a count of correctly named letters within four minutes) with an ANOVA, did not reveal the same effects in terms of group differences and interactions (all $p > .1$).

5. Discussion

In this study, we evaluated the effectiveness of a newly created version of GraphoGame for Dutch-speaking beginning readers, employing active (math game) and passive (no game) control conditions in 16 first-grade classrooms in the Netherlands and Flanders. The main purpose of this game was to intensify exposure to relevant early reading materials and to provide additional training for struggling beginning readers on top of mainstream reading instruction in the classroom. The novelty in our study was not only a newly created Dutch adaptation of an existing game-based literacy framework, but also the inclusion of entire first-grade classrooms, in-game assessments and exposure parameters yielding individualised metrics allowing accurate monitoring of individual progress throughout the game. Beyond an overall evaluation of response to intervention, we explored three factors to~~, possibly~~ determine~~ing~~ the effectiveness of digital game-based learning in early literacy training within the framework of a single study: assessment tools, participant characteristics, and intervention properties.

5.1. Response to intervention

For our first and main research question, we wanted to evaluate the effect of playing GraphoGame-NL for up to seven weeks at the onset of formal reading instruction, and hypothesised that children playing GraphoGame-NL would show a larger response to intervention compared to children playing a control game or children who did not play at all. This hypothesis was partly confirmed. In the Belgian sample, children who played the literacy game improved their letter-sound knowledge more than the math and passive control groups (as measured by the accuracy in the timed letter-sound identification task). In addition, we observed faster word reading fluency in this group at T2 with small to medium-sized effects compared to the children assigned to the passive control and, to a lesser extent, the math condition. For the Dutch sample, there was a trend towards faster responses in the timed letter-sound identification task for the reading group compared to the math group. Recombining both samples revealed a nonlinear interaction of exposure to the game and phonological awareness scores at T1 for word reading fluency. Children who scored high on phonological awareness prior to training and played extensively were more fluent readers than could be expected based on phonological awareness and rapid automatised naming alone.

5.2. Participant characteristics

As for the second research question about participant characteristics, we asked whether there are certain subgroups of children who benefit more from GraphoGame-NL exposure than others and hypothesised that poor performers would benefit more. However, the effects relating to gaming condition ~~that we found ~~were mostly main effects, ~~which ~~indicating~~es~~ that there were no systematic differences between participants' proficiency levels across the three experimental groups. The only exception, pointing in the opposite direction to~~as~~ our hypothesis, ~~being the~~was that few children who performed above average in phonological awareness skills at pre-test who were comparatively faster readers when they had above average exposure to the reading game. We also anticipated that certain subgroups of children, like those at familial risk or those speaking a different language at home, might perform worse at pre-test and exhibit a different outcome from exposure to the game, but we did not find evidence for that either.

Most studies use an inclusion criterion based on scores in reading-related tests (e.g., Saine et al., 2010, 2011), the nomination by class teachers (e.g., Kyle et al., 2013), or socioeconomic status (SES; e.g., Rosas, Escobar, Ramírez, Meneses, & Guajardo, 2017). While the rationale for such inclusion criteria is clear, all these approaches pose certain difficulties. In the case of the test-based or SES-based approach, there is the question of finding the right cut-off score. Children scoring at the lower end of the population scale are more likely to perform closer to average at the next assessment, a phenomenon known as regression to the mean (Morten & Torgerson, 2004). Furthermore, teacher ratings may be subjective and based on the assessment of skills unrelated to a child’s reading abilities (Begeny, Krouse, Brown, & Mann, 2011).

To prevent such sampling bias in the present study, we invited all children from 16 classrooms to play, independent of their reading performance~~ on reading-related tasks~~, and investigated the effect of pre-test scores on training-induced skill improvement. Our approach was unintentionally strengthened further because of the large pre-test differences between the Dutch and Belgian children in our sample. These differences appear to stem from the different preschool systems, where Belgium has a stricter separation of pre-school and school, compared to the more gradual transition into formal instruction from four years of age onwards in the Netherlands. Similar differences between these two neighbouring countries have been observed in early numeracy skills (Torbeyns, Van den Noortgate, Ghesquière, Verschaffel, Van de Rijt, & Van Luit, 2002). Ultimately, this gave even further spread to the preliteracy skills in our sample and allowed us to evaluate the impact of factors such as age, familial risk for dyslexia, sex, home language environment, handedness, and an intelligence measure (i.e., ~~like~~ abstract reasoning) more exhaustively than ~~has been done in~~ previous research.

At first sight, one could argue that, due to the absence of interactions of pre-test scores and outcome, the intervention was equally effective for all children. However, when comparing results stratified by country, it is apparent that the weaker beginning readers in Belgium showed ~~overall~~ more intervention effects (both in letter-sound knowledge and word reading fluency), whereas in the more advanced Dutch sample, we found fewer effects (limited to grapheme-phoneme correspondence automation). This can be taken as evidence that individual starting levels matter for GraphoGame-NL intervention outcomes, ~~which is~~ in line with most previous studies. Training poor performers at an early stage in their literacy development usually yields group-wide benefits in easily trainable skills like letter knowledge (e.g., Brem et al., 2010; Rosas et al., 2017), and, in longer interventions, also decoding and reading (e.g., Saine et al., 2010; 2011). However, the opposite effect, that children with high pre-test scores have an increased benefit, has also been reported~~ before~~. Ruiz et al. (2017) found a small but significant advantage for~~of~~ early readers who already scored high at pre-test in timed letter knowledge. The few studies that trained entire classrooms (e.g., Jere-Folotiya et al., 2014; Koikkalainen 2015; Ronimus & Lyytinen, 2015)~~ did~~ unfortunately did not consider interaction terms with pre-test scores in their analyses, thus provide~~ing~~ no reference point for comparisons. Regarding the general role of pre-test scores as predictors for intervention outcomes, conventional reading interventions found that reading-related skills are poor predictors for the response to intervention. Improvements were rather related to levels of short-term memory and vocabulary (Byrne, Shankweiler, & Hine, 2008) - two variables which were not measured in the present study and are not routinely collected and used as covariates in analyses of reading interventions.

5.2.1. Familial risk of dyslexia

For effects relating to familial risk of dyslexia, we found that at-risk children score slightly lower across both assessment points only with respect to phonological skills, and that status of familial risk did not influence the training effectiveness. The former is somewhat surprising, given that other studies also reported weaker performance in other reading precursors for children at familial risk like rapid automatised naming (van Bergen et al., 2012; Lyytinen et al., 2004). So far, only two studies have specifically investigated the role of familial risk in GraphoGame effectiveness. While a study by Brem and colleagues (2010) did not find any distinct effects relating to familial risk either, a study by Blomert and Willems (2010) found that at-risk children did not improve as much as ~~their~~ peers ~~did~~. The authors concluded that familial risk of dyslexia is characterised by a letter-sound association and integration deficit, which the data from the present study does not support. The fact that the present study did not find ~~any~~ distinct training effects attributable to familial risk may be due to the small number of at-risk children in each condition (varying from seven to 18) or the rather weak self-report questionnaire asking for reading failure in the close family, but without requesting proof of a formal diagnosis in first degree relatives.

5.2.2. Sex

In our sample, boys had significantly poorer letter-sound knowledge and phonological awareness skills compared to girls at the start of first grade. This appears to be the onset of a constant difference which extends throughout school into adolescence, where girls outperform their male peers in terms of reading (OECD, 2010; Ming Chui & McBride-Chang, 2006; Tops, Glatz, Premchand, Callens, & Brysbaert, 2019; Torppa, Eklund, van Bergen, & Lyytinen, 2015). Sex differences therefore warrant scrutiny in literacy digital game-based learning research, also given that boys generally play more games and show a stronger preference for game-based learning than girls~~their female peers~~ (Admiraal, Huizenga, Heemskerk, Kuiper, Volman, & ten Dam, 2013; Gwee, San Chee, & Tan, 2011; Bonanno & Kommers, 2007). Ideally, studies should therefore control for sex or previous game experience in their analyses, which is currently almost never done in the field (e.g., for studies reported in McTigue et al., 2019).

5.3. Intervention properties

Concerning the third research question of intervention properties, we asked whether in-game metrics are relevant predictors for response to GraphoGame-NL intervention. Studies reporting positive GraphoGame-related effects used training durations ranging from one up to 28 weeks with an intensity of two to five training sessions per week (McTigue et al., 2019; Richardson & Lyytinen, 2014). However, whether training duration and intensity act as independent variables modifying digital game-based learning outcomes, or whether the overall exposure to the game (in hours) is a better predictor of training effectiveness, remain~~s an~~ open ~~empirical~~ questions. Furthermore, the ideal training duration and intensity may ~~differ~~ depend~~ing~~ on population properties and training goals, which raises the obligation to investigate possible interactions of training and population properties.

Previous literacy digital game-based learning studies using GraphoGame usually rely~~ied~~ exclusively on the number of gaming sessions, or the time spent playing as a measure of training intensity. Only a few studies communicate treatment fidelity measures such as attrition rates, which can be as high as 46% (Jere-Folotiya et al., 2014). We therefore extracted additional game-exposure measures, such as the highest level ~~that was reached~~attained, or total number of seen items which might capture the actual gameplay better than mere time on task. For example, even though all children played in the range of 20-30 sessions, the number of items seen within the training period had a much wider range from 5000 to 20000. This is a result of the speed and accuracy of children: responding faster ~~will yield~~results in more levels, responses and seen items, while being less accurate results in being exposed to fewer items during the same task time~~ on task~~. Due to the adaptivity of modern games which constantly adjust the difficulty level to the individual learner, different children are exposed to different content, making exposure comparisons difficult, even within the same study. Response patterns also vary over time depending on the complexity (simpler, more familiar content vs. more complex new information) of consecutive levels (Nja, 2019). We therefore hypothesised that characteristics from the gaming process itself might help explain variance in the intervention outcome. Our study provides some evidence in this direction, in that exposure to the game was positively related to reading fluency outcome, although this only applied to the participants with above average phonological awareness skills. This suggests that data extracted from in-game behaviour can indeed be used for dynamic individual assessment which implies that GraphoGame-NL can serve a diagnostic function by identifying non-responders at an early stage of literacy development (Koikkalainen et al., 2015; Puolakanaho & Latvala, 2017).

Possibly, the rather strict inclusion criterion of at least 20 playing sessions made the present sample too homogenous to find ~~any~~ interactions between intervention outcomes and exposure. Upon re-inclusion of children who played less than 20 sessions and by ~~fusing these~~combining exposure measures with a principal component analysis, we found that game exposure modulated reading fluency when phonological awareness and rapid automatised naming were statistically controlled ~~for~~. For the maturation of literacy skills, we ~~would~~ argue that the time-course of development of phonological skills plays a crucial role for the benefits of GraphoGame-NL. Playing beyond mastery of grapheme-phoneme correspondences has little impact on reading fluency when phonological skills are poor. Children with good phonological skills at pre-test benefited more from the exposure to GraphoGame-NL than children with poor phonological skills. We suggest reducing the weekly playing intensity once letter-sound knowledge accuracy reaches ceiling, and ~~instead~~ extending the overall training period. This might allow poor performers to get more out of the game, especially to give more time for maturation of phonological skills (see Borleffs, Glatz, Daulay, Richardson, Zwarts, & Maassen, 2018 for a similar suggestion based on data from GraphoGame for Standard Indonesian). Future studies should furthermore focus on identifying ~~those~~ factors that best contribute to training phonological skills. In Appendix 1, we provide a detailed description of the games used in this research, as we believe this to be crucial in enabling future research to uncover the mechanisms of (more) successful interventions.

5.4. Assessment tools

~~The~~ Our final research question ~~of the current study~~ focused on the impact of assessment tools on examining intervention effects, i.e., in-game assessment tools versus pen-and-paper assessment tools. With in-game assessments, we were able to detect intervention-related improvements in letter-sound knowledge~~,~~; however, these assessments could not confirm the results we found for reading fluency using pen-and-paper tests. We note that the implementation of our in-game written lexical decision task had poor reliability ~~measures~~ and only weak associations with other variables (see 3.4.2). Therefore, it might ~~not~~ be an unsuitable tool to capture reading skills, at least in this group of beginning readers. Another possible reason why an in-game effect for reading fluency was not observed can be seen as a question of the distance of learning transfer. Measuring (timed) letter-sound knowledge before and after a training of letter-sound correspondences can be considered a near training transfer because both are closely related, whereas evaluating changes in reading fluency based on a combined training of letter-sound correspondences and phonological awareness can be considered a far learning transfer. These skills are not directly related, might require intermediate developmental steps, may take longer and be overall smaller (Froyen, Bonte, Atteveldt, & Blomert, 2009; Vaessen & Blomert, 2010). Indeed, improvements in letter knowledge are almost unanimously reported in literacy digital game-based learning research (Richardson & Lyytinen, 2014) as this skill is easily trainable and measurable, while improvements in phonological awareness and reading fluency are ~~rather~~ the exception (e.g., studies reported in McTigue et al., 2019; Carvalhais, Limpo, Richardson, & Castro, 2020; Lovio, Halttunen, Lyytinen, Näätänen, & Kujala, 2012; Ktisti, 2015).

In addition to learning transfer, how reading and reading-related skills are measured in intervention studies can have ~~a~~ considerable impact on the conclusions ~~we draw~~. This is ~~also an aspect that can be seen~~important in our~~the~~ findings ~~of the present study~~. When analysing the letter-sound knowledge task at the level of the individual letter, we find intervention-related gains which were not captured using aggregated data from the same task, which would arguably correspond to a conventional pen-and-paper assessment. In the case of letter-sound knowledge for example, pen-and-paper tests are typically administered without time pressure and reach ceiling within the first few months of school (Blomert & Willems, 2010), thus losing predictive and evaluative power. By administering a speeded letter-sound knowledge task which measures response times on top of accuracy, the fluency with which letter-sound associations are retrieved from memory can also be assessed. Such a task is indeed more specifically related to the fluency of multimodal processing of audio-visual information (Blomert, 2011; Hahn at al., 2014). Thus, these in-game assessments, by measuring accuracy and response times at the item level, tap into the domain of automatisation to an extent which conventional pen-and-paper tests are ~~not~~ unable to capture. Our study therefore demonstrates that the evaluation of response to intervention depends on the choice of outcome measure and the accompanying assessment tool and the statistical analysis. We further showed that in-game behaviour in serious games provides potentially sensitive measures to dynamically assess (pre)literacy skills.

5.5. Limitations

We acknowledge several limitations in the design and procedure, which should be considered when interpreting ~~the~~ our results and analyses ~~presented above~~. The unexpectedly large pre-test differences forced us to split our sample by country, which led to smaller groups and reduced power compared to the study we initially conceived. Due to significant group differences at T1, we cannot rule out regression to the mean as a possible explanation for some of our effects ~~described above~~. The analyses ~~presented here~~ also tested the inclusion of a wide range of measures as covariates in a conservative, yet exploratory fashion. We highly recommend replication of our study with other cohorts of Dutch and Flemish children.

An additional weakness is that we only measured reading fluency at T2. Due to an earlier pilot showing floor results and due to time constraints for testing at schools, we decided not to collect such data at T1. As a result, we could not directly test interactions between reading fluency improvement and other factors. However, by controlling reading fluency outcome for reading precursors at T1 (letter knowledge, phonological awareness, rapid automatised naming, and age), these results are nevertheless relevant and meaningful.

Another issue arises from the fact that the teachers who participated were favourable, or at least open, to~~wards~~ the use of digital tools in their classrooms, and were furthermore not blinded to the gaming conditions. Thus, they knew their treatment allocation. This may have changed their teaching style ~~in one way or another~~, which is ~~something that is hard~~difficult to control or correct for. To balance ~~out~~ the impact single classrooms may have on intervention effects, children should ideally be randomised individually (for example, one third of a classroom playing the reading game, one third playing a control game and one third not playing). From our experience, this is hard to implement in classrooms and it would also negatively affect classroom atmosphere if some children were not allowed to play. An~~other~~ alternative ~~could be~~is to implement the playing at home, which would come with its own set of challenges, like how to ensure daily playing or prevent excessively long gaming sessions (Ronimus & Lyytinen, 2015).

Finally, the math game may not have been the best control condition. Through data collected with an auditory EEG paradigm from a subset of the children in the present study, it became apparent that playing the math game might also contribute to the development of phonological awareness skills (Glatz, 2018). As arithmetic representations are also phonological in nature (De Smedt & Boets, 2010; De Smedt, Taylor, Archibald, & Ansari, 2010), both games ultimately promote careful listening and fast access to phonological representations. Future research on computerised literacy training should therefore ~~try to make use of~~implement an active control condition where the improvements of video gaming can be expected in the visual or motor domain rather than in verbal and/or auditory learning.

7. Conclusion

We conducted one of the first literacy digital game-based learning studies relying on single-trial data from in-game tasks to evaluate its effectiveness. Playing GraphoGame-NL led to an increase in mastery of grapheme-phoneme correspondences and to small ~~effects~~ improvements in reading fluency. Demographic characteristics such as familial risk of dyslexia or languages/dialects spoken at home had little impact on response to intervention. This study presented evidence that GraphoGame-NL can serve a diagnostic function and thus replace or extend assessment by means of conventional tests of literacy skills.

8. References

Admiraal, W., Huizenga, J., Heemskerk, I., Kuiper, E., Volman, M., & ten Dam, G. (2014). Gender-inclusive game-based learning in secondary education. *International Journal of Inclusive Education*, *18*(11), 1208-1218. doi: 10.1080/13603116.2014.885592

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723. doi: 10.1109/TAC.1974.1100705

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub. doi: 10.1176/appi.books.9780890425596

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi: 10.18637/jss.v067.i01

Begeny, J. C., Krouse, H. E., Brown, K. G., & Mann, C. M. (2011). Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. *School Psychology Review*, *40*(1), 23-38. doi: 10.1080/02796015.2011.12087726

van Bergen, E., de Jong, P. F., Plakas, A., Maassen, B., & van der Leij, A. (2012). Child and parental literacy levels within families with a history of dyslexia. Journal of Child Psychology and Psychiatry, 53(1), 28-36. doi: 10.1111/j.1469-7610.2011.02418.x

Bergmann, J., & Wimmer, H. (2008). A dual-route perspective on poor reading in a regular orthography: Evidence from phonological and orthographic lexical decisions. *Cognitive Neuropsychology*, *25*, 653-676. doi: 10.1080/02643290802221404

Blomert, L. (2011). The neural signature of orthographic–phonological binding in successful and failing reading development. *Neuroimage*, *57*(3), 695-703. doi: 10.1016/j.neuroimage.2010.11.003

Blomert, L., & Willems, G. (2010). Is there a causal link from a phonological awareness deficit to reading failure in children at familial risk for dyslexia?. *Dyslexia*, *16*(4), 300-317. doi: 10.1002/dys.405

Bonanno, P., & Kommers, P. (2007, July). Exploring the Influence of Group characteristics on Interactions during Collaborative Gaming. In *IADIS International Conference: e-Learning held in Lisbon, Portugal*.

Borleffs, E., Glatz, T. K., Daulay, D. A., Richardson, U., Zwarts, F., & Maassen, B. A. (2018). GraphoGame SI: the development of a technology-enhanced literacy learning tool for

Standard Indonesian. *European Journal of Psychology of Education*, *33*(4), 595-613. doi: 10.1007/s10212-017-0354-9

Borleffs, E., Maassen, B. A. M., Lyytinen, H., & Zwarts, F. (2017). Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: A narrative review. *Reading and Writing*, 1-22, doi: 10.1007/s11145-017-9741-5

van den Bos, K. P. (1998). IQ, phonological awareness and continuous-naming speed related to Dutch poor decoding children's performance on two word identification tests. *Dyslexia*, *4*(2), 73-89.

van den Bos, K. P., Spelberg, H., Scheepsma, A., & De Vries, J. (1994). De Klepel. Vorm A en B. Een test voor de leesvaardigheid van pseudowoorden. Verantwoording, handleiding, diagnostiek en behandeling. Berkhout, Nijmegen, The Netherlands.

van den Bos, K. P. (2003). Snel Serieel Benoemen; Experimentele versie. [Rapid naming; Experimental version]. Groningen: University of Groningen.

van den Bos, K. P., & Lutje Spelberg, H. C. (2010). CB&WL Continu Benoemen & Woorden Lezen. Verantwoording [Continuous Naming & Reading Words. Technical Manual]. Amsterdam: Boom Testuitgevers.

Brem, S., Bach, S., Kucian, K., Kujala, J. V., Guttorm, T. K., Martin, E., ... & 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. doi: 10.1073/pnas.0904402107

Brus, B., & Voeten, M. (1991). Een-minuut-test vorm A en B, schoolvorderingstest voor de technische leesvaardigheid bestemd voor groep 4 tot en met 8 van het basisonderwijs. Verantwoording en handleiding. Lisse: Swets & Zeitlinger.

Byrne, B., Shankweiler, D., & Hine, D. W. (2008). Reading development in children at risk for dyslexia. In M. Mody, & E. R. Slliman (Eds.), Brain, behavior and learning in language and reading disorders (pp. 240–270). NY: The Guilford Press.

Carvalhais, L., Limpo, T., Richardson, U., & Castro, S. L. (2020). Effects of the Portuguese Graphogame on reading, spelling, and phonological awareness in second graders struggling to read. *The Journal of Writing Research*, *12*(1). doi: 10.17239/jowr-2020.12.01.02

Chambers, B., Abrami, P., Tucker, B., Slavin, R. E., Madden, N. A., Cheung, A., & Gifford, R. (2008). Computer-assisted tutoring in Success for All: Reading outcomes for first graders. *Journal of Research on Educational Effectiveness*, *1*(2), 120-137. doi: 10.1080/19345740801941357

Desoete, A., Praet, M., Van de Velde, C., De Craene, B., & Hantson, E. (2016) Enhancing mathematical skills through interventions with virtual manipulatives. In Patricia S. Moyer-Packenham (Eds.) International Perspectives on Teaching and Learning Mathematics with Virtual Manipulatives (pp.171-187). Springer: Switserland. doi: 10.1007/978-3-319-32718-1

D’hondt, M., Desoete, A., Schittekatte, M., Kort, W., Compaan, E., Neyt, F., ... & Surdiacourt, S. (2008). De CELF-4-NL: een opvolger voor de TvK. *Signaal*, *65*, 4-16.

Elen, R. (2006). Proef Fonologisch Bewustzijn (PFB): handleiding, materiaal, scoreformulieren. *Vlaamse Vereniging voor Logopedisten.*

Froyen, D. J., Bonte, M. L., van Atteveldt, N., & Blomert, L. (2009). The long road to automation: neurocognitive development of letter–speech sound processing. *Journal of Cognitive Neuroscience, 21*(3), 567-580. doi: 10.1162/jocn.2009.21061

Geelhoed, J. W., & Reitsma, P. (1999). PI-dictee.

Glatz, T. (2018). Serious games as a level playing field for early literacy: A behavioural and neurophysiological evaluation (Doctoral dissertation). Retrieved from the library of the University of Groningen.

Gwee, S., San Chee, Y., & Tan, E. M. (2011). The role of gender in mobile game-based learning. *International Journal of Mobile and Blended Learning (IJMBL)*, *3*(4), 19-37. doi: 10.4018/978-1-4666-2139-8.ch016

Hahn, N., Foxe, J. J., & Molholm, S. (2014). Impairments of multisensory integration and cross-sensory learning as pathways to dyslexia. *Neuroscience & Biobehavioral Reviews, 47*, 384-392. doi: 10.1016/j.neubiorev.2014.09.007

Jere-Folotiya, J., Chansa-Kabali, T., Munachaka, J. C., Sampa, F., Yalukanda, C., Westerholm, J., ... & Lyytinen, H. (2014). The effect of using a mobile literacy game to improve literacy levels of grade one students in Zambian schools. Educational Technology *Research and Development*, *62*(4), 417-436. doi: 10.1007/S11423-014-9342-9

Kahan, B. C., Jairath, V., Doré, C. J., & Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomised trials: an assessment of 12 outcomes from 8 studies. *Trials*, *15*(1), 1-7. doi: 10.1186/1745-6215-15-139

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior research methods*, *42*(3), 627-633. doi: 10.3758/BRM.42.3.627

Koikkalainen, M. (2015). Computerised reading fluency assessment: Task validity and the strongest discriminators of fluency skills among second-graders. [Master's thesis] University of Jyväskylä. http://urn.fi/URN:NBN:fi:jyu-201510023300

Kort, W., Schittekatte, M., & Compaan, E. (2008). *CELF-4-NL: clinical evaluation of language fundamentals*. [Dutch version]. Pearson.

Ktisti, C. (2015). *Computer-based remediation for reading difficulties in a consistent orthography: comparing the effects of two theory-driven programs* (Doctoral dissertation). Retrieved from the library of the University of Cyprus.

Kujala, J. V., Richardson, U., & Lyytinen, H. (2010). A Bayesian-optimal principle for learner-friendly adaptation in learning games. *Journal of Mathematical Psychology*, *54*(2), 247-255. doi: 10.1016/j.jmp.2009.10.001

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. doi: 10.1002/rrq.038

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H., Lohvansuu, K., ... & Kunze, S. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry, 54*(6), 686-694. doi: 10.1111/jcpp.12029

van der Leij, A., Bergen, E., Zuijen, T., Jong, P., Maurits, N., & Maassen, B. (2013). Precursors of developmental dyslexia: an overview of the longitudinal Dutch dyslexia programme study. *Dyslexia*, *19*(4), 191-213. doi: 10.1002/dys.1463

Lovio, R., Halttunen, A., Lyytinen, H., Näätänen, R., & Kujala, T. (2012). Reading skill and neural processing accuracy improvement after a 3-hour intervention in preschoolers with difficulties in reading-related skills. *Brain research*, *1448*, 42-55. doi: 10.1016/j.brainres.2012.01.071

Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). A definition of dyslexia. *Annals of dyslexia*, *53*(1), 1-14. doi: 10.1007/s11881-003-0001-9

Lyytinen, H., Aro, M., Eklund, K., Erskine, J., Guttorm, T., Laakso, M. L., ... & Torppa, M. (2004). The development of children at familial risk for dyslexia: birth to early school age. *Annals of dyslexia*, *54*(2), 184-220. doi: 10.1007/s11881-004-0010-3

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. doi: 10.1111/j.1467-9450.2009.00791.x

Mascheretti, S., Bureau, A., Battaglia, M., Simone, D., Quadrelli, E., Croteau, J., ... & Marino, C. (2013). An assessment of gene-by-environment interactions in developmental dyslexia-related phenotypes. *Genes*, *Brain and Behavior*, *12*(1), 47-55. doi: 10.1111/gbb.12000

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. doi: 10.1002/rrq.256

Ming Chui, M. & McBride-Chang, C. (2006). Gender, context, and reading: A comparison of students in 43 countries. Scientific Studies of Reading, 10(4), 331–362. doi:10.1207/s1532799xssr1004_1.

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., ... & Tóth, D. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learning and Instruction*, *29*, 65-77. doi: 10.1016/j.learninstruc.2013.09.003

Mommers, M. J. C., Verhoeven, L., & Van der Linden, S. (1990). Veilig leren lezen. *Zwijsen, Tilburg*.

Morton, V., & Torgerson, D. J. (2003). Effect of regression to the mean on decision making in health care. *Bmj*, *326*(7398), 1083-1084. doi: 10.1136/bmj.326.7398.1083

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution 4*(2), 133-142. 10.1111/j.2041-210x.2012.00261.x

Nerbonne, J., Heeringa, W., Van den Hout, E., Van der Kooi, P., Otten, S., & Van de Vis, W. (1996). Phonetic distance between Dutch dialects. In *CLIN VI: Proceedings of the sixth CLIN meeting* (pp. 185-202).

Nietfeld, J. L., Shores, L. R., & Hoffmann, K. F. (2014). Self-regulation and gender within a game-based learning environment. *Journal of Educational Psychology*, *106*(4), 961. doi: 10.1037/a0037116

Njå, M. (2019). Players' progression through GraphoGame, an early literacy game: influence of game design and context of play. *Human Technology*, *15*(2). doi: 10.17011/ht/urn.201906123157

OECD (2010). *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)*, PISA, OECD Publishing. doi:10.1787/9789264091450-en

Patel, P., Torppa, M., Aro, M., Richardson, U., & Lyytinen, H. (2018). GraphoLearn India: The effectiveness of a computer-assisted reading intervention in supporting struggling readers of english. *Frontiers in psychology*, *9*, 1045.

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition*, *101*(2), 385-413. doi: 10.1016/j.cognition.2006.04.008

Piquette, N. A., Savage, R. S., & Abrami, P. C. (2014). A cluster randomised control field trial of the ABRACADABRA web-based reading technology: replication and extension of basic findings. *Frontiers in psychology*, *5*, 1413. doi: 10.3389/fpsyg.2014.01413

Potocki, A., Ecalle, J., & Magnan, A. (2013). Effects of computer-assisted comprehension training in less skilled comprehenders in second grade: A one-year follow-up study. *Computers & Education*, *63*, 131-140. doi: 10.1016/j.compedu.2012.12.011

Praet, M., & Desoete, A. (2014). Number line estimation from kindergarten to grade 2: a longitudinal study. *Learning and Instruction*, *33*, 19-28. doi: 10.1016/j.learninstruc.2014.02.003

Puolakanaho, A., & Latvala, J. M. (2017). Embedding Preschool Assessment Methods into Digital Learning Games to Predict Early Reading Skills. *Human Technology*, *13*, 216-236. doi: 10.17011/ht/urn.201711104212

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. retrieved from: http://www.R-project.org/

Regtvoort, A., Zijlstra, H., & van der Leij, A. (2013). The Effectiveness of a 2-year Supplementary Tutor-assisted Computerised Intervention on the Reading Development of Beginning Readers at Risk for Reading Difficulties: A Randomised Controlled Trial. *Dyslexia*, *19*(4), 256-280. doi: 10.1002/dys.1465

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. doi: 10.17011/ht/urn.201405281859

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. doi: 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. *Infancia y Aprendizaje*, *40*(1), 158–188. https://doi.org/10.1080/02103702.2016.1263451

Ruiz, J. P., Lassault, J., Sprenger-Charolles, L., Richardson, U., Lyytinen, H., & Ziegler, J. C. (2017). GraphoGame: un outil numerique pour enfant en difficultes d'apprentissage de la lecture. *ANAE Approche Neuropsychologique des Apprentissages chez l'Enfant, 148*, 333-343.

Rutter, M., Caspi, A., Fergusson, D., Horwood, L. J., Goodman, R., Maughan, B., ... & Carroll, J. (2004). Sex differences in developmental reading disability: new findings from 4 epidemiological studies. *Jama*, *291*(16), 2007-2012. doi: 10.1001/jama.291.16.2007

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

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The statistician*, 169-178. doi: 10.2307/2348250

Savage, R., Abrami, P. C., Piquette, N., Wood, E., Deleveaux, G., Sanghera-Sidhu, S., & Burgos, G. (2013). A (Pan-Canadian) cluster randomised control effectiveness trial of the ABRACADABRA web-based literacy program. *Journal of Educational Psychology*, *105*(2), 310. doi: 10.1037/a0031025

Schaerlaekens, A. M., Kohnstamm, G. A., Lejaegere, M., & Vries, A. K. (1999). *Streeflijst woordenschat voor zesjarigen: gebaseerd op nieuw onderzoek in Nederland en België*. Swets & Zeitlinger.

Schneider, W., Roth, E., & Ennemoser, M. (2000). Training phonological skills and letter knowledge in children at risk for dyslexia: a comparison of three kindergarten intervention programs. *Journal of Educational Psychology*, *92*(2), 284. doi: 10.1037/0022-0663.92.2.284

Schulte-Körne, G., Deimel, W., Bartling, J., & Remschmidt, H., (1998). Auditory processing and dyslexia: evidence for a specific speech processing deficit. *Neuroreport*, *9*, 337-340. doi: 10.1097/00001756-199801260-00029

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*, *8*(1), 18. doi: 10.1136/bmj.c332

Schumacher, J., Hoffmann, P., Schmäl, C., Schulte-Körne, G., & Nöthen, M. M. (2007). Genetics of dyslexia: the evolving landscape. *Journal of medical genetics*, *44*(5), 289-297. doi: 10.1136/jmg.2006.046516

Seymour, P. H., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of psychology*, *94*(2), 143-174. doi: 10.1348/000712603321661859

De Smedt, B., & Boets, B. (2010). Phonological processing and arithmetic fact retrieval: evidence from developmental dyslexia. *Neuropsychologia, 48*(14), 3973-3981. doi: 10.1016/j.neuropsychologia.2010.10.018

De Smedt, B., Taylor, J., Archibald, L., & Ansari, D. (2010). How is phonological processing related to individual differences in children's arithmetic skills?. *Developmental Science, 13*(3), 508-520. doi: 10.1111/j.1467-7687.2009.00897.x

Snowling, M. J., & Melby-Lervåg, M. (2016). Oral Language Deficits in Familial Dyslexia: A Meta-Analysis and Review. *Psychological bulletin, 142*(5), 498-545. doi: 10.1037/bul0000037

Tellegen, P. J., & Laros, J. A. (2014). *SON-R 6-40. Snijders-Oomen non-verbal intelligence test*. Göttingen, Germany: Hogrefe

Tops, W., Glatz, T., Premchand, A., Callens, M., & Brysbaert, M. (2019). Study strategies of first-year undergraduates with and without dyslexia and the effect of gender. *Journal of Special Needs Education, 35*, 1-16. doi: 10.1080/08856257.2019.1703580

Torbeyns, J., Van den Noortgate, W., Ghesquière, P., Verschaffel, L., Van de Rijt, B. A., & Van Luit, J. E. (2002). Development of early numeracy in 5-to 7-year-old children: A comparison between Flanders and The Netherlands. *Educational Research and Evaluation*, *8*(3), 249-275. doi: 10.1076/edre.8.3.249.3855

Torppa, M., Eklund, K., van Bergen, E., & Lyytinen, H. (2015). Late-emerging and resolving dyslexia: A follow-up study from age 3 to 14. *Journal of Abnormal Child Psychology*, *43*(7), 1389-1401. doi: 10.1007/s10802-015-0003-1

Vaessen, A., Bertrand, D., Tóth, D., Csépe, V., Faísca, L., Reis, A., & Blomert, L. (2010). Cognitive development of fluent word reading does not qualitatively differ between transparent and opaque orthographies. Journal of Educational Psychology, 102(4), 827. doi: 10.1037/a0019465

Vaessen, A., & Blomert, L. (2010). Long-term cognitive dynamics of fluent reading development. Journal of experimental child psychology, 105(3), 213-231. ISO 690. doi: 10.1016/j.jecp.2009.11.005

van Viersen, S., de Bree, E. H., Zee, M., Maassen, B., van der Leij, A., & de Jong, P. F. (2018). Pathways into literacy: The role of early oral language abilities and family risk for dyslexia. *Psychological Science, 29*(3), 418-428. doi: 10.1177/0956797617736886

Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Worth, J., Nelson, J., Harland, J., Bernardinelli, D., & Styles, B. (2018). GraphoGame Rime. Evaluation report and executive summary. *National Foundation for Educational Research.* This is not peer-reviewed publication, but an independent evaluation report.

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. doi: 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. doi: 10.1037/0033-2909.131.1.3

Zijlstra, H., van Bergen, E., Regtvoort, A., de Jong, P. F., & van der Leij, A. (2020, July 13). Prevention of Reading Difficulties in Children With and Without Familial Risk: Short- and Long-Term Effects of an Early Intervention. *Journal of Educational Psychology.* Advance online publication. http://dx.doi.org/10.1037/edu0000489