Learning and Instruction 68 (2020) 101348

Contents lists available at ScienceDirect

Learning and Instruction

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

Can children's instructional gameplay activity be used as a predictive indicator of reading skills?

Jenny M. Thomson[a,c,*], Njål Foldnes[b,c], Per Henning Uppstad[c], Morten Njå[c], Oddny Judith Solheim[c], Kjersti Lundetræ[c]

[a] Human Communication Sciences, University of Sheffield, UK
[b] BI Norwegian Business School, Norway
[c] Norwegian Reading Centre, University of Stavanger, Norway

A B S T R A C T

For children who may face reading difficulties, early intervention is a societal priority. However, early intervention requires early detection. While much research has approached the issue of identification through measuring component skills at single timepoints, an alternative is the utilisation of dynamic assessment. To this point, few initiatives have explored the potential for identification through progress data from play in digital literacy games. This study explored how well growth curves from progress data in a digital intervention can predict reading performance after gameplay compared to measuring component skills at a single timepoint (school entry). 137 six-year-old students played the digital Graphogame for 25 weeks. Latent growth curve analyses showed that variation in trajectories explained variation in literacy performance to a greater extent than risk status at school entry. Findings point to a potential for non-intrusive reading assessment in the application of a serious digital game in first grade.

1. Rationale

Learning to read is one of the most important skills children will acquire in the early years of school and difficulties in acquiring this skill can have adverse educational outcomes (McLaughlin, Speirs, & Shenassa, 2014), vocational outcomes (McLaughlin et al., 2014; OECD, 2013) as well as a negative impact on both physical and mental health (DeWalt, Berkman, Sheridan, Lohr, & Pignone, 2004). Thus, for children who may face reading difficulties, early intervention is a societal priority. However, early intervention requires early detection. The importance of acting early is indicated by research showing the positive effects of early literacy interventions (Catts, Nielsen, Bridges, Liu, & Bontempo, 2015; Dion, Brodeur, Gosselin, Campeau, & Fuchs, 2010; Solheim, Fritjers, Lundetræ, & Uppstad, 2018).

Despite ever-increasing knowledge in the field of reading assessment, early detection of difficulty remains an error-prone process. Successful reading is dependent on the integrity of a number of different perceptual, cognitive and linguistic skills (Pennington et al., 2012) and so, typically, assessments that aim to identify the risk, or overt manifestation of a reading difficulty need to measure a number of component skills, including phonological awareness, letter knowledge or word decoding, verbal short-term memory, rapid automatised naming and oral language (Pennington & Lefly, 2001; Thompson et al., 2015). While interdependent, each of these skills will have a specific developmental timeline, potentially with uneven rates of change – letter knowledge, for example is an assessment measure that may only be fully sensitive to individual differences during the first few months of a child's literacy instruction. However, within that optimal time window measurements of letter knowledge may allow for strong prediction of word reading ability in subsequent school years (Georgiou, Torppa, Manolitsis, Lyytinen, & Parrila, 2012; Puolakanaho et al., 2008). This example also points to the fact that the relative predictive ability of certain measures in relation to others may change over time (Solheim, Torppa, Uppstad, & Lerkkanen, 2020). For these reasons, while our ability to predict children's risk of reading failure is arguably stronger than it has ever been, relying on measures recorded at a single point in time, to characterize a dynamic and constantly changing skill can still result in over- or under-identification of risk (see e.g. Speece, 2005).

Partially in response to this challenge of accurate detection, identification of specific reading disabilities in schools has moved towards a model in which detection of a difficulty is defined not in terms of assessments carried out at a single time point, but rather, in terms of an individual's response to intervention, or “RTI” (Gersten, 2009; IDEA, 2004). Considering a specific reading disability such as developmental dyslexia, a disability of neurobiological origin characterised by impairments in decoding, word reading accuracy and fluency (Lyon, Shaywitz, & Shaywitz, 2003), the definition provided by the DSM-5

* Corresponding author. Division of Human Communication Sciences, Health Sciences School, University of Sheffield, 362 Mushroom Lane, S10 2TS, Sheffield, UK.
E-mail address: j.m.thomson@sheffield.ac.uk (J.M. Thomson).

https://doi.org/10.1016/j.learninstruc.2020.101348
Received 1 November 2019; Received in revised form 28 April 2020; Accepted 30 April 2020
Available online 22 May 2020

(American Psychiatric Association, 2013) reiterates the need to consider a response to intervention in identification, stating that a diagnosis of dyslexia can only be made if difficulties have persisted for at least 6 months despite the provision of extra help or targeted instruction. This approach to identification of difficulties thus takes into account dynamic assessment data from multiple time points.

While research that has evaluated the effectiveness of this approach is generally supportive (Gersten et al., 2009; Gersten, Newman-Gonchar, Haymond, & Dimino, 2017) it is very dependent upon both the nature of the intervention used as well as reliable assessment measures that have the ability to sensitively and specifically capture the development in reading abilities brought about (Gersten et al., 2017). One way to increase the degree of alignment between intervention and assessment is to collect and analyse progress data from within the intervention itself in order to more directly observe response to intervention. Digital interventions arguably make within-activity progress data easier to automatically capture and researchers have started to successfully exploit this approach within education (Shute, Leighton, Jang, & Chu, 2016; Shute, Wang, Greiff, Zhao, & Moore, 2016); however, this approach has only to a small extent been implemented within literacy instruction. In this study we capitalised upon the progress data generated by a digital reading intervention called Graphogame to explore this methodology further. The novelty of the current study is validated by a recent review of the literature on this specific game showing that no existing study had taken advantage of the available progress data (McTigue, Solheim, Zimmer, & Uppstad, 2020).

GraphoGame is a play-like, internet based learning platform that provides children with training in phoneme awareness, letter-sound and early word decoding training. It was originally devised by researchers at the University of Jyväskylä in Finland with the aim of free delivery to the end user (Lyytinen, Erskine, Kujala, Ojanen, & Richardson, 2009; Lyytinen, Ronimus, Alanko, Poikkeus, & Taanila, 2007). Since its inception in Finland and promising initial findings, the game has subsequently been adapted for at least 10 alphabetic languages of varying orthographic depth, across more than 20 countries in four continents (Africa, Europe, North America, South America). The flexibility of the web-based platform means that while the basic game content remains constant across languages, researchers from countries adapting the platform can work with the Finnish developers to determine the educational/linguistic progression through letters, syllables, and words, as well as the level of challenge and adaptation.

The content adapts to the individual player according to actual performance in identifying letters, syllables or words matching auditory stimuli played through headphones. The adaptation algorithm of the game ensures a consistent balance in trials between challenge and mastery, based on the individual player's previous performance. At a certain proficiency level the algorithm provides timed target items and distractors, pushing the player to faster identification. Thus, during the course of game play, a child has the opportunity to progress to more difficult items, if and when, they demonstrate mastery of more foundational content. Graphogame is one of the minority of computerised reading interventions that has an emerging evidence-base exploring its efficacy (McTigue et al., 2020).

2. Study objectives

The intent of this study was to take advantage of the extensive progress data a digital intervention can provide and look at the utility of process data to predict future reading performance.

This study was carried out as part of the larger 'On Track' study (n = 1199), which investigated the effects of early intervention for children at risk for reading difficulty (Lundetræ, Solheim, Schwippert, & Uppstad, 2017). The study was located in Norway, where children start school when they are six years old and assessment for reading difficulties typically occurs at the end of the first year of schooling. In order to try and reduce the incidence of reading difficulties, the aims of the 'On Track' project were to develop screening tools to detect reading difficulty risk at an earlier stage - at school entry - as well as measure the effects of reading interventions carried out in the first year of schooling. Details of the screening measures, which included traditional predictors of reading risk such as letter knowledge, rapid automatized naming and phonological awareness, are provided in the Methods section below. Children's performance on the screening test was used to create an overall risk index, which was used within the current study as a variable with which to compare to game-play progress.

Game-play progress itself was captured from digital log data, obtained for a mixed-ability group of 137 six-year-old children playing Graphogame regularly over a 25 week period. The children played the game during their regular classroom literacy time at a similar level of frequency and intensity for all children in the class, thus it could be seen as equivalent to a Tier 1 intervention. It was predicted that initial risk status could explain variation in children's game progress trajectories. This study sought first, to validate this prediction, but then secondly and more crucially, to explore whether initial risk status or game progress data was the more accurate predictor of literacy skills measured at the end of the first year of schooling.

As part of the validation, the 'On Track' sample also allowed us to look at the explanatory power of other learner characteristics that could influence game progress trajectories: Firstly, the sample included a proportion of children who had parents who did not speak a Scandinavian language at home (Norwegian, Swedish or Danish), for whom Norwegian would be a second language (L2), as opposed to a first language (L1). Across both consistent and inconsistent alphabetic orthographies (Verhoeven, 2000) research suggests that children learning to read in an L2 typically have equivalent decoding skills to their L1 peers (August & Shanahan, 2017; Lesaux, Rupp, & Siegel, 2007), though often poorer oral language (Lervåg & Aukrust, 2010; Proctor, Carlo, August, & Snow, 2005). Given that Graphogame has a primary focus upon decoding skills, we predicted that the growth curves of game progression for children with different home language backgrounds would not differ. Secondly, wider research into engagement with computer games has demonstrated gender preference factors which can negatively impact levels of engagement and motivation to play for both female serious game players (Alserri, Zin, & Wook, 2018) as well as non-serious game players (Chou & Tsai, 2007). While many of the existing studies focus upon young adults, it was felt important in this study to explore whether student gender was a significant factor in explaining variation in game progression.

Accordingly, this study asked:

1. To what extent a) children's risk status, as measured at school entry, b) second language status (SLS) and c) gender, explain variation in growth curves of game progression?
2. In comparison to risk status, as measured at school entry, how well does variation in growth curves of game progression predict literacy performance measured after gameplay?

3. Methods

3.1. Participants

The children in this study were all enrolled as part of the larger 'On Track' study (n = 1199), which investigated the effects of early intervention for children at risk for reading difficulty (Lundetræ, Solheim, Schwippert, & Uppstad, 2017). Children were on average 6.2 years old when the study began at school entry (range 5.5–6.7). The schools were a convenience sample within close traveling distance of the region, and were recruited during the spring of 2014. Their scores on the national reading tests had been close to the national mean (1.5 ± 0.1 on a scale from 1 to 3) in two of the three previous years. The On Track sample included 19 schools, whereof 17 were included in a randomized controlled trial. The two remaining schools were included in the present

study: the “On Track GraphoGame Extension”. All first grade students from the two schools (7 classrooms) were invited to participate in the study. Parental consent was given for 97.7% of the students. Children with reported hearing difficulties, as identified by parent report, were excluded from the sample. The final sample included 137 students.

3.2. Procedure

At the beginning of the study, the reading readiness skills of all children were screened during the first four weeks after school entry in Grade 1. Parents answered a questionnaire relating to demographics, home literacy environment, familial risk of RD, the student's language background, and child health. Regarding language background, information was obtained regarding the languages each parent spoke at home and whether these were Scandinavian or non-Scandinavian. For the analyses reported here, given the small overall number of children exposed to a language other than Scandinavian at home, this variable was dichotomized whereby children were divided into those where either no parents or one parent who spoke a non-Scandinavian language at home (n = 119; 86.9%) versus children where both parents spoke a non-Scandinavian language at home (n = 18; 13.1%).

Risk status for reading difficulties was determined by combining screening scores on four individually-administered, tablet-based measures of pre-literacy skills, to generate a student risk index:

3.2.1. Letter-sound knowledge

Using a 15-item multiple choice format, pre-recorded letter sounds were presented and the student was to identify the corresponding upper-case letter. The student responded by pressing one of four letters appearing on the screen. Reliability in the overall 'On Track' sample as measured by Cronbach's alpha was .85.

3.2.2. Rapid automatized naming (RAN)

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

3.2.3. Phonemic awareness (PA)

PA was measured by means of eight phoneme-isolation and eight phoneme-blending tasks. Both tasks were ordered by difficulty (easiest first) and was automatically discontinued after two subsequent errors. Phoneme-blending required the student to blend a sequence of phonemes into a word. Pre-recorded stimuli were presented at a rate of one phoneme per second: “Here you see pictures of/ri/,/rips/,/ris/, and/ring/[English: ‘ride’, ‘redcurrant’, ‘rice’, ‘ring’]. Listen carefully and press the picture that goes with:/r//i//s/“. The tester pointed at the objects shown in the pictures as they were named. Students responded by pressing one of the four pictures. Reliability was $\alpha = 0.87$.

The phoneme-isolation task required the student to isolate and pronounce the initial phoneme in words. Students responded orally, and the tester scored the response on the tablet. Reliability in the 'On Track 'sample (Cronbach's α) was 0.92.

3.2.4. Determination of risk status for reading difficulties

Children who scored below the 30th percentile in any of these tests accumulated one risk point. The children also got an additional risk point if at least two close relatives reported having reading difficulties, resulting in a risk score of 0–5. Because the study sample was selected to be representative of the typical range of ability observed within primary school classrooms, over half the sample (51.8%) exhibited no risk behaviours, with increasingly smaller groups of children exhibiting cumulative risk scores, and no child receiving a score of 5. Given the small number of children scoring four risk points (n = 5), this group was combined with those scoring three risk points, to avoid having analysis subgroups with very small sample sizes. Table 1 documents the background variables for the study participants, in terms of gender, language background and risk status at school entry.

Table 1
Sample characteristics (n = 137).

	Percentage
Gender: Female	53.3
Second language status: Children with both parents speaking a non-Scandinavian language	13.1
Reading risk status:	
0 risk points	51.8
1 risk point	21.2
2 risk points	13.1
3+ risk points	13.9

3.2.5. Graphogame intervention

Starting within the same school term, all children commenced upon a schedule of playing the early literacy serious game, Graphogame, 10 min a day, four times a week, over a 25 week period. Schools were provided with tablets, loaded with the Graphogame software by the research team. Teachers were advised to include children's Graphogame within regular classroom literacy time, and all game play was automatically logged. The version reported here is the Norwegian version of Graphogame, adapted by the researchers from both the University of Jyväskylä and the Norwegian Reading Centre, University of Stavanger. The Norwegian version of GraphoGame consists of nine mini-games with immediate feedback and a motivational reward system (each mini-game presents the same content but in different play scenarios to maintain engagement). The reward system is managed via a personal avatar, created at the very start of the game. Further details about the technical specifications of the Norwegian Graphogame are reported in Njå (2019).

To operationalize progress through the game we first segmented the 25 weeks of game play into five measuring periods of five weeks (excluding holidays). While progress could be measured at even finer gradations, e.g. daily, the decision to use five-weekly intervals, provided enough time points to enable growth curve modelling, while at the same time accommodated the intensity of the data extraction process for aspects where manual input was needed. Given the individualized and adaptive nature of play within each of these periods, children's progression through subsets of content would vary.

Game data from five evenly spaced time intervals, 5 weeks apart was extracted for the purposes of this study. Firstly basic data on the amount of game play was exported from the website Grapholearn.com – for each child this included the number of days played, number of trials played and time spent playing trials in the game. The second set of data was a manual extraction from the database server, where additional aggregated data was available. This data extraction included children's progress in terms of items known by the last play session within each time period. Within the game, subsets of items - letters, syllables or words – are incrementally added to game play in a consistent order. These subsets largely increase in difficulty over the course of game play and introduction of a new subset is contingent upon performance mastery of existing subsets. The letter content is organized in three subsets of eight letters each (total letters = 24), the syllable content is grouped into 22 subsets (total syllables = 272) whilst the word content is grouped into 90 subsets (total words = 434). For each child, items known at end of each time period were indexed in terms of content type

- letter, syllable or word – as well as subset number. For the purposes of looking at game progression the three content types are kept separate. This is due to a) the significantly different number of subsets for each content type and b) differential probabilities of receiving letter, syllable or word content due to parameters set into the original game design. For example, until a player demonstrates mastery of at least 40 percent of the letter content, they will not be exposed to syllable content. Once exposed to syllable content, the level of mastery here will influence the probability of receiving either letter content or (if doing well), word content.

In this article we are reporting the analyses that focus upon progression with word level content. The majority of children moved very quickly through the letter level of the game and so this data was deemed less informative for measuring change over time. All the analyses reported here were subsequently carried out at both syllable and word levels, yielding the same pattern of findings for both. Given that the word level of content has the largest item pool and allows us to observe the most advanced level of progress children make within the game as a whole, it was decided to focus the current analysis on word level progress. Specifically, progress is operationalized as number of 'words known' by the end of each 5 week measurement period. This variable is determined by the proportion of correct responses accrued for specific word targets presented in the game sequence, representing within-game skill mastery. To reduce skewness and kurtosis in word level scores, the raw scores were rescaled into thirteen levels of within-game reading proficiency, ranging from level 1 to level 13. We refer to these reading variables as W_1 to W_5, W_i refers to reading level at the ith wave of measurement. In addition to reading level as measured by W_1, W_2, ..., W_5 we also include in Table 2 descriptive statistics for number of hours spent playing GG in each of the five periods. We refer to these variables as T_1, T_2, ..., T_5. Table 2 indicates, as expected, that reading scores improve over time. In contrast, the time spent playing GG does not increase or decrease over time, which was also expected. The median/mean total time spent playing GG was approximately 8.5 h, while the first and third quartiles were, 7.5 and 9.5 h, respectively. The children therefore spent on average 1.5 h less time playing GG (10 min four times a week for 25 weeks totals 10 h of playing time) than expected from the instructions. Table 2 also contains statistics for the post game literacy (PGL) measure, further described in the next subsection. The excess kurtosis and skewness values reported in Table 2 suggest that the distribution of each of $W_1 - W_5$, $T_1 - T_5$ and PGL may be considered to approximately follow a normal distribution. We also tested the multivariate vector comprised by these variables with Mardia's test for multivariate kurtosis, and found no evidence for a departure from multivariate normality ($z = -0.97$, p-value = 0.33).

Pearson correlations among the longitudinal variables, in addition to At-risk, are given in Table 3. Note that time spent playing GG is overall weakly positively associated with increasing reading scores, which suggest that we should ultimately control for time spent playing GG when considering the longitudinal development of reading levels.

3.2.6. Reading assessment at the end of grade 1 (post game literacy - PGL)

At the end of grade 1, the children's word reading was assessed using a subtest from the Norwegian National assessment test. The subtest consisted of 14 items, with a time limit of 2 min. Each item consisted of a picture followed by four visually similar words, whereof one corresponded to the picture. Following a practice item, the child was asked to read the words as fast as possible and to check the word that matched the picture. E.g. a picture of a fish ('fisk' in Norwegian) followed by 'fiske', 'fikse', 'fiks' and 'fisk'. The correct stimuli was presented in a random order. Number of correct words was measured (maximum = 14).

4. Analysis

The central concern of longitudinal research revolves around the description of change over time, and to find determinants of this change. That is, we want to understand interindividual differences in intraindividual change, and latent growth curve modeling (Bollen & Curran, 2006) is well-suited for this. This approach uses latent variables (e.g, α and β_1) to account for variation within and between individuals. Time is accounted for by fixing certain parameters in the model. Variation in individual starting point is accounted for by α, whose effect on the observed scores is fixed to 1 across time. Variation in growth between individuals is accounted for by latent variables β_1 (linear growth) and β_2 (quadratic growth). The effect of β_1 and β_2 on the observed scores are fixed to values that reflect linear and quadratic growth, reflectively. To answer our research questions, we therefore fit a series of latent growth models using the package lavaan (Rosseel, 2012) in the R software environment.

Given the lack of evidence for multivariate non-normality reported in the previous section, we employed normal-theory maximum likelihood estimation for all our models, while fit statistics were based on the normal-theory based chi-square statistic.

First, in order to establish whether a linear growth trajectory is sufficient to describe the data, we fit the unconditional linear growth model, referred to as M_1, depicted in Figure 1a. In this model, the trajectories are assumed to follow a linear trend, entirely explained by the random coefficients latent intercept (α) and slope (β) variables. The second model is the unconditional quadratic growth model, referred to as M_2, and depicted in Fig. 1b. This model is an extension of the base model M_1, where an additional random coefficient β_2 is included to allow the trajectories to follow a quadratic curve. The relative model fit of M_1 and M_2 may be compared statistically with a nested chi-square test in order to establish whether M_2 is a significant improvement over M_1.

After establishing whether a linear or a quadratic form is most suitable for the observed trajectories,[1] we fit a conditional model, in which the random coefficients are predicted by the time-constant variables At-risk, gender and SLS, see Figure 2a . We refer to this model as M_3. Finally, to take into account the variation in time spent playing GG, time-varying covariates T_1, ...T_5 were embedded into model M_3, and we refer to the resulting model as M_4, see Fig. 2b. Whether M_4 fits the data substantively better than M_3 may then be decided with a nested chi-square test. In models M_3 and M_4, of primary interest are the effects of At-risk, gender and SLS on the random coefficients α and β, since these relate directly to our first research question.

To answer our second research question, we added a measurement of post-game literacy performance to the best fitting model of M_3 and

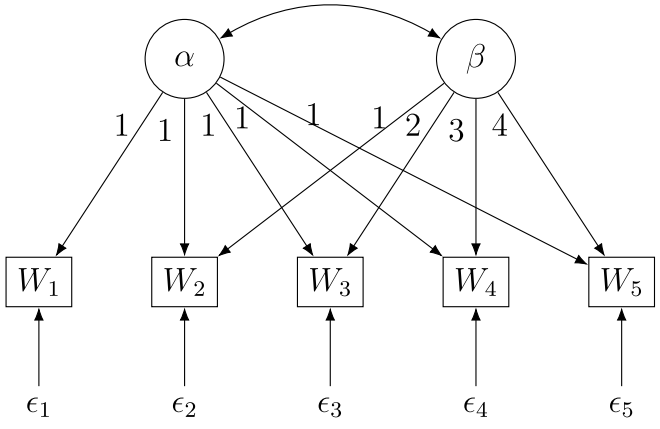
Table 2
Descriptive statistics. W_1, ..., W_5 denote the five waves of reading level measurement. T_1, ..., T_5 denote hours spent playing GG prior to the five waves of reading levels measurement. SLS = second-language status. PGL = post-game literacy.

	n	mean	sd	median	min	max	skew	kurtosis
Gender	137	0.53	0.50	1	0	1	−0.13	−2.00
SLS	137	0.87	0.34	1	0	1	−2.16	2.68
At-risk	137	0.89	1.10	0	0	3	0.85	−0.73
W_1	137	2.80	1.81	2	1	8	0.79	−0.47
W_2	137	4.39	2.38	4	1	10	0.31	−0.86
W_3	137	5.99	2.78	6	1	12	0.12	−0.74
W_3	137	7.10	2.94	7	1	13	−0.13	−0.67
W_5	137	7.67	2.91	8	1	13	−0.29	−0.53
T_1	137	1.88	0.30	1.88	1.01	2.62	−0.09	−0.11
T_3	137	1.54	0.27	1.56	0.70	2.27	−0.36	0.41
T_3	137	1.89	0.35	1.90	0.68	2.73	−0.31	0.57
T_4	137	1.80	0.46	1.74	0.45	2.84	0.03	−0.53
T_5	137	1.41	0.49	1.46	0.22	2.53	−0.18	−0.80
PGL	137	7.29	3.87	7	0	14	0.27	−1.03

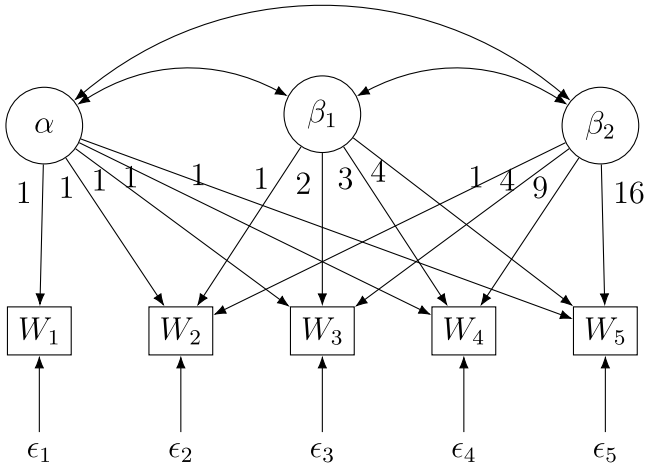
[1] Cubic trajectories were also estimated, but did not lead to improved model fit compared to the quadratic model.

Table 3
*Pearson correlations of longitudinal variables. W_1, ..., W_5 denote the five waves of reading level measurement. T_1, ..., T_5 denote hours spent playing GG prior to the five waves of reading levels measurement. *p<0.05, **p<0.01,***p<0.001.*

	At-risk	W_1	W_2	W_3	W_4	W_5	T_1	T_2	T_3	T_4	T_5
At-risk											
W_1	−0.52***										
W_2	−0.59***	0.88***									
W_3	−0.62***	0.83***	0.92***								
W_4	−0.64***	0.78***	0.86***	0.94***							
W_5	−0.62***	0.75***	0.84***	0.92***	0.96***						
T_1	−0.13	0.26**	0.19*	0.20*	0.20*	0.18*					
T_2	0.14	0.01	0.17*	0.14	0.13	0.12	0.02				
T_3	−0.11	0.17*	0.19*	0.28***	0.27**	0.29***	0.28***	0.41***			
T_4	−0.05	0.07	0.08	0.17	0.24**	0.28***	0.21*	0.32***	0.67***		
T_5	−0.12	0.10	0.08	0.17*	0.22**	0.27**	0.18*	0.23**	0.49***	0.74***	
PGL	−0.44***	0.60***	0.64***	0.68***	0.67***	0.66***	0.00	0.02	0.10	−0.02	0.00

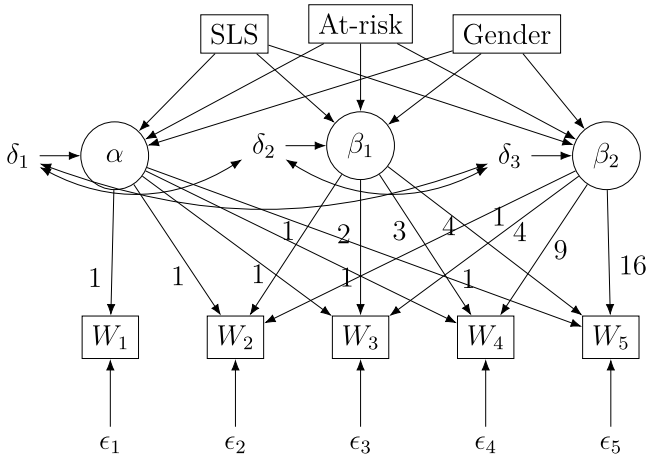


(a) M_1: Linear uncondtional model

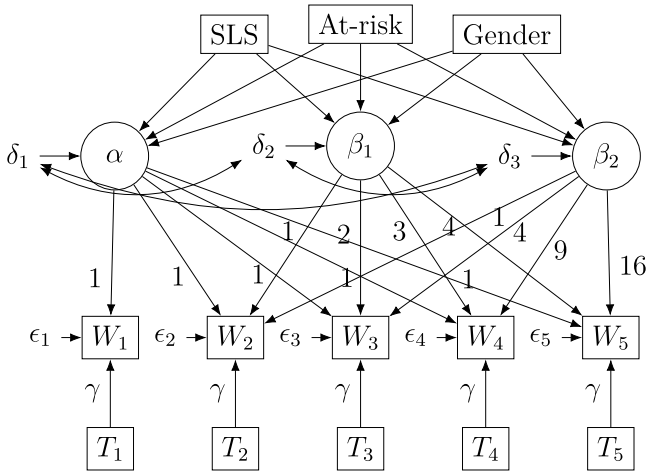


(b) M_2: Quadratic unconditional model

Fig. 1. Linear and quadratic unconditional latent growth models.



(a) M_3: Quadratic model with time-invariant predictors.



(b) M_4: Quadratic model with time-invariant predictors and time-varying covariates.

Fig. 2. Conditional quadratic growth models.

M_4, referred to as M_5. In this model the growth curve coefficients (α and β) are specified as predictors of post-game literacy performance, and of primary interest is the effects that these have on post-game literacy. A simplified path-diagram of M_5 is presented in Fig. 3.

To assess the goodness-of-fit of the sequence of models $M_1 - M_5$ in order to choose the best-fitting model, we rely on comparing fit indices like RMSEA, CFI and SRMR across models. In addition, we also took note of Akaikes Information Criterion (AIC). Formal tests of the equality constraints imposed when moving from one model to another were conducted using the chi-square test of nested models.

5. Results

Our first step was to compare the fit of the linear unconditional model M_1 to the quadratic unconditional model M_2. In Table 4 are

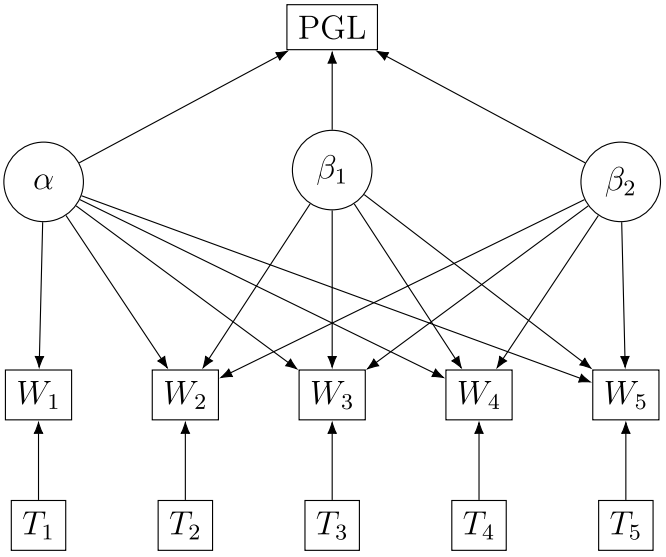


Fig. 3. Model M_5. Fixed factor loadings, residuals and covariances among them have been removed to improve clarity.

Table 4
Fit statistics associated with models $M_1 - M_5$. χ^2= test statistic of overall model fit. df = model degrees of freedom. p-value = p-value associated with χ^2. RMSEA = Root Mean Squared Error of Approximation. CFI = Comparative Fit Index. SRMR = Standardized Root Mean Square Residual. AIC = Akaike Information Criterion.

Model	χ^2	df	p-value	RMSEA	CFI	SRMR	AIC
M_1	207.43	10	0.00	0.38	0.82	0.14	2314.90
M_2	16.69	6	0.01	0.11	0.99	0.03	2132.16
M_3	20.03	8	0.01	0.10	0.99	0.02	2066.42
M_4	59.62	32	0.00	0.08	0.98	0.07	2017.53
M_5	68.43	38	0.00	0.08	0.98	0.09	2770.20

Table 5
Subset of parameter estimates for M_3. α, β_1 and β_2 refer to the intercept, linear and quadratic terms of the growth curve. At-risk = Indicates whether student was classified as at-risk for reading difficulties. Gender = 1 for females, 0 for males. SLS = Second-language status.

	Estimate	Std. Err.	z	p-value
Regression Slopes				
α				
At-risk	−0.86	0.09	−9.12	.000
Gender	−0.12	0.27	−0.44	.661
SLS	0.17	0.40	0.41	.683
β_1				
At-risk	−0.53	0.07	−7.38	.000
Gender	−0.27	0.17	−1.59	.111
SLS	0.30	0.19	1.63	.102
β_2				
At-risk	0.08	0.01	6.11	.000
Gender	0.06	0.03	1.91	.056
SLS	−0.04	0.04	−0.92	.355
Latent Intercepts				
α	3.47	0.44	7.82	.000
β_1	2.32	0.19	12.38	.000
β_2	−0.27	0.04	−6.50	.000
Latent Variances				
α	2.30	0.34	6.80	.000
β_1	0.80	0.21	3.80	.000
β_2	0.03	0.01	3.71	.000
Latent Covariances				
α w/β_1	0.20	0.20	0.99	.323
α w/β_2	−0.05	0.04	−1.49	.136
β_1 w/β_2	−0.14	0.04	−3.46	.001

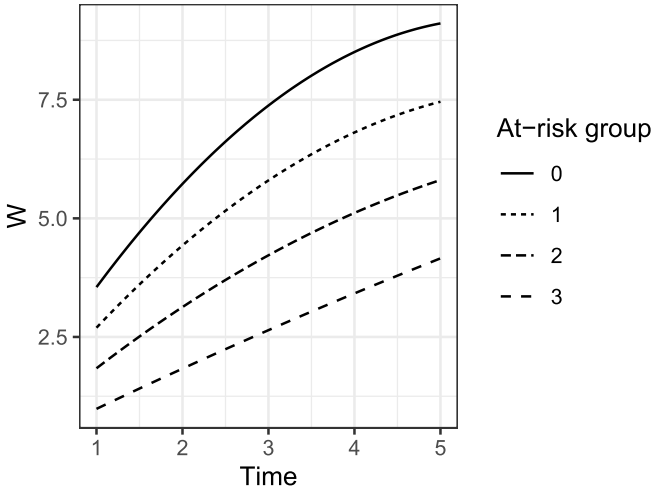


Fig. 4. Model-implied growth curves from M_3 at different levels of At-risk.

presented fit measures for all the estimated models. Clearly, the quadratic model fits the data substantially better than M_1. Also, the chi-square difference test of nested models rejected the linear model relative to the quadratic model ($\Delta\chi^2(4) = 174.44$, p-value <0.001). We therefore decided to proceed to the conditional models using a quadratic trajectory model. Next, we estimated the conditional quadratic model M_3, see Table 5 for model estimation results. Neither SLS nor gender predicts any of the random coefficients α, β_1 and β_2 at the 5% level of significance. In contrast, At-risk is a significant predictor of all three coefficients. In the following models, we therefore exclude SLS and gender as predictors, and retain At-risk, for parsimony. A formal chi-square test of nested models confirmed that this removal did not diminish model fit ($\Delta\chi^2(6) = 6.03$, p-value = 0.42). The resulting model-implied growth trajectories, one for each level of At-risk, are plotted in Fig. 4. The Figure suggests that the more at risk a student is for reading difficulties, the lower the expected trajectory starts out, and the slower the progress is expected to be.

In order to take into account the amount of game playing, we next fitted model M_4. The fit statistics in Table 4 suggest that M_4 has a fit similar to that of M_3, since two of the statistics, RMSEA and AIC, favours M_4, while the other two statistics, CFI and SRMR, favour M_3. Model estimation results for M_4 are given in Table 6. The regression coefficient γ relating T_i to W_i for $i = 1, \ldots, 5$ is highly significant and implies that an increase in playing time of 1 h during the interval between any two measurements will on average be associated with an increase in reading score of 0.65 units. Note that taking into account the effect of time spent playing GG, the effect of At-risk on the random growth coefficients does not substantively change compared to M_3.

To shed light on our second research question, we estimated model M_5. This model assesses the effect of growth trajectory on post-game literacy. The model estimates are presented in Table 7. It is seen that the form of trajectory is significantly related to post-game literacy. For instance, an increase in initial GG score of one unit will on average be associated with an increase in post-game literacy score of 1.19. Likewise, the slope and quadratic coefficients are significant predictors of post-game literacy. Importantly, the R^2 of the dependent variable *PGL* in model M_5 was 0.56. That is, the model accounts for more than half of the variation in post-game literacy. We also estimated a linear regression model with post-game literacy score as dependent variable and at-risk status as independent variable. In this model R^2 was 0.20. That is, while at-risk status can explain 20% of the variation in post-game literacy scores, the GG growth trajectory accounts for 56% of the variation in post-game literacy scores.

Table 6
Subset of parameter estimates for M_4. W_1, ..., W_5 denote the five waves of reading level measurement. T_1, ..., T_5 denote hours spent playing GG prior to the five waves of reading levels measurement. α, β_1 and β_2 refer to the intercept, linear and quadratic terms of the growth curve. At-risk = Indicates whether student was classified as at-risk for reading difficulties. Gender = 1 for females, 0 for males. SLS = Second-language status.

	Estimate	Std. Err.	z	p
Regression Slopes				
α				
At-risk	−0.83	0.09	−9.20	.000
β_1				
At-risk	−0.54	0.07	−8.24	.000
β_2				
At-risk	0.09	0.01	7.03	.000
W_i				
T_i	0.65	0.08	7.98	.000
Latent Intercepts				
α	2.32	0.25	9.35	.000
β_1	2.39	0.10	24.71	.000
β_2	−0.23	0.02	−11.51	.000
Latent Variances				
α	2.23	0.33	6.72	.000
β_1	0.79	0.19	4.08	.000
β_2	0.03	0.01	3.78	.000
Latent Covariances				
α w/β_1	0.19	0.18	1.05	.293
α w/β_2	−0.05	0.03	−1.55	.121
β_1 w/β_2	−0.14	0.04	−3.67	.000

Table 7
Subset of parameter estimates for M_5. W_1, ..., W_5 denote the five waves of reading level measurement. T_1, ..., T_5 denote hours spent playing GG prior to the five waves of reading levels measurement. α, β_1 and β_2 refer to the intercept, linear and quadratic terms of the growth curve. PGL = Post-Game Literacy.

	Estimate	Std. Err.	z	p
Regression Slopes				
W_i				
T_i	0.63	0.08	7.99	.000
PGL				
α	1.19	0.14	8.64	.000
β_1	3.80	0.30	12.55	.000
β_2	13.74	2.73	5.04	.000
Latent Intercepts				
α	1.61	0.21	7.82	.000
β_1	1.92	0.09	20.29	.000
β_2	−0.16	0.02	−8.66	.000

6. Discussion

This study set out to explore the extent to which variation in growth curves of Graphogame progression could predict literacy performance measured post-gameplay. Through the application of growth curve modelling, it was found that variation in trajectories predicted literacy performance post game-play to a much greater extent than did risk status as measured at school entry. Furthermore, as part of an initial validation of the relationship between risk status and game play progress, it was found that while neither second language status, nor gender explained significant variation in the growth curve parameters, children's risk status was a significant predictor of all three growth coefficients: the more at risk a student is for reading difficulties, as measured at school entry, the lower the expected trajectory starts out after an initial five weeks of play, and the lower the subsequent progress is expected to be.

To our knowledge, this is the first study to apply growth modelling to digital literacy instruction data, in order to better understand the trajectories of children's progress through the game, and factors that may influence this. Digital reading programmes are increasingly being used internationally to support early reading instruction for children both with and without risk for reading difficulties. Findings from the present study point to a potential for an additional use of gameplay, i.e. utilizing data on children's progress from within a game for assessment purposes. A non-intrusive assessment like this could reduce the time currently spent on assessing students and leave more time for their learning. However, more research will be needed to fully explore this possibility.

Digital learning programmes themselves are often treated like a “black box” (Latour, 1987), with attention paid to the outcomes of the play (Boyle et al., 2016; Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012), rather than how the games work and interact with users (Gaydos, 2015; Lämsä, Hämäläinen, Aro, Koskimaa, & Äyrämö, 2018; Njå, 2019). The data presented here offers an initial glimpse into the black box of a specific game. Graphogame has been designed as a preventative tool and in its original inception was designed for children at risk of dyslexia (Lyytinen et al., 2009). In the present study the children designated as most at risk of reading difficulties made the least progress within a 25 week period. Such a finding raises many questions and prompts us to actively consider what a successful response to intervention is for struggling readers. It is firstly important to acknowledge that without more fine-grained investigation, the direct relationship between game progress and generalisable literacy learning is not fully quantifiable. However, this slower progress nonethelesss provides noteworthy information with alternative interpretations available. One possibility is that this is an encouraging finding - for a group of children with demonstrated difficulties in reading-related skills, they have been able to progress through the game and reach the word level of the game which requires a certain level of mastery of both letters and syllables. Alternatively, we can ask, could game parameters in terms of e.g. the challenge level, rate of lexical progression and type of feedback be further optimised for this group. We hope that the analysis here can act as a catalyst for subsequent interrogation of the game data at a micro-level in order to yield further answers.

This is an explicit and novel example of using the ‘big data’ that serious games yield to clearly document children's response to intervention, and again, goes one step beyond approaches that rely more solely on more isolated measures of pre- and post-intervention performance. The trajectories used in the analysis spanned 25 weeks of playing time, yet scrutiny of the growth curves (see Fig. 4) suggests that group differences could potentially be determined within a shorter interval of play. A future combination of screening for literacy skills at school entry, alongside a focused period of serious game play could provide new sensitivity and specificity to the identification of reading difficulty risk, as well as providing reinforcement and practice of essential early literacy skills.

We turn now to the variables that did not predict children's growth trajectories. Regarding second language status, previous research using single time-point assessment data supports the notion that children with L2 and no other risk factors typically have equivalent decoding skills to their L1 peers in the face of potential vulnerabilities in reading comprehension e.g. (August & Shanahan, 2017). This dataset goes one step further and suggests that the trajectory of progress for L2 children as a group, through an instructional game, is also not distinguishable from an L1 child. Most crucially, this observation is in contrast to findings for L1 or L2 children with distinct risk factors, as measured by school-entry assessments of letter-sound knowledge, phonological awareness and rapid naming, as well as indicators of familial risk; as Fig. 4 shows, cumulative risk status significantly, and deleteriously impacts a child's progress trajectory (in the sample reported here, 66.67% of the L2 group were in risk groups 0–1, 16.67% were in risk group 2 and 16.67% were in risk group 3).

The study further found that while reading risk status had a significant impact on children's progress through the game, their gender did not. This is taken as a positive finding in terms of equality of

learning outcomes. No equivalent data is available looking at young children's progress through a literacy serious game, though studies of older youth have reported that boys can be more motivated to play computer games (Chou & Tsai, 2007) and what gender differences are present in the type of games that appeal to boys versus girls, with boys tending to prefer action/fighting games, with girls more drawn to social games and virtual worlds (Alserri et al., 2018). Potentially Graphogame is well-placed between these extremes and so does not necessarily play explicitly towards the playing preferences of one gender over another. Or alternatively, the playing context here, where there was not a choice of an alternative game, and periods of play were managed overall by the lesson time allocated, meant that any possible gender preferences that were present within the children did not have an opportunity to manifest in the data collected. It is also important to note that within this sample, there were no significant gender differences in post-play literacy performance, measured at the end of grade 1.

7. Limitations

One limitation of current study is the relatively small sample size (n = 137), and in addition the small proportion of children for whom both parents spoke a non-Scandinavian language (n = 18). The ability of children's at risk status to significantly predict all three growth coefficients within the current sample size potentially points to the robustness of the effect, however a further study with a larger population is warranted. Regarding the second language status of the sample, the current design was a convenience sample, with the proportion of children living in homes where both parents spoke a non-Scandinavian language equivalent to local norms. However, in order to more systematically validate the findings reported here, it would be important to try to actively recruit more children exposed to non-Scandinavian languages at home, to allow comparison of more equally sized groups. It would also be valuable to look more specifically at the role of oral language ability on game progress, across the ability spectrum.

In addition, as noted above, a challenge for any intervention research is the issue of inferring consolidated, generalisable learning from the successful completion of game activities. The variable of game progress used in this study was that of the number of 'words known' by the end of each 5 week measurement period, determined by game algorithms from the proportion of correct responses accrued for specific word targets presented in the game sequence. A further step in this work would be to see how reading performance for the same words outside of the game was impacted by within-game progress.

8. Conclusion

This study provides a first attempt to use the extensive progress data a digital intervention can provide, to predict future reading performance. The progress data reported here yielded critical new insights into the impact of reading risk status on progress through a digital literacy intervention in Grade 1. It also confirmed the predictive role of response to intervention in understanding trajectories of learning to read.

Acknowledgements

This article is part of the project, On Track, led by Oddny Judith Solheim, Per Henning Uppstad and Kjersti Lundetræ at the University of Stavanger. The project is funded by the Research Council of Norway's FINNUT research programme, grant no. 237861.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2020.101348.

References

Alserri, S. A., Zin, N. A. M., & Wook, T. S. M. T. (2018). Gender-based engagement model for serious games. *International Journal of Advanced Science, Engineering and Information Technology, 8*(4), 1350–1357.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Association.

August, D., & Shanahan, T. (2017). *Developing literacy in second-language learners: Report of the national literacy panel on language-minority children and youth.* Routledge.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective, Vol. 467*. John Wiley & Sons.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., ... Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education, 94*, 178–192.

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early identification of reading disabilities within an rti framework. *Journal of Learning Disabilities, 48*(3), 281–297.

Chou, C., & Tsai, M.-J. (2007). Gender differences in taiwan high school students' computer game playing. *Computers in Human Behavior, 23*(1), 812–824.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*(2), 661–686.

DeWalt, D. A., Berkman, N. D., Sheridan, S., Lohr, K. N., & Pignone, M. P. (2004). Literacy and health outcomes. *Journal of General Internal Medicine, 19*(12), 1228–1239.

Dion, E., Brodeur, M., Gosselin, C., Campeau, M.-È., & Fuchs, D. (2010). Implementing research-based instruction to prevent reading problems among low-income students: Is earlier better? *Learning Disabilities Research & Practice, 25*(2), 87–96.

Gaydos, M. (2015). Seriously considering design in educational games. *Educational Researcher, 44*(9), 478–483.

Georgiou, G. K., Torppa, M., Manolitsis, G., Lyytinen, H., & Parrila, R. (2012). Longitudinal predictors of reading and spelling across languages varying in orthographic consistency. *Reading and Writing, 25*(2), 321–346.

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. D. (2009). *Assisting students struggling with reading: Response to intervention and multi-tier intervention in the primary grades. A practice guide*.

Gersten, R., Jayanthi, M., & Dimino, J. (2017a). Too much, too soon? Unanswered questions from national response to intervention evaluation. *Exceptional Children, 83*(3), 244–254.

Gersten, R., Newman-Gonchar, R., Haymond, K. S., & Dimino, J. (2017b). *What is the evidence base to support reading interventions for improving student outcomes in grades 1-3? Rel 2017-271*. Regional Educational Laboratory Southeast.

Lämsä, J., Hämäläinen, R., Aro, M., Koskimaa, R., & Äyrämö, S.-M. (2018). Games for enhancing basic reading and maths skills: A systematic review of educational game design in supporting learning by people with learning disabilities. *British Journal Of Educational Technology, 49*(4), 596–607.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* Harvard university press.

Lervåg, A., & Aukrust, V. G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry, 51*(5), 612–620.

Lesaux, N. K., Rupp, A. A., & Siegel, L. S. (2007). Growth in reading skills of children from diverse linguistic backgrounds: Findings from a 5-year longitudinal study. *Journal of Educational Psychology, 99*(4), 821.

Lundetræ, Kjersti, Solheim, Oddny Judith, Schwippert, Knut, & Uppstad, Per Henning (2017). Protocol: 'On Track', a group-randomized controlled trial of an early reading intervention. *International Journal of Educational Research, 86*, 87–95. https://doi.org/10.1016/j.ijer.2017.08.011.

Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). A definition of dyslexia. *Annals of Dyslexia, 53*(1), 1–14.

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.

Lyytinen, H., Ronimus, M., Alanko, A., Poikkeus, A.-M., & Taanila, M. (2007). Early identification of dyslexia and the use of computer game-based practice to support reading acquisition. *Nordic Psychology, 59*(2), 109–126.

McLaughlin, M. J., Speirs, K. E., & Shenassa, E. D. (2014). Reading disability and adult attained education and income: Evidence from a 30-year longitudinal study of a population-based sample. *Journal of Learning Disabilities, 47*(4), 374–386.

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

Njå, M. (2019). Players' progression through graphogame, an early literacy game: Influence of game design and context of play. *Human Technology, 15*(2), 226–255.

OECD (2013). Oecd skills outlook 2013. Retrieved from https://www.oecd-ilibrary.org/content/publication/9789264204256-enhttps://doi.org/10.1787/9789264204256-en.

Pennington, B. F., & Lefly, D. L. (2001). Early reading development in children at family risk for dyslexia. *Child Development, 72*(3), 816–833.

Pennington, B. F., Santerre-Lemmon, L., Rosenberg, J., MacDonald, B., Boada, R., Friend, A., ... Olson, R. K. (2012). Individual prediction of dyslexia by single versus multiple deficit models. *Journal of Abnormal Psychology, 121*(1), 212–224.

Proctor, C. P., Carlo, M., August, D., & Snow, C. (2005). Native Spanish-speaking children reading in English: Toward a model of comprehension. *Journal of Educational*

Psychology, 97(2), 246.
Puolakanaho, A., Ahonen, T., Aro, M., Eklund, K., Leppänen, P. H., Poikkeus, A. M., ... Lyytinen, H. (2008). Developmental links of very early phonological and language skills to second grade reading outcomes: Strong to accuracy but only minor to fluency. *Journal of Learning Disabilities, 41*(4), 353–370.
Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.
Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the science of assessment. jan *Educational Assessment, 21*(1), 34–59.
Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. oct *Computers in Human Behavior, 63*, 106–117.
Solheim, O. J., Fritjers, J., Lundetræ, K., & Uppstad, P. H. (2018). Effectiveness of an early reading intervention in a semi-transparent orthography: A group randomized controlled trial. *Learning and Instruction, 58*, 65–79.
Solheim, O. J., Torppa, M., Uppstad, P. H., & Lerkkanen, M.-K. (2020). Screening for slow reading acquisition in Norway and Finland - a quest for context-spesific predictors. *Scandinavian Journal of Educational Research*. https://doi.org/10.1080/00313831.2020.1739130.
Thompson, P. A., Hulme, C., Nash, H. M., Gooch, D., Hayiou-Thomas, E., & Snowling, M. J. (2015). Developmental dyslexia: predicting individual risk. *Journal of Child Psychology and Psychiatry, 56*(9), 976–987.
Verhoeven, L. (2000). Components in early second language reading and spelling. *Scientific Studies of Reading, 4*(4), 313–330.