Dynamic assessment of the effectiveness of digital game-based literacy training in beginning readers

Toivo Glatz Corresp., 1, 2, 3 , **Wim Tops** 1 , **Elisabeth Borleffs** 1 , **Ulla Richardson** 4 , **Natasha Maurits** 2, 5 , **Annemie Desoete** 6 , **Ben Maassen** 1, 2, 7

1 Center for Language and Cognition (CLCG), Faculty of Arts, University of Groningen, Groningen, Netherlands

2 Research School of Behavioural and Cognitive Neuroscience (BCN), University Medical Center Groningen (UMCG), Groningen, 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 Department of Neuroscience, University Medical Center Groningen (UMCG), Groningen, Netherlands

Corresponding Author: Toivo Glatz
Email address: t.k.glatz@rug.nl

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 randomized control 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 4 to 7 weeks. Our outcome measures included reading fluency, as well as purpose built in-game proficiency levels to measure written lexical decision and letter speech 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 other two conditions. In addition to a small classroom wide benefit in terms of reading fluency, we furthermore discovered that children who scored high on phonological awareness prior to training were more fluent readers after extensive exposure to the reading game. This study is among the first to exploit game generated data for the evaluation of DGBL for literacy interventions.

Dynamic assessment of the effectiveness of digital game-based literacy training in beginning readers

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

[1]Center for Language and Cognition (CLCG), Faculty of Arts, University of Groningen, Groningen, Netherlands

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

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

[4]Department of Neurology, University Medical Center Groningen (UMCG), Groningen, Netherlands

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

[6]Department of Neuroscience, University Medical Center Groningen (UMCG), Groningen, Netherlands

Corresponding author: Toivo Glatz[1,2]

Email address: toivo.glatz@gmail.com

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. What tools do we have or need that help promote reading skills, detect problems at an early stage and prevent struggling readers from developing more serious literacy problems such as dyslexia? In this context, digital game-based learning shows potential.

Developmental dyslexia or specific learning disorder in reading (DSM-5; APA, 2013; henceforth, 'dyslexia') is a developmental disorder characterized by persistent difficulties in word recognition (reading) and/or spelling. 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 countries depending on their characteristics (Bergmann & Wimmer, 2008; Ziegler & Goswami, 2005). As a consequence 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 transparent orthographies (Seymour, Aro, & Erskine, 2003; Bergmann & Wimmer, 2008; Ziegler & Goswami, 2005).

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; they, for instance, interact with environmental factors such as parental socioeconomic status (Mascheretti et al., 2013). There are certain early childhood cognitive and behavioural precursors of future reading skills. The most prominent ones are letter knowledge as a measure of grapheme-phoneme correspondences knowledge, phonological awareness, and rapid automatized naming of familiar objects. Early below-average performance on tasks targeting these skills increases the risk of dyslexia (van der Leij, Bergen, Zuijen, Jong, Maurits, & Maassen, 2013; Lyon, Shaywitz, & Shaywitz, 2003; Lyytinen et al., 2004; Lyytinen, Erskine, Kujala, Ojanen, & Richardson, 2009). Phonological awareness and rapid automatized naming have been shown to predict reading speed and accuracy across languages (Moll et al., 2014). However, certain cross-linguistic variability exists with respect to the relative weight of each of the cognitive and behavioural precursors of reading acquisition (Landerl et al., 2013; Ziegler et al., 2010): letter knowledge being most predictive in Finnish with its extreme letter-sound consistency (Lyytinen et al., 2009), rapid automatized naming being the best long-term predictor in German (Brem et al., 2013), and letter knowledge, rapid automatized naming and phonological awareness being important indicators in Dutch (van Bergen, Jong, Plakas, Maassen, & van der Leij, 2012). An important distinction has to be made for letter knowledge, which just reflects the availability of letter-sound associations that quickly reach ceiling and therefore has limited use for long-term predictions. Adding time pressure within a speeded letter-speech sound identification task yields a measure which is even more specifically related to the fluency of multimodal processing of audio-visual information (Blomert, 2011; Hahn, Foxe, & Molholm, 2014). In addition to the availability of letter-sound association this 'timed' letter knowledge also assesses the fluency in which these associations can be retrieved from memory.

Given its multifactorial aetiology and early indicators such as poor letter knowledge, phonological awareness and rapid automatized naming, the question arises whether timely training of these skills might help remediate or even prevent reading difficulties. During the past decade, it has been shown that one promising way to deliver such a training is by computerized gaming (Chambers et al., 2008; Richardson and Lyytinen, 2014). While there is no commonly agreed definition for digital game-based learning or so called serious games (Susi, Johannesson, & Backlund, 2007) such tools have a range of interesting characteristics (De Freitas, 2006; Kiili, 2005; Prensky, 2001), as they: i) offer multimodal learning environment (Potocki, Ecalle, & Magnan, 2013), ii) provide immediate feedback for improved learning, iii) can adapt to individual learners depending on their responses, iv) are highly motivating for the players (e.g. Desoete, Praet, Van de Velde, De Craene, & Hantson, 2016), and v) can monitor the development of accuracy and response times in task-relevant contexts and provide researchers with longitudinal data (e.g. Praet & Desoete, 2014; Puolakanaho & Latvala, 2017).

GraphoGame is one such promising computerized training method targeting reading-related skills (for a review, see Richardson & Lyytinen, 2014). It is an adaptive, child-friendly digital learning environment that aims to help beginning readers. GraphoGame was originally designed for the very transparent orthography of Finnish as a tool to boost grapheme-phoneme correspondence knowledge in beginning readers by establishing accurate phonemic representations, connecting these to the orthographic stimuli, and establishing a fluent association between the two. More recent versions of the game also train phonological awareness, spelling, and reading fluency (Richardson & Lyytinen, 2014). A number of experimental studies using GraphoGame have been conducted in over 20 different languages, with overall mixed results. While most interventions led to improvements in terms of letter knowledge, only a few demonstrated actual benefits for reading skills. A recent meta-analysis revealed that the average gain in reading performance across 19 GraphoGame studies was close to zero (McTigue, Solheim, Zimmer, Uppstad, 2019). However, all these studies also differed in various methodological aspects. Be it in terms of population properties like selection criteria, age (5 to 10 years), the number and types of control groups or sample sizes per experimental group (N = 10 to 185). Similarly, there are drastic differences in terms of training implementation, i.e. whether training took place in school or at home, during hours or after hours, with or without adult engagement, how long the training period was (ranging from one to 28 weeks) and the overall exposure to the game (one to eight hours on task). Linguistic properties form a third pillar of often discussed differences, with orthographies ranging from transparent to opaque and each language having its own unique mix of cognitive precursors of reading. Therefore, training goals and content, as well as assessments of language skills will inevitably differ. An additional issue regarding the latter is that most standardised (reading) tests are not designed to detect the subtle changes that occur over a few weeks of training. The large variation of all these factors makes it challenging to grasp the characteristics that make up a successful intervention. Unfortunately, many studies also lack relevant details, which further impede comparison, be it in form of a detailed outline of the actual training material, the cognitive skills a game is meant to train, or the training environment. Note that for comparability sake we limit our discussion to GraphoGame studies, but the same questions arise with many other literacy interventions using digital game-based learning, such as ABRACADABRA (Savage et al., 2013; Piquette, Savage, & Abrami, 2014) or Bouw! (Regtvoort, Zijlstra, & Leij, 2013; Zijlstra, Van Bergen, Regtvoort, De Jong, & Van Der Leij, 2020).

In sum, the divergent outcomes of digital game-based learning studies in general, and GraphoGame studies in particular, raise the question which circumstances impact the effectiveness of literacy digital game-based learning. Deepening our understanding of how progress can be optimally measured, how game content and training frequency and intensity modulate training outcomes, and for which populations of children this approach works best might help us to improve the effectiveness of digital game-based learning for early literacy. Therefore, in this work we focus on i) the tools that are generally used to assess reading-related cognitive skills, ii) the characteristics of the populations studied, and iii) training properties such as duration, intensity, and content.

Present study: research questions and hypotheses

Our main aim is to evaluate the effectiveness of a newly created Dutch version of GraphoGame. More specifically, we want to investigate possible impacts of assessment tools, population characteristics, and intervention properties on training outcomes. For this purpose, we invited entire first grade classrooms to play GraphoGame-NL for five to seven weeks. We used traditional paper-and-pencil as well as in-game tests to track their performance.

Research Question 1: *Does the evaluation of game effectiveness depend on the choice of **assessment tools** used to track changes in performance?* First, we tested the effects of gaming on different reading-related skills: i.e., direct assessment of word reading, on the one hand, and reading related skills, on the other hand. Second, we integrated comparisons of different assessment tools for the same skill (see Methods). We hypothesized that tests of skills prerequisite for reading were more sensitive for measuring training efficiency than tests of reading fluency itself. Moreover, we presumed that online measures such as response times would be overall more sensitive than offline metrics (e.g., paper-and-pencil test).

Research Question 2: *How do **population characteristics** impact intervention effectiveness?* We hypothesized that children who had below average performance in letter knowledge and phonological awareness prior to the training would benefit most from GraphoGame-NL. We also expected that certain subgroups of children perform below average at pre-test. These would presumably be children at familial risk for dyslexia, younger children, or those speaking a foreign language at home.

Research Question 3: *How do **intervention properties** contribute to training effectiveness?* From the training phase itself, we acquired multiple metrics of child's gaming process: i.e., how many levels were played (incl. levels that had to be repeated), the highest level reached by the child, and how many targets and distractors they saw throughout the gaming. We hypothesized that these game characteristics contributed to the variability in training outcomes.

Materials and methods

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. We aimed to follow the CONSORT standards of randomized control trials (Schulz, Altman, & Moher, 2010).

Participants

Mainstream primary schools in the northern region (Groningen area) of the Netherlands (NL) and the western region (Ghent area) of the Dutch-speaking part of Belgium (B) were approached by phone or letter in which they were invited to join our study. Initial requirements to participate were that i) schools needed to have enough computers with headphones available allowing their first-grade students to play the game on a daily basis, ii) class teachers expressed their willingness to motivate and enable each student to play the game 10-15 minutes per day for at least five weeks, iii) additional behavioural tests could be carried out at school during regular school hours, and iv) teachers accepted to adhere to the (at that moment still unknown) experimental condition their classroom would be assigned to. We found eight schools willing and eligible to participate (three in the Netherlands, five in Belgium), with a total of 16 first-grade classrooms (four in the Netherlands, 12 in Belgium) and a potential of 312 children (107 in the Netherlands, 205 in Belgium).

All parents were asked for their written informed consent for the gaming and additional behavioural tests. Furthermore, they had to complete a questionnaire inquiring about their children's handedness, language(s)/dialect(s) spoken at home, reading problems in the child and/or in close family members, presence of confirmed neurological problems in the child and potential medication. To enable as many children as possible to play the game there were no initial exclusion criteria, and parents were also given the option to consent to the gaming without additional assessments. A few families made use of that latter option (*N* = 4) while some more did not consent to participation at all (*N* = 26). For final analysis we excluded data of those children that were repeating the first grade (*N* = 2), one individual that was one year older than its peers without repeating first grade, those children that were diagnosed with a neurodevelopmental disorder (*N* = 5), those that missed an assessment session (*N* = 8), and those who failed to play at least 20 sessions (corresponding to four weeks of daily playing) or alternatively accumulate at least 2.5 hours of game exposure (*N* = 19). The details of this sample are specified in the CONSORT flow diagram (Figure 1).

To avoid that children would play the wrong game, they were assigned to an experimental condition by classroom (i.e., cluster randomized) and not individually. Classrooms 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). Importantly, even the children of this passive group took part in two gaming sessions to complete the in-game assessments at pre- and post-test. Where possible, a within-school design was set up: three schools participated with three classrooms, so each classroom per school was randomly assigned to one of the three experimental conditions. Another two schools joined with two classrooms, which were randomly assigned to the reading and math conditions. The final three schools joined with one classroom and each classroom was once more randomly assigned to one of the three conditions. See Table 1 for an overview of the number of children per classroom and their experimental conditions, and Table 2 for pre-test results.

Computerised training

The Dutch version GraphoGame-NL was created specifically by our research group for the present literacy study, the content of which was selected from Veilig leren lezen (VLL; 'Learning to read safely'; Mommers, Verhoeven, & van der Linden, 1990) a widely used literacy teaching method in the Netherlands and a vocabulary achievement list for six-year olds (Schaerlaekens, Kohnstamm, Lajaegere, & Vries, 1999). The game included 650 items, ranging from simple and complex graphemes (e.g. ⟨n⟩, ⟨r⟩, ⟨ui⟩), to 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 or CVCC consonant clusters (e.g. *prijs*, 'price'; *zwart*, 'black'). For a detailed description of the tasks and materials used within the game, see 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 Wuggy (Keuleers & Brysbaert, 2010). Five female students reading linguistics or speech-language pathology at the University of Groningen spoke the auditory stimuli. Naive native speakers of Dutch subsequently evaluated all items with respect to their prototypicality; the most prototypical items were then included in the game, yielding one to four different spoken realizations per target. It needs to be noted here that there are some systematic differences in pronunciation between the Dutch spoken in the Netherlands and the Dutch variety or Flemish spoken in 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 Flemish children also have exposure 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 now 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.

Assessment

The following offline paper-and-pencil tests were used as outcome measures:

Reading fluency: Participants 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 untrained items (words that did not occur in the game or in any other assessment). Words 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 did not statistically differ between lists in any of our analyses, we took the average of both lists and z-transformed the result.

Phonological Awareness 1: All phonological awareness subtests of the CELF-4-NL (Kort, Schittekatte, & Compaan, 2008) test battery were 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 norm scores, acting both as dependent and independent variables.

Phonological Awareness 2: All phonological awareness tasks of the Proef Fonologisch Bewustzijn (PFB, Elen, 2006) were 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. We analysed both, *z*-transformed raw scores as well as norm scores as dependent and independent variables.

To measure the children's progress in terms of accuracy and response times with tasks that we had incorporated within the game itself, the following game-based tests are additional outcome measures. A detailed description of these following tasks with screenshots can be found in Appendix 1.

Timed letter-speech-sound identification: 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 10 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 fast children saw all 42 targets, while slower children may have only been able to see 20 or 30. Based on results of 270 children, internal consistency was high as indicated by Cronbach *α* (0.87 and 0.93) as well as 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 as a covariate.

Written lexical decision: 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, i.e. considering accuracy and response times for each target. In line with 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 with 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 also featured a high neighbourhood density of real words at an edit distance of one grapheme, ranging from eight 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.

Finally, the following tests were administered as co-variates for the analyses:

Abstract reasoning: The analogies and categories subtests of the SON-R 6-40 (Tellegen & Laros, 2014) were used as an estimate of nonverbal fluid intelligence. Within the analogies subtest children have to identify a pattern that changes one geometrical figure into another and apply this principle to a new figure. The categories subtest presents three pictures with a common characteristic and children have to pick two additional pictures (out of five), which also possess this characteristic. 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 resulting raw scores of both subtests were averaged and *z*-transformed.

Rapid Automatized Naming (objects and colours): 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 (van den Bos, 2003). 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).

Training properties: We extracted six variables related to game progress and learning opportunity: the number of played sessions, hours and levels, as well as the overall number of seen items and given responses, and the maximum level that was reached at the end of the training. While the reading and math game are based on the same framework and mini-games, the math game generally features less distractors and shorter levels compared to the reading game, which means that even though both groups had the same exposure in terms of sessions and hours on task, children in the math group were exposed to more levels, gave more responses, and were overall exposed to less items on screen. Furthermore, the number of levels differed in both games (265 in the reading and 178 in the math game).

Procedure

Pre-tests (T1) commenced in September, three to six weeks after the start of the new school term, followed by a five to seven-week playing phase in October and November, with post-testing (T2) being conducted in November and December. All tests mentioned above were administered twice, except for abstract reasoning (T1 only), reading fluency (T2 only), and the in-game written lexical-decision task (T2 only). In addition to parents giving written informed consent prior to the start of the study, all children were asked for verbal consent before the assessments started. The behavioural tests took up to an hour each and were administered by undergraduate and graduate students in speech-language pathology or linguistics during school time.

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 10 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 had to take 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 asked them about their progress and encouraged them to give the game another try when the content became more difficult at least once a week.

All three groups completed the game-based assessments, the children in the experimental groups during the first and last playing sessions and the passive controls at some point during September or October (T1) and November or December (T2).

Data analysis

Differences at T1 were checked using two-way ANOVAs including experimental condition (reading, math, passive control) and country (Netherlands, Belgium) as predictors. Significant main effects or interactions were then followed up with a *t*-test or Tukey HSD. We found several main effects of country with the Dutch children consistently outperforming the Belgian children in letter knowledge, phonological awareness and rapid automatized naming (for details see Results section). This finding was unexpected and made our planned analyses obsolete. We thus decided to split up the analysis of this study into three parts: i) the Belgian sample, to evaluate research questions one (assessment tools) and two (population characteristics) in beginning readers with active and passive control groups, ii) the Dutch sample, to evaluate research questions one and two in a population of more advanced readers limited to an active control group, and iii) a combined sample of the reading and math groups from both countries, to specifically evaluate research question three (intervention properties) for a broad range of reading abilities.

Further statistical analyses were carried out using linear mixed effects regression in R 4.0.4 (R Core Team, 2021; Bates, Maechler, Bolker, & Walker, 2015) and non-linear mixed regression using generalized additive models (Wood, 2006). Separate models were fit for dependent variables of reading fluency, phonological awareness, letter speech sound identification (accuracy & response time) and written lexical decision (accuracy and response time) separately for both countries. Therefore, to answer our three research questions, a total of 16 mixed models were built. Due to the high number of models and effects, we will focus on those models which show effects relating to experimental conditions and research questions (see Table 3 for an overview of all models).

To identify the best model based on the Akaike's Information Criterion (AIC; Akaike, 1974), we tested the stepwise inclusion of main effects and interactions for fixed effects (fitted with maximum likelihood), as well as random intercepts and slopes for the random effects structure (fitted with restricted maximum likelihood estimation). A predictor (or more generally an effect) was only kept if it reduced AIC by at least two, thus indicating a better model fit while penalizing the increase in complexity of the model. In case the best model ended up without covering our research questions, i.e. if gaming condition was not a term in the best model, we could infer that there is no effect of gaming condition. Regardless, driven by our research questions, in these cases we still added gaming condition as a main effect to the best model to be able to report measures of significance and effect size.

The following variables were considered during the model building: age at T1, gender, gaming condition, hours played, handedness, familial risk for dyslexia, language spoken at home (mono/multilingual), abstract reasoning, letter knowledge, log transformed rapid naming speed of colours and objects, and phonological awareness. Where available, we always tested for inclusion of raw and percentile/norm scores as predictor (e.g. for CELF phonological awareness we had both raw scores and percentiles, tested the inclusion of both, and picked the one that explained more variance). For in-game measures we also tested inclusion of previous trial response time, current trial response time (in case of accuracy models) and trial number as predictors to remove autocorrelation of observations. Due to their highly skewed nature, response times were always box-cox transformed (Sakia, 1992). For analyses of game exposure and learning opportunity we furthermore considered the variables of played sessions, hours and levels, as well as numbers of given responses and items seen over gameplay and finally, the highest level that was reached at the end of the training. Apart from these fixed effects, we added random intercepts and slopes for variables such as items, subjects, classrooms, and schools. To facilitate interpretation, raw scores were centred and *z*-transformed where possible, so that the model coefficient β is identical to the effect size Cohen's *d*.

Finally, for each resulting model we trimmed outliers based on residuals beyond ±2 standard deviations of the model prediction and refitted the model to ensure that presented effects are not carried by outliers. Every model then underwent a model criticism to ensure that reported models fulfil test 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), a marginal and conditional R^2 was calculated, the former being an estimation of the fixed-effects structure alone, while the latter incorporates both fixed and random effects.

Due to data trimming and cases of missing observations, most analyses were carried out on smaller subsets of the data, and 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 the post-test are not available for 66 children ($N_{\text{Read}} = 31$, $N_{\text{Math}} = 33$, $N_{\text{Passive}} = 2$).

Results

At pre-test 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 automatized naming measures (colours: $F_{1,242} = 31.40$, $p < .001$; objects: $F_{1,241} = 16.58$, $p < .001$; see Table 2). For this reason, separate analyses were carried out within each country. Within the Belgian sample, there as a main effect of condition for rapid automatized naming colours ($F_{2,158} = 3.06$, $p = .050$) where the math group was significantly faster in than the reading group (post-hoc with TukeyHSD: $p = .039$), as well as 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 with TukeyHSD: $p = .035$). For the Dutch sample, the math group had significantly more multilingual children than the reading group (Fisher's exact test: $p = .018$) but otherwise the groups did not differ in any other pre-test measures.

Research Question 1: Assessment tools

To answer the first research question, we evaluated word reading fluency, phonological awareness and the two in-game tests of letter-speech sound identification and written lexical decision. **Word reading fluency** was assessed with two one-minute reading lists at T2. Whereas we did not find any effects associated with gaming condition in the Dutch sample, there were effects in the Belgian group (see Figure 2). 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 Belgian children (N_{Passive} = 48, N_{Math} = 52, N_{Read} = 50, trimmed seven observations or 4.3% of data), controlled for the covariates letter knowledge at T1, CELF phonological awareness at T1, log transformed rapid automatized naming 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.

Phonological awareness was measured using the nine subtests of the CELF and four subtests of the PFB. Neither for the Belgian nor for the Dutch sample did we find any effects related to gaming condition in either of the two tests.

Letter-speech sound identification assessment was embedded into the game itself. We found that the reading game boosted accuracy in the Belgian sample and a trend towards boosting response speed in the more advanced 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 session × condition (see Figure 3). At T1 the control group knew significantly more letters than the math (β = 0.53, z = 2.37, p = .018) and reading groups (β = 0.51, z = 2.60, p = .009). While we did find an effect of session for the math group (β = 1.78, z = 12.69, p < .001) this was smaller for the passive group (β = -0.49, z = -3.02, p = .003) and marginally bigger for the reading group (β = 0.34, z = 1.88, p = .060). The gain of the reading group therefore also far exceeded the passive group (β = 0.83, z = 5.79, p < .001). This best fitting model was controlling for level type, CELF phonological awareness at T1 and trial response time. The random effect structure consisted of random intercepts per subject and target, and random slopes for previous 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.

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 session × condition interaction (see Figure 4). At T1 the two groups did not differ from one another (β = 0.01, t = 0.59, p = .597) but we found a significant main effect of session (β = 0.04, t = 7.76, p < .001) and a marginally significant interaction of the two (β = 0.01, t = 1.98, p = .052) with the effect of speeding up from T1 to T2 being bigger for the reading than the math cohort. 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 session by subject and random slopes for session by target.

Written Lexical Decision assessment was embedded into the game itself and did not show any differences between the gaming conditions in terms of accuracy or response times.

Research Question 2: Population characteristics

To answer the second research question, we were looking for possible interactions of gaming condition with pre-test scores, as well as age, gender, familial risk, abstract reasoning and home language environment. In almost all analyses, population characteristics explained unique variance as covariates and thus helped to describe more robust and generalizable intervention effects. However, we did not find any interactions of population characteristics and intervention type, which suggests that improvements are comparable across subpopulations. **Familial risk** for dyslexia was assessed by parental questionnaires inquiring about the occurrence of reading difficulties in first grade relatives. Familial risk was a relevant predictor for PROEF phonological awareness scores in the Belgian sample, reflected in slightly lower scores for children with a familial risk for dyslexia (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 effect on pre-test scores or response to intervention. **Age** was a relevant covariate in analyses of in-game response times of letter-speech sound identification (β = 0.02, t = 2.22, p = .030) and written lexical decision tasks (β = -0.20, t = -3.17, p = .002). In both cases, younger children took, on average, longer to reply than older children. No effect sizes for these in-game assessments could be computed from the mixed models. **Gender** was a relevant covariate for letter-speech sound identification accuracy in the Dutch sample (β = -1.62, t = -4.59, p < .001) and word reading fluency in the combined sample (F = 8.775, p < .001). In both cases, girls outperformed their male peers by a significant margin. Again, no effect sizes for these in-game assessments could be computed. **Nonverbal intelligence** as measured by the SON-R 6-40 was a relevant co-variate for the CELF phonological awareness in the Belgian sample (β = 0.17, t = 2.65, p = .009) with higher nonverbal intelligence resulting in higher phonological awareness scores and a small effect size (*d = 0.17)*. **Home language environment** and **handedness** never came up as relevant predictors.

Research Question 3: Intervention properties

To answer the third research question, we were looking for possible interactions of gaming condition with exposure measures. The inclusion of six game progress and achievement measures was tested in all models fitted for research questions one and two, but none of them turned out being relevant, suggesting that training effects are independent of training properties. One possible explanation for this could be the exclusion of children who did not reach 20 playing sessions, which reduces variance in exposure. We therefore re-included children who played less than 20 sessions (*N* = 15), and combined children from both countries to increase statistical power, putting the available sample size for this analysis at *N* = 210. The resulting groups did not differ in any test scores at T1 (see Table 4 for sample characteristics).

The best model predicting *z*-transformed one-minute reading fluency scores at T2 for 196 children (N_{Math} = 95, N_{Read} = 101, trimmed 14 observations or 6.7% of the data) described two nonlinear interaction surfaces of CELF phonological awareness at T1 and the first principal component of exposure per group. For the reading group this nonlinear interaction was significant ($F = 2.99$, $p = .009$) while for the math group it was not ($F = 0.20$, $p = .653$; see Figure 5). 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. When phonological awareness is kept stable (e.g. at -1 or +1 *z*-scores) exposure modulates reading outcome, but only when exposure is above average. The difference between these two surfaces (subplot C) indicates that children with good phonological awareness skills and a lot of exposure to the reading game are more proficient readers with a medium-sized effect of Cohen's $d = 0.5$ than their peers from the math group. With additional main effects for country, as well as two nonlinear smooths for rapid automatized naming colours time at T2 and CELF phonological awareness at T1, this model explained 49.8% of variance in the reading fluency scores.

To investigate whether this effect within the reading group was carried by specific subpopulations we included gender and familial risk for dyslexia as covariates. This led to a similar model ($N = 98$, N_{male} = 51, N_{female} = 47, trimmed six observations or 5.8% of data) with two nonlinear interaction surfaces of CELF phonological awareness at T1 and game exposure for males ($F = 6.27$, $p < .001$) and females ($F = 8.28$, $p < .001$). Upon visualization (see Figure 5) it appears that the pattern of girls mirrors that seen for the reading game in general (subplot B vs. E). The difference between boys and girls who played the reading game, which was also significant ($F = 6.32$, $p < .001$, subplot F), shows a somewhat diffuse pattern without a clear interpretation. This model further included a nonlinear smooth for CELF phonological awareness at T1 and explained 80.1% of variance in reading fluency scores. Notably, the effects described above are only carried by a handful of children each and the model split by status of familial risk of dyslexia did not reveal additional effects, probably because of the limited number of children at familial risk for dyslexia in the current sample.

Discussion

In this study, we evaluated the effectiveness of a newly created version of GraphoGame for Dutch-speaking beginning readers, employing an active (math game) and a passive (no game) control condition in 16 first-grade classrooms in the Netherlands and Flanders, the Dutch-speaking part of Belgium. The main purpose of this game was to intensify exposure to relevant early reading materials and to provide additional training for struggling beginning readers. First of all, we found large differences between the two countries at both testing points irrespective of training. Within the Belgian sample, the children who played the literacy game improved their letter knowledge more than the other two groups (as measured by the accuracy in the timed letter-speech sound identification task) and we observed somewhat faster reading fluency in this group at post-test compared to the other two conditions with small to medium-sized effects. Within the overall more advanced Dutch sample, there was a trend towards faster responses in timed letter knowledge within the reading group compared to the math group. Combining both samples revealed that children who played extensively and scored high on phonological awareness prior to training were more fluent readers than could be expected based on other reading precursors and based on phonological awareness alone. Beyond an overall evaluation of effectiveness, our secondary aim was to conduct an exploration of further factors, possibly determining the effectiveness of digital game-based learning in early literacy training within the framework of a single study. Thus, three potential factors contributing to digital game-based learning effectiveness were investigated: i) assessment tools, ii) population characteristics, and iii) intervention properties.

Research Question 1: Assessment tools

We asked whether the evaluation of game effectiveness partly depends on the choice of assessment tools used to track changes in performance. We measured reading-related skills as well as reading itself, by using different assessments for the same skill, and also combining online and offline measures. To evaluate reading-related skills studies typically use assessments like word (e.g., EMT; Brus & Voeten, 1991) and pseudo word reading tests (e.g., De Klepel; van den Bos, Spelberg, Scheepsma, & De Vries, 1994), word dictation (e.g. PI-dictee; Geelhoed & Reitsma, 1999) or tests for vocabulary, phonological awareness, or rapid automatized naming (e.g. CELF-4-NL; Kort et al., 2008). However, these paper-and-pencil tests are usually not designed to detect the subtle changes in performance occurring during a few weeks of digital game-based learning. If an improvement is not big enough, it may simply get lost within variance related to sensitivity, specificity or test-retest reliability of a given test. For example, the CELF-4-NL manual states reliabilities in the range of 0.71 to 0.86 for subscales, and 0.88 to 0.92 for composite scores as well as a test-retest reliability over a 5-month period of 0.75. In addition to potential reliability limitations, such assessments are not designed to be used multiple times within a short period of time to capture small improvements. With the number of test items being limited, the likelihood that items are remembered from previous sessions is high. Due to regression to the mean, children scoring at the lower or higher end of the population scale are more likely to perform closer to average at the next assessment (Morten & Torgerson, 2004). Taken together, there is a lack of sensitivity to change, which may be one of the main reasons why shorter training studies fail to find significant improvements in reading related skills even though the games trained exactly these skills.

Two more issues need to be discussed in this context. Firstly, poor sensitivity to change could be seen as a question of learning transfer: measuring (timed) letter knowledge before and after a training of grapheme-phoneme correspondence can be considered a near training transfer, whereas evaluating changes in reading fluency based on a combined training of grapheme-phoneme correspondences and phonological awareness can be considered a far transfer, which may take longer and be overall smaller (Froyen, Bonte, Atteveldt, & Blomert, 2009; Vaessen & Blomert, 2010). And 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). To assess phonological awareness in the present study two paper-and-pencil tests were used which differed in nature. Whereas the CELF-IV (Kort et al., 2008) tests nine different abilities one by one (e.g. segmentation, blending, phoneme identification, deletion and replacement), the PROEF (Elen, 2006) constantly alternates between one of four abilities (rime, segmentation, blending, pseudoword repetition). Our assumption was that the PROEF would better reflect the automatization of phonological processing, but we did not find any training effects with either test. While not being a far learning transfer per se, training of phonological skills and/or measuring potential improvements appear to be difficult to achieve for most short-term interventions.

Secondly, poor sensitivity to change is also a matter of how skills are assessed. For letter knowledge, paper-and-pencil 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. On top of accuracy, one can also measure response times and add time pressure, which within a speeded letter-sound association task is even more specifically related to the fluency of multimodal processing of audio-visual information (Blomert, 2011; Hahn at al., 2014). Whereas letter knowledge just assesses the availability of letter-sound associations, letter-speech sound identification (here also called 'timed' letter knowledge) additionally assesses the fluency with which these associations can be retrieved from memory. Evaluation of digital game-based learning effectiveness might therefore better be based on data provided by the game itself and by separate assessment levels in-between training levels which track changes in accuracy and response times for individual learners. For this reason, we incorporated separate test units within the game to be played at the start and the end of the training period to measure progress in reading-related skills. These in-game assessments measure accuracy and response times at the item level, and thereby tap into the domain of automatization to an extent which offline paper-and-pencil tests are not able to capture. This research further contributes to the field, by making use of mixed effects regression of single trial data which allows to take into account that test items differ regarding their difficulty and the time it takes to respond.

We found that children who played the literacy game made more pronounced progress in letter knowledge than their peers who played the math game and those who did not play any game. Although it can be argued that most children would eventually have attained the same skill levels without the game, our findings confirm that GraphoGame can speed up acquisition of grapheme-phoneme correspondences across children of all abilities. Our attempt to measure reading fluency in form of a written lexical decision task inside the game was not successful. The reliability measures were poor and there were only few associations between the outcome of that assessment and other measured variables. The task was perhaps too difficult for starting readers and might only become a relevant measure for reading fluency at a later stage. Even beyond that, the use of in-game assessments remains methodologically problematic (Puolakanaho & Latvala, 2017). For instance, when individual children do not understand the task or if they find it too difficult or boring, they may just randomly click around to pass the level, achieving very fast response times but correspondingly low accuracy scores. This is easily controlled for in group-wide analyses by excluding the data of children performing below chance level, although it may be difficult to set a chance level because weighting the complexity and confusability of the many items presented simultaneously may pose a problem (Kujala, Richardson, & Lyytinen, 2010). To obtain useable data, in our experience it is useful to have an adult supervise the assessments in small groups of children. If performed individually and unsupervised, the assessments may generate little useable data because the tests may not have been performed as intended. Possibly, assessments can be repeated at certain intervals, but ultimately, it would be most convenient to collect dynamic in-game data that considers the entire gameplay by continuously tracking a child's progress, possibly even precluding the need for dedicated assessment levels.

Research Question 2: Population characteristics

Our question was whether population characteristics impact intervention effectiveness. We hypothesized that poor performers would benefit most from GraphoGame, but we did not find any evidence for that. Most effects relating to gaming condition were main effects, which indicates that there were no systematic differences within the three experimental groups. The only exception, albeit pointing the opposite direction, being the few children who performed above average in phonological awareness skills at pre-test, who were comparatively faster readers conditional to more extensive 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 also exhibit a different outcome from exposure to the game, but we did not find evidence for that either.

Literacy interventions usually target poor performers, who are a generic group of children in whom the underlying mechanism of reading-related difficulties may vary drastically. To account for this variability, both reading-related performance and the presence/absence of familial risk for dyslexia need to be taken into account as such children are more likely to share a common underlying deficit accounting for their reading deficiencies (Snowling & Melby-Lervag, 2016; van Viersen, de Bree, Zee, Maassen, van der Leij, de Jong, 2018). The questions whether poor performers and children at familial risk of dyslexia can profit from digital game-based learning training, and whether or how these groups differ from each other are of high clinical relevance. Unfortunately, such a question remains difficult to answer if **inclusion criteria** vary across studies and only certain children take part. 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 case of the test-based or SES based approach, there is the question of finding the right cut-off score. Furthermore, due to **regression to the mean**, children scoring at the lower end of the population scale are more likely to perform closer to average at the next assessment (Morten & Torgerson, 2004). On the other hand, 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 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 drastic 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, often requiring a physical change of school around the age of six, the Netherlands has a more gradual transition into formal instruction from four years of age onwards within the same institution. This interpretation is also supported by the fact that 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 cognitive measures in our sample and allowed us to evaluate the impact of factors such as age, abstract reasoning, familial risk for dyslexia, gender, language(s)/dialects spoken at home, and handedness more exhaustively than has been done in previous literacy digital game-based learning 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 much weaker beginning readers in Belgium showed overall more intervention effects (in letter knowledge and reading fluency), whereas for 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 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). 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 of early readers who already scored high at pre-test in timed letter knowledge. The few studies who trained entire classrooms (e.g. Jere-Folotiya et al., 2014; Koikkalainen 2015; Ronimus & Lyytinen, 2015) did unfortunately not consider interaction terms in their analyses, thus providing 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 actually 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 to control for confounding in analyses of reading interventions.

For effects relating to **familial risk** of dyslexia, we found that at-risk children had slightly lower phonological skills, but the training effectiveness was not influenced by status of familial risk. The former is somewhat surprising, given that other studies also reported weaker performance in other reading precursors for children at familial risk (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. Whereas 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 risk children did not improve as much as their peers. The author's concluded that familial risk is characterized by a letter-speech-sound association and integration deficit, which the data from the present study does not support. Both studies had shortcomings preventing the authors from drawing firm conclusions about the effects of familial risk on GraphoGame effectiveness which merit mentioning. Including as few as 32 children (14 risk, 18 no risk) across two experimental groups, the study by Brem and colleagues (2010) may have suffered from a lack of power. In addition, playing GraphoGame followed by a math control game or vice versa in a crossover design, the children spent systematically less time on the second game. Blomert and Willems (2010) suggested that the absence of improvements in timed letter knowledge, phonological awareness and reading skill in the risk children in their study might have been due to the young (preschool) age of these (familial risk) children being exposed to reading materials that were too difficult. The fact that the present study did not find any distinct training effects attributable to status of 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 grade relatives.

An interesting insight from **gender** effects in game-based learning in general is that previous gaming experience may predict in-game achievement, which puts girls at a disadvantage (Nietfeld, Shores, & Hoffmann, 2014). Ideally, studies should therefore control for gender 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). While creators of game-based learning tools should aim to build gender neutral and inclusive games, considering that developmental dyslexia is diagnosed 1.5 to three times more often in boys than in girls (Rutter et al., 2004), a slight male preference for game-based learning might actually be an asset. In our sample, boys had significantly poorer letter knowledge and phonological awareness skills compared to girls at the onset 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; Torppa, Eklund, van Bergen, & Lyytinen, 2015). We also found that the observed benefits in terms of reading fluency when phonological awareness and game exposure were high was mostly carried by girls. Thus, at the group level, the boys and girls in our sample were in slightly different stages of reading acquisition. In sum, we feel that gender differences warrant further scrutiny in literacy digital game-based learning research, also given that males generally play more games, show a stronger preference for game-based learning and are more open towards technology and computers than their female peers (Admiraal, Huizenga, Heemskerk, Kuiper, Volman, & ten Dam, 2013; Gwee, San Chee, & Tan, 2011; Bonanno & Kommers, 2007).

Research Question 3: Intervention properties

Finally, we asked how intervention properties contribute to training effectiveness and we hypothesized that characteristics from the gaming process itself might help explain variance in the intervention outcome. Our study provides only limited insights in this regard. Previous literacy digital game-based learning studies using GraphoGame usually relied on the number of gaming sessions or the time spent playing as a measure of training intensity, and only few communicate treatment fidelity measures such as attrition rates, which can be as high as 46% (Jere-Folotiya et al., 2014). Studies reporting positive 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). 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 remains an open empirical question. Furthermore, the ideal training duration and intensity may differ depending on population properties and training goals, which raises the obligation to investigate possible interactions of training and population properties.

We therefore extracted additional game-exposure measures, such as the highest level that was reached, 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 speed and accuracy of children: responding faster will yield more levels, responses and seen items, while being less accurate results in being exposed to less items during the same period, due to the game's adaptivity. However, individual response patterns do also vary over time depending on the complexity (simpler, more familiar content vs. more complex new information) of consecutive levels (Nja, 2019). Individually, these additional measures did not seem to be related to response to intervention in the present study, but rather reflect pre-test skills. This confirms that data extracted from in-game behaviour can be used as for dynamic assessment (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 interactions with exposure. Upon re-inclusion of children who played less than 20 sessions and by fusing these exposure measures with a principal component analysis, we found that learning opportunity and phonological awareness modulated reading fluency when other reading pre-cursors were kept stable. Therefore, the time-course of development of phonological skills plays a crucial role for the benefits of our intervention. Playing beyond mastery of grapheme-phoneme correspondences has little impact on reading fluency when phonological skills are poor, and we did not find evidence that the current game promotes phonological skills at all. This is problematic, as combined letter-sound training and phonological awareness training were found to be more successful in boosting reading and spelling skills than either of them in isolation (Schneider, Roth, & Ennemoser, 2000). We therefore suggest reducing the weekly playing intensity once letter knowledge accuracy reaches ceiling, and instead extend 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. Future studies should furthermore focus on identifying how to best train phonological skills in literacy digital game-based learning interventions.

In addition to ongoing discussions evolving around cross-linguistic differences (Landerl et al., 2013; Moll et al., 2014; Vaessen et al., 2010) which influence reading development trajectories and remediation efforts, the field should also pay attention to differences relating to the conception of interventions as subsequent decisions for game content and parameters might also explain the wide range of outcomes seen in literacy DBGL studies. Taking this issue even further, in modern adaptive games the difficulty level is constantly adjusted to the individual learner, so different children will be exposed to different types of content (making comparisons difficult, even within the same study). This makes qualitative comparisons of games, training content and training properties very important. While such an analysis is beyond the scope of this study, it is further hampered because usually games and their underlying intentions, decisions, settings and materials are not openly shared. We therefore believe it is crucial to share detailed information on game-design (see Appendix 1 for a detailed description of the games used in this research) to allow future research to achieve better intervention.

Limitations

As with all studies, we acknowledge several limitations in the design and procedure, which should be considered when interpreting the results and analyses presented above. First of all, the unexpectedly large pre-test differences forced us to split our sample up by country, which led to smaller groups and reduced power compared to the study we initially conceived. 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 results with other cohorts of Dutch and Flemish children. An additional weakness is that we only measured reading fluency at post-test. 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 pre-test. As a result, we could not directly test interactions between reading fluency improvement and other factors, but by controlling reading fluency outcome for reading precursors at pre-test (letter knowledge, phonological awareness, rapid automatized naming and age) we are still convinced that our results are robust and meaningful. Another issue arises from the fact that the teachers who participated were favourable, or at least open, towards the use of digital tools in their classrooms, and were furthermore not blinded to the experimental conditions, and thus knew their treatment allocation. This may have changed their teaching style in one way or another, which is something that is hard to control for or correct. To balance out the impact single classrooms may have on intervention effects, children should ideally be randomized individually, i.e. 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. Another alternative could be 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, through ERP data collected from a subset of the children in the present sample it became apparent that playing the math game may have actually also contributed to the development of phonological awareness skills (Glatz, 2018). Ultimately, both games promoted careful listening and fast access to phonological representations. As arithmetic representations are indeed also phonological in nature (De Smedt & Boets, 2010; De Smedt, Taylor, Archibald, & Ansari, 2010), the mathematics game may not have been the ideal active control condition for the present research. Future research on computerized literacy training should therefore try to make use of an active control condition where the improvements of video gaming can be expected in the visual or motor domain (like described by Green & Bavelier, 2003) rather than in verbal and/or auditory learning.

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 a robust increase in mastery of grapheme-phoneme correspondences and to small to medium sized effects in reading fluency. Biographical characteristics such as familial risk of dyslexia or languages/dialects spoken at home had little impact on response to intervention and additional research investigating larger groups of children at familial risk of dyslexia is needed. Follow up studies will need to evaluate the longer-term effects of such a brief computer-assisted literacy training in first graders learning to read the semi-transparent Dutch orthography. It is unclear whether our findings are generalizable to more opaque (e.g. English) or more transparent orthographies (e.g. Finnish and Greek). Studies employing GraphoGame in Dutch are ongoing, with a focus on struggling readers and an exploration of new learning materials and tasks.

Acknowledgements

We are very grateful to all the children, parents, and teachers who participated in our research. We are indebted to Iivo Kapanen for helping to bring GraphoGame-NL to life, Sabien van Dyke and Vanessa Janssens for test administration in Belgium, and Anastasia Glushko for feedback on early drafts of this manuscript.

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.

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

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub.

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.

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.

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., 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., 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).

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.

De Freitas, S. (2006). Learning in immersive worlds: A review of game-based learning.

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.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423*(6939), 534.

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.

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.

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.

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

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *The Internet and higher education, 8*(1), 13-24.

Koikkalainen, M. (2015). Computerized 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.

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.

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.

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.

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.

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.

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

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

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition, 101*(2), 385-413.

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

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

Prensky, M. (2001). Fun, play and games: What makes games engaging. *Digital game-based learning, 5*(1), 5-31.

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

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 Computerized Intervention on the Reading Development of Beginning Readers at Risk for Reading Difficulties: A Randomized Controlled Trial. *Dyslexia, 19*(4), 256-280.

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.

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.

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

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.

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.

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.

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.

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

Susi, T., Johannesson, M., & Backlund, P. (2007). Serious games: An overview.

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.

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

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.

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.

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

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.

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

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.

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.

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

Figure 1

CONSORT flow diagram of the GG-NL randomized control trial.

Consolidated Standards of Reporting Trials (CONSORT) flow diagram showing the randomized GG-NL control trial at the subject level.

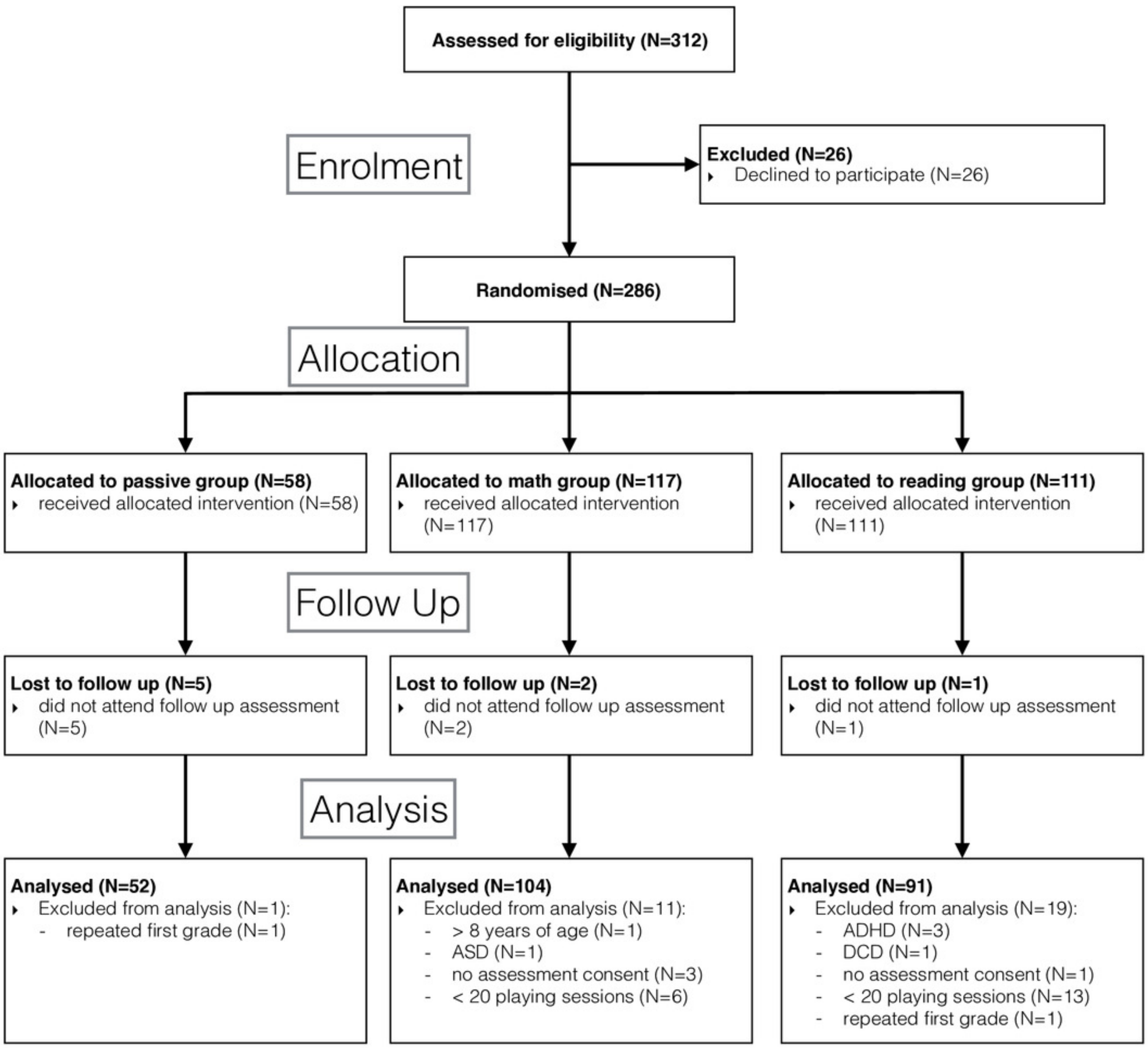
Assessed for eligibility (N=312)
Enrolment
Excluded (N=26)
Declined to participate (N=26)
Randomised (N=286)
Allocation
Allocated to passive group (N=58)
received allocated intervention (N=58)
Allocated to math group (N=117)
received allocated intervention (N=117)
Allocated to reading group (N=111)
received allocated intervention (N=111)
Follow Up
Lost to follow up (N=5)
did not attend follow up assessment (N=5)
Lost to follow up (N=2)
did not attend follow up assessment (N=2)
Lost to follow up (N=1)
did not attend follow up assessment (N=1)
Analysis
Analysed (N=52)
Excluded from analysis (N=1):
- repeated first grade (N=1)
Analysed (N=104)
Excluded from analysis (N=11):
- > 8 years of age (N=1)
- ASD (N=1)
- no assessment consent (N=3)
- < 20 playing sessions (N=6)
Analysed (N=91)
Excluded from analysis (N=19):
- ADHD (N=3)
- DCD (N=1)
- no assessment consent (N=1)
- < 20 playing sessions (N=13)
- repeated first grade (N=1)

Figure 2

Reading fluency of the Belgian sample at T2.

Model predicted z-transformed reading fluency scores of the Belgian sample at T2. Whiskers represent 95% confidence intervals.

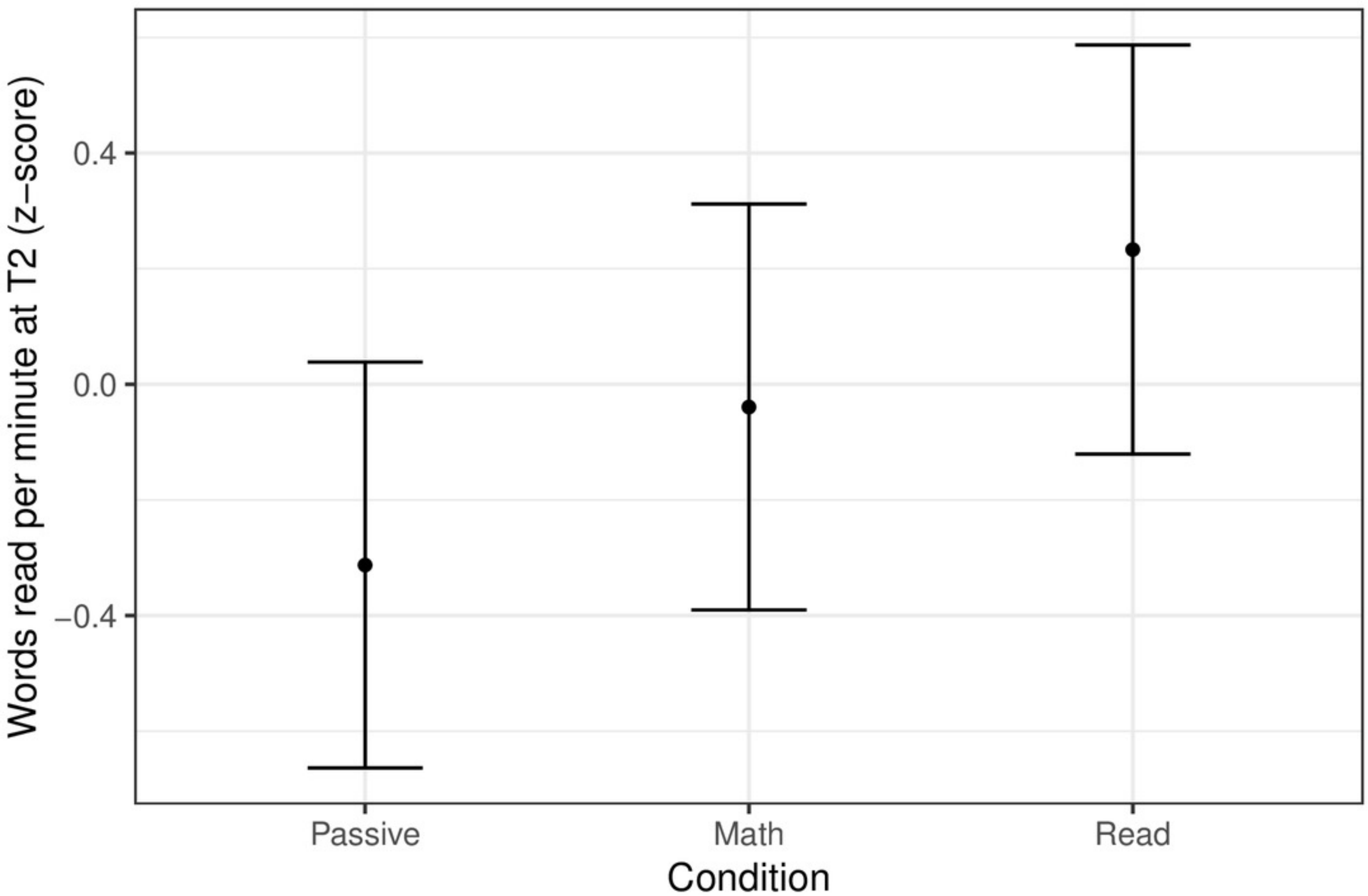

Figure 3

LSSI accuracy scores of the Belgian sample across sessions.

Model predicted accuracy for the in-game letter speech sound identification task of the Belgian sample. Whiskers represent 95% confidence intervals.

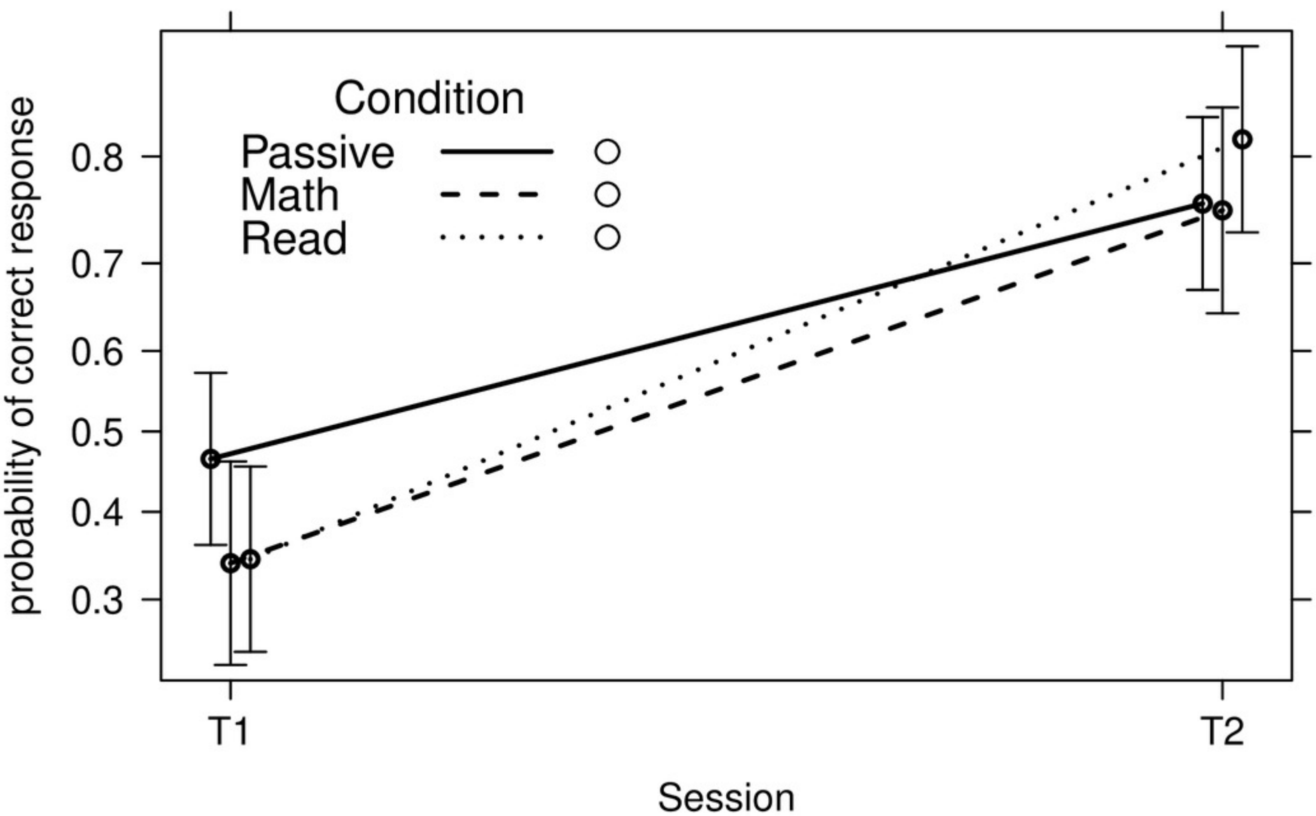

Figure 4

LSSI response times of the Dutch sample across sessions.

Model predicted letter speech sound identification response time of the Dutch sample. Whiskers represent 95% confidence intervals.

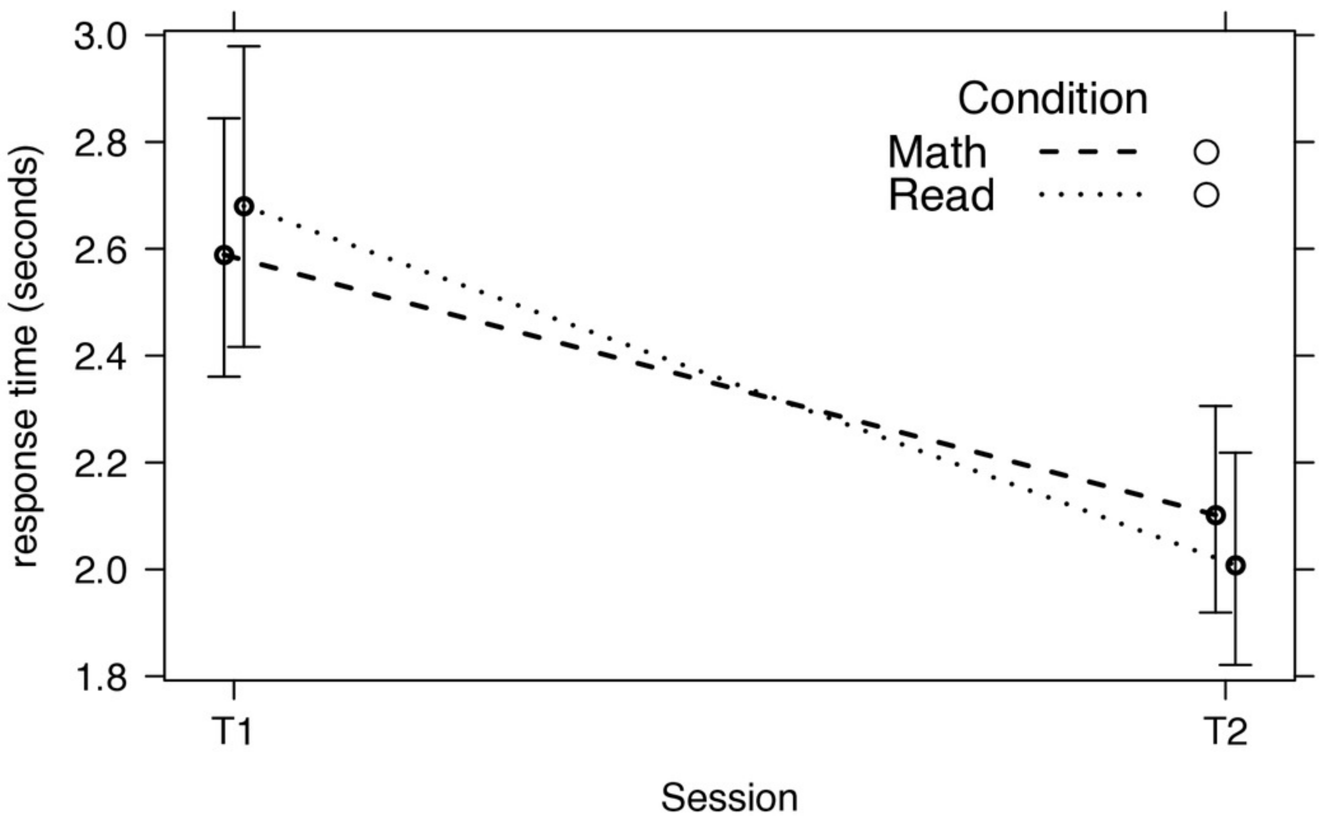

Figure 5

Reading fluency of the two experimental groups after combining both countries at T2.

Model predicted T2 reading fluency composite score (colour coded) of the two experimental groups after combining both countries using a game exposure × CELF PA at T1 interaction. All variables are z-transformed. Dots indicate available observations. (A) Math group. (B) Reading group. (C) Difference of both games. (D) Males within reading group. (E) Females within reading group. (F) Difference of both genders within the reading group.

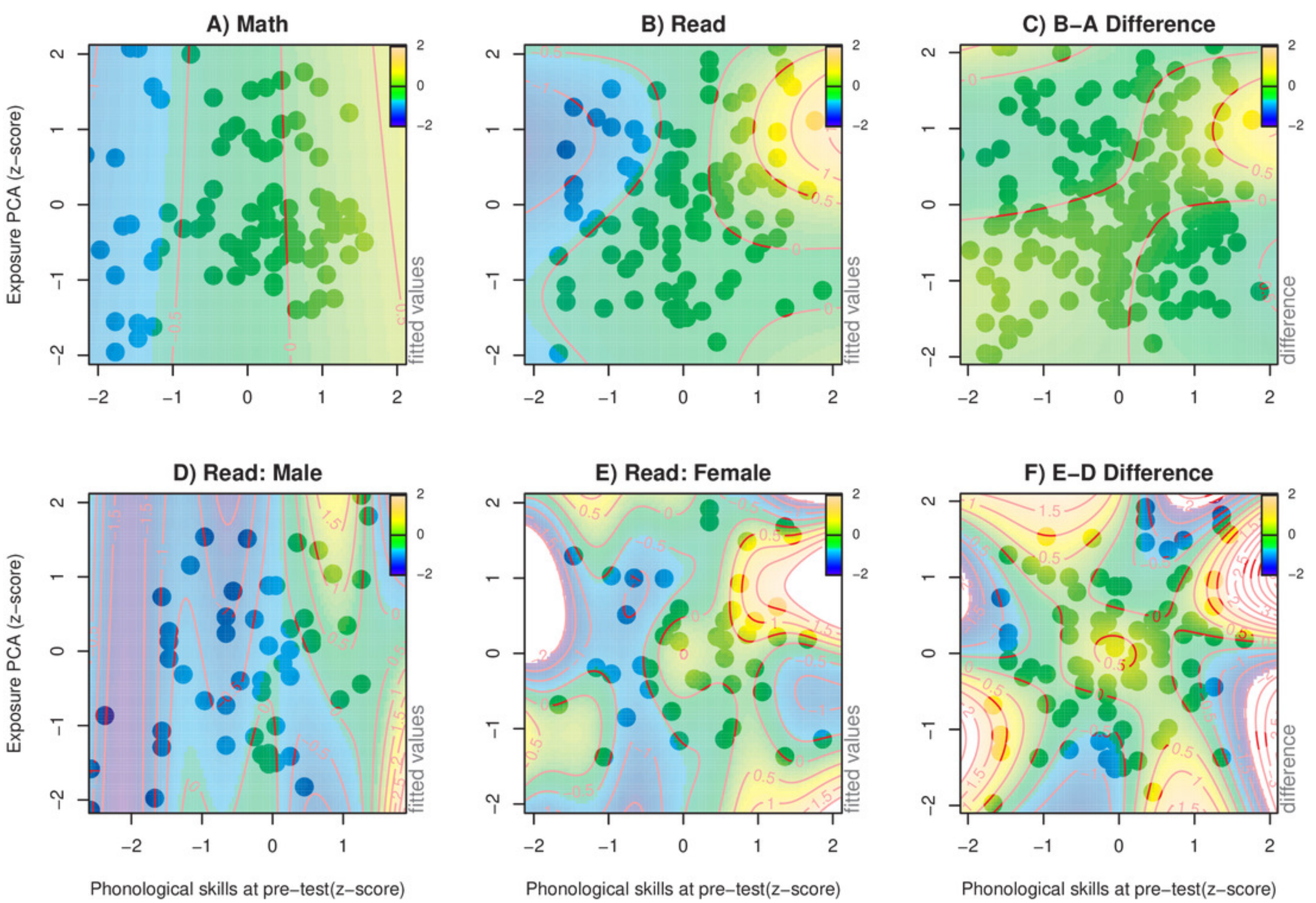

Table 1(on next page)

Classroom assignment to experimental conditions.

Classroom assignment to experimental conditions and number of children per classroom. Numbers represent count data of children who were analyzed / total number children per classroom, school and country.

School	Country	Condition		
		Passive	**Math**	**Read**
A: 33 / 46		-	19 / 24	14 / 22
B: 24 / 28	**Netherlands** 86 / 107	-	-	24 / 28
C: 29 / 33		-	29 / 33	-
D: 64 / 71		22 / 24	22 / 24	20 / 23
E: 18 / 47		6 / 11	8 / 18	4 / 18
F: 47 / 49	**Belgium** 161 / 205	15 / 17	16 / 16	16 / 16
G: 23 / 28		-	10 / 14	13 / 14
H: 9 / 10		9 / 10	-	-

Table 2 (on next page)

Participant characteristics

Descriptive statistics of the three experimental groups split up country at T1 and T2 (N=247). Values represent counts or means (standard deviations).

	Passive		Math				Read			
Country	B		NL		B		NL		B	
N	52		48		56		38		53	
Gender (f / m)	21 / 31		23 / 25		21 / 35		23 / 15		24 / 29	
Age (years)	6.20 *(0.30)*		6.25 *(0.32)*		6.31 *(0.31)*		6.19 *(0.29)*		6.26 *(0.37)*	
Familial risk (yes / no)	7 / 45		11 / 37		7 / 49		6 / 32		10 / 43	
Handedness (l / r)	9 / 43		3 / 45		5 / 51		6 / 32		4 / 49	
Monolingual (y / n)	46 / 6		36 / 12[Δ]		53 / 3		36 / 2[Δ]		51 / 2	
Abstract reasoning[†] (z-score)	-0.29 *(0.99)*		0.41 *(0.81)*		-0.15 *(1.02)*		0.48 *(1.07)*		-0.06 *(0.88)*	
Session	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**
Letter knowledge[†]	11.27[Δ] *(5.30)*	21.32 *(4.92)*	24.50 *(4.74)*	28.46 *(3.45)*	9.36 *(6.42)*	20.78 *(5.97)*	22.24 *(6.36)*	25.26 *(6.11)*	8.23[Δ] *(6.70)*	22.58 *(6.84)*
CELF PA[†] (percentile)	41.14 *(23.91)*	50.50 *(24.37)*	59.85 *(19.35)*	72.50 *(19.26)*	38.22 *(23.68)*	55.71 *(25.32)*	63.68 *(22.22)*	73.46 *(17.63)*	35.74 *(20.05)*	50.87 *(20.49)*
PROEF PA[†] (percentile)	49.52 *(25.94)*	55.91 *(29.02)*	64.84 *(26.55)*	73.80 *(19.63)*	45.67 *(29.28)*	61.03 *(26.92)*	72.50 *(22.80)*	75.92 *(22.75)*	44.29 *(25.37)*	48.73 *(24.33)*
RAN colours[†] (seconds)	70.88 *(15.92)*	57.37 *(12.57)*	58.44 *(11.80)*	52.46 *(10.17)*	67.50[Δ] *(15.37)*	55.88 *(11.42)*	59.39 *(14.71)*	53.11 *(11.87)*	74.96[Δ] *(16.02)*	59.40 *(14.01)*
RAN objects[†] (seconds)	75.56 *(17.49)*	68.08 *(19.37)*	70.46 *(15.11)*	66.54 *(17.76)*	76.25 *(17.33)*	69.93 *(21.50)*	65.68 *(12.13)*	63.67 *(15.70)*	81.77 *(23.82)*	73.55 *(22.18)*
Word reading (words per minute)		10.05 *(4.59)*		23.78 *(11.59)*		12.02 *(7.79)*		19.30 *(10.19)*		12.62 *(7.43)*

Note: NL: Netherlands; B: Belgium; PA: phonological awareness; RAN: rapid automatized naming; T1: pre-test; T2: post-test; [†]significant difference between countries at T1 ($p < .05$); [Δ]significant difference between conditions within countries at T1 ($p < .05$).

Table 3 (on next page)

Overview of fitted models and results

Overview of the 16 models that were fitted to answer the research questions. For each of the seven outcome measures two models were fitted: one for each country and then an additional two models combining the two countries. Most models describe null results, so the results section focusses on those models that show an effect of condition.

Outcome variable at T2	Country	N	R^2	Effect of condition	Included relevant co-variates (according to AIC)						
					Age	Gender	FR	T1 LK	T1 PA	T1 RAN	T1 IQ
Reading fluency		150	0.39	Read > Passive				✓	✓	✓	
CELF PA		152	0.64	n.s.					✓		✓
PROEF PA		150	0.46	n.s.			✓		✓		
LSSI accuracy	B	104	0.47	Read > Math > Passive					✓		
LSSI speed		108	NA	n.s.				✓		✓	
WLD accuracy		104	0.16	n.s.					✓		
WLD speed		103	0.42	n.s.	✓			✓	✓		
Reading fluency		78	0.43	n.s.				✓	✓		
CELF PA		83	0.45	n.s.				✓	✓		
PROEF PA		81	0.52	n.s.					✓		
LSSI accuracy	NL	75	0.81	n.s.		✓			✓		
LSSI speed		75	NA	Read > Math	✓				✓		
WLD accuracy		75	0.65	n.s.							
WLD speed		75	0.54	n.s.				✓	✓		
Reading fluency	NL + B	196	0.49	Read > Math					✓	✓	
Reading fluency		98	0.80	NA (only read)		Females > Males			✓		

Note: NL: Netherlands, B: Belgium, PA: phonological awareness, LSSI: letter speech sound identification, WLD: written lexical decision, LK: letter knowledge, RAN: rapid automatized naming, IQ: abstract reasoning. NA: not available (R^2 is not computable for some LSSI response time models due to presence of random slopes), T1: pre-test, T2: post-test.

Table 4(on next page)

Participant characteristics at T1 for analysis 3.

Descriptive statistics of the two experimental groups at T1 after combining both countries (N=210). Values represent counts or means (standard deviations).

	Math (N = 106)	Read (N = 104)
Country (NL / B)	46 / 60	41 / 63
Gender (f / m)	41 / 65	50 / 54
Monolingual (y / n)	91 / 15	93 / 11
Handedness (r / l)	100 / 6	97 / 7
FR (y / n)	18 / 88	17 / 87
Age (years)	6.28 *(0.30)*	6.25 *(0.34)*
Letter knowledge	16.00 *(9.24)*	14.12 *(9.33)*
CELF PA (percentile)	47.26 *(24.96)*	44.21 *(25.62)*
PROEF PA (percentile)	52.95 *(29.14)*	53.96 *(28.28)*
RAN objects (seconds)	74.77 *(17.36)*	76.27 *(21.21)*
RAN colors (seconds)	64.41 *(14.86)*	68.71 *(16.76)*
Abstract reasoning (*z*-score)	0.02 *(0.97)*	-0.02 *(1.03)*
Sessions played	26.20 *(4.61)*	26.14 *(6.43)*
Hours played	3.30 *(0.93)*	3.32 *(1.01)*
Levels played$^{\Delta}$	393 *(156)*	220 *(69)*
Maximum level reached$^{\Delta}$	134 *(30)*	117 *(58)*
Items seen$^{\Delta}$	8642 *(4213)*	10188 *(4236)*
Responses given$^{\Delta}$	3716 *(2059)*	2472 *(910)*

Note. NL: Netherlands, B: Belgium, FR: familial risk for dyslexia, PA: phonological awareness, RAN: rapid automatized naming. $^{\Delta}$significant difference between groups at $p < .05$.

1