Human TECHNOLOGY

ISSN: 1795-6889

https://ht.csr-pub.eu Volume 21(3), December 2025, 548–575

HOW SERIOUS ARE SERIOUS GAME ASSESSMENTS OF READING AND READING-RELATED SKILLS?

Riikka Heikkilä
Niilo Mäki Institute
Finland
ORCID 0000-0001-9037-2788

Jarkko Hautala
Niilo Mäki Institute
Finland
ORCID 0000-0002-7402-6364

Vesa Rantanen
Centre for Applied Language Studies, University of Jyväskylä
Jyväskylän kenttäurheilijat (JKU)
Finland

Lea Nieminen
Centre for Applied Language Studies, University of Jyväskylä
Finland
ORCID 0000-0003-2286-7409

Maija Pocknell
Pirha - The Wellbeing Services County of Pirkanmaa
Finland

Juha-Matti Latvala
Niilo Mäki Institute
Finland
ORCID 0000-0001-9359-6517

Ulla Richardson
Centre for Applied Language Studies, University of Jyväskylä
Finland
ORCID 0000-0001-9181-5106

Abstract: ***Purpose****: This study examined the validity of serious game–based assessments (SGAs) for measuring Finnish primary school children's (Grades 1–4; n = 735) reading, spelling, and related cognitive skills.* ***Methods****: Performance in the digital SGAs was compared with corresponding paper-and-pencil tasks assessing word and pseudoword reading, sentence reading fluency, spelling, and underlying skills including phonological processing, rapid automatized naming, short-term memory, receptive vocabulary, and associative learning.* ***Results****: The SGAs showed good concurrent and construct validity for reading fluency, reading accuracy, spelling, rapid automatized naming, and vocabulary. The SGA tasks explained 72–80% of the variance in traditional reading fluency measures and 51–66% in reading accuracy.* ***Conclusions****: The findings indicate that serious game–based assessments provide a valid and engaging alternative to traditional literacy assessment tools. Implications for digital assessment and intervention in reading development are discussed.*

Keywords: *reading fluency, digital serious game assessment, assessment of reading and reading-related skills, transparent orthography, GraphoLearn.*

DOI: https://doi.org/10.14254/1795-6889.2025.21-3.4

INTRODUCTION

Around 5–15% of school-aged children face difficulties in learning to read (American Psychiatric Association, 2013). The importance of early intervention of reading difficulties is indisputable: the earlier the identification and adequate support, the less time and effort is needed to reduce the risk of later reading problems (see Fletcher, Francis, Foorman, & Schatschneider, 2021 for summary). Early identification and support to prevent learning disabilities is also at the core of the Response to Intervention (RTI) framework (Fuchs & Fuchs, 2005), a model that has also been applied in Finland where the current study was conducted (Björn, Aro, Koponen, Fuchs, & Fuchs, 2016). Though teachers in Finland have a central role in providing support for students in this three-tiered support system, they can often struggle to find time to provide support adequately to all in need, a task that also involves the completion of accompanying documentation (Eklund, Sundqvist, Lindell, & Toppinen, 2021). Crucially providing early and adequate support requires valid and practical skill assessments. Although some standardised paper-and pencil assessments exist, they require a qualified administrator and usually take up a plenty of time with one-to-one assessments and for scoring the data. What is urgently needed is a reliable and practical assessment tool applicable to teachers (Virinkoski, Lerkkanen, Holopainen, Eklund, & Aro, 2018).

One way to enable practical skill assessment is to use digital versions of standardised skill assessments. Indeed, studies investigating comparability of digital assessments to pencil and paper tasks have shown that although the differences between the two modes are evident the digital versions of the assessments can be a valid way to assess skills (e.g., Kroehne, Buerger, Hahnel, & Goldhammer, 2019). A benefit of digital assessment over traditional one-to-one assessment is that at best it can provide a cost-effective way to secure, scientifically valid, and reliable assessment data without requiring a professional administrator as it also saves time by scoring and interpreting data automatically (e.g., Wang, Jiao, Young, Brooks, & Olson, 2008). Digital assessments can also provide new kinds of methods for assessment, data analyses, and feedback at both the individual and institutional levels that have been virtually impossible via traditional tools (Bennett, 2015). Furthermore, some usability studies have also indicated that both teachers and students seem to prefer digital assessments over the traditional assessment method (e,g., Paleczek, Seifert, & Schöfl, 2021; Seifert, & Paleczek 2021).

In the present study, we took two further steps to critically examine the potential of digital skill assessment specifically in terms of reading and reading related skills in Finnish speaking primary school students. First, instead of merely digitalising existing standardised paper and pencil tasks (PP) we designed and developed tasks directly applicable into the serious game GraphoLearn[1] (e.g., Richardson & Lyytinen, 2014). In other words, our aim was to develop serious game assessment (SGA) tasks that utilised the multimodality of a digital gaming format at the same time as assessing the skills assessed in the standardised PP tasks. Second, we critically examined whether the developed tasks provided similar assessment data as in their counterpart paper and pencil tasks using different validity constructs. Thus, our aim was to investigate if and to what extent the SGA tasks provide comparable and valid data to PP tasks. To our knowledge, investigations of validity of reading and reading-related skill assessments within a serious-game framework does not exist prior to our investigation. The closest alternative to SGA are computerised assessments. Therefore, next we will provide a short

background to our study by looking into previous research evidence provided by studies on computerised assessment of reading and related skills.

Computerised Assessment of Reading and Related Skills

The majority of the studies concerning digital, computerised assessment (CA) of reading and literacy skills have investigated reading comprehension skills of high school or university students, and only a relatively small number of studies have explored CA of basic literacy skills, such as decoding and spelling in younger school-aged children (for reviews, see Blok, Oostdam, Otter, & Overmaat, 2002; Kingston, 2008; Wang et al., 2008). In addition, most of the studies have been conducted in the opaque English orthography (Carson, Boustead & Gillon, 2014; Clemens et al., 2015; DeGraff, 2005; Ma et al., 2025; Merrell & Tymms, 2007; Sainsbury & Benton, 2011; Singleton, Thomas & Horne, 2000; Yeatman et al., 2024), for which the emphasis of reading assessment is mostly on reading accuracy, reading comprehension and phonological skills. In transparent orthographies like Finnish, German, Italian, and Greek, learners obtain high reading accuracy already during first Grade (Seymour, Aro, & Erskine, 2003). What is more challenging for Finnish and other learners in transparent orthographies, however, is the development of reading fluency (Eklund, Torppa, Aro, Leppänen, & Lyytinen, 2015; Landerl & Wimmer, 2008), which highlights the need for CA methods to assess fluency as well as accuracy (Protopapas & Skaloumbakas, 2007). Some previous studies have successfully developed CA methods for reading fluency assessment for French (Auphan, Ecalle, & Magnan, 2019), Greek (Protopapas & Skaloumbakas, 2007; Protopapas, Skaloumbakas, & Bali, 2008) and Finnish (Authors, 2020; Heikkilä et al., 2023; Niskakoski, Määttä, Korpivaara, & Westerholm, 2020; Paananen et al., 2019).

To deepen our understanding of reading skill development, assessing also reading-related skills may be important. Specifically reading-related assessments can broaden the understanding of the potential cognitive underpinnings of reading difficulties (Furnes & Samuelsson, 2011; Parrila, Kirby, & McQuarrie, 2004; Vaessen & Blomert, 2010). Some recent studies have addressed the digital assessment of reading related skills in Greek (Zygouris, Vlachos, Styliaras, Tziallas, & Avramidis, 2025), but they did not iclude latency-data nor reading itself. Some currently available CA methods serve as a more comprehensive battery for assessing reading and reading-related skills, but are administered on a one-to-one basis (Jiménez, García, & Balade, 2024; Nergård-Nilssen & Friborg, 2022). While this approach allows for close observation of the child's performance, it is time-consuming and resource-intensive, and does not take full advantage of the efficiency and scalability offered by group-administered digital assessments..

As in reading, accuracy measures for many underlying skills, such as letter knowledge, phonological awareness, and spelling accuracy, are prone to the ceiling effect in transparent orthographies, which highlights the need to assess fluency in addition to accuracy in these measures (Georgiou et al., 2020; Torppa, Georgiou, Niemi, Lerkkanen, & Poikkeus, 2017; Vaessen, Gerretsen, & Blomert, 2009). Conveniently CA provides an opportunity to measure response times for individual responses, which we will utilise in this study to measure fluency for reading and spelling. Next, we will briefly review the connection between reading and underlying skills from the perspectives of reading development and orthography, and of previous studies on the CA of reading-related skills.

Poor readers typically make spelling errors, yet to a lesser degree in transparent than in more opaque orthographies (Furnes & Samuelsson, 2010). Previous studies on CA of spelling skills have shown weaker correlations between CA and PP tasks in transparent (.40-.42; Protopapas et al., 2008) than in opaque orthography (.60-.71; DeGraff, 2005). However, adding latency data may expand the accuracy of discrimination in transparent languages (Zygouris et al., 2025).

Letter knowledge skills of pre-school children have been shown to be a strong predictor of later reading and spelling skills across orthographies (e.g. Caravolas, 2004; Lervåg & Hulme, 2010; Torppa et al., 2013) but seems to lose its discriminatory power as soon as basic reading skills have been acquired (Leppänen, Niemi, Aunola, & Nurmi, 2006). CA letter knowledge, however, has previously been found to serve as a highly sensitive measure for screening later risk for reading problems measured in kindergarten (Puolakanaho & Latvala, 2017).

Phonological awareness (PA) refers to an individual's ability to understand the sound structure of a word. PA has been consistently shown to predict reading skills, especially reading accuracy (e.g. Melby-Lervåg, Lyster & Hulme, 2012) and word and pseudoword spelling (e.g. Landerl & Wimmer, 2008; Lervåg & Hulme, 2010; Torppa et al., 2013). However, the connection between PA and reading seems to vary depending on orthographic complexity and other variables (Landerl et al., 2019) and diminish as children learn to decode words accurately (e.g. de Jong & van der Leij, 1999; Juul, Poulsen, & Elbro, 2014). Phonological skills might still, however, differentiate those with reading difficulties to a small degree (Eklund et al., 2015) and predict spelling accuracy (Furnes & Samuelsson, 2010; Lervåg & Hulme, 2010). Previous studies of PA assessments have reported generally moderate or low correlations (< .60) between CA and PP tasks (Singleton et al., 2000; DeGraff, 2005). In a semi-transparent Norwegian orthography, individually administered PA measures have shown at least acceptable reliability (Nergård-Nilssen & Friborg, 2022). To our knowledge, validated CA of PA accuracy has not been studied in transparent orthographies, which highlights the need for further development and validation of such tasks.

Rapid automatised naming (RAN), the ability to retrieve and fluently name serially presented familiar items, is a known precursor of reading skills, especially fluency (for reviews, see e.g. Kirby, Georgiou, Martinussen, & Parrila, 2010; Norton & Wolf, 2012). The connection between RAN and reading fluency seems to be consistent across orthographies (e.g. Landerl et al., 2019) and remains strong across age groups (Furnes & Samuelsson, 2010; Torppa, Lyytinen, Erskine, Eklund, & Lyytinen, 2010; Van den Bos, Zijlstra, & Spelberg, 2002). To our knowledge, no previous studies comparing CA and PP RAN exist, probably because RAN is problematic for a CA presentation domain, as it essentially requires naming items aloud. For our study, however, we did include RAN CA tasks by developing and employing a speech-recognition system for the tasks.

Studies on the association between short-term memory (STM) or verbal STM (VSTM) and reading skills suggest that STM taps the verbal processes that are connected with reading and that children with poor reading skills often have difficulties in domain-specific STM tasks (Swanson, Zheng, & Jerman, 2009). In some studies, with transparent orthographies, STM has also differentiated between average and poor spellers (Brandenburg et al., 2015; Lervåg & Hulme, 2010; Wimmer & Schurz, 2010; Zygouris et al., 2025). In previous CA studies on STM, the CA span task was associated with reading and spelling accuracy (Protopapas &

Skaloumbakas, 2007), and the PP and CA tasks were moderately correlated (r=.53; Paul et al., 2005). However, STM was not among the best discriminators between good and poor readers (Protopapas et al., 2008).

There are not many studies on the connection between receptive vocabulary and reading, but the results refer to a stronger connection in opaque (Ouellette, 2006) than transparent orthographies (Torppa et al., 2013). The CA and PP versions of vocabulary tasks are typically very similar, requiring a user to recognise an auditorily presented target and select a corresponding picture out of four alternatives. A very strong correlation of 0.82 has been reported between CA and PP assessments with differing test items (Merrell & Tymms, 2007; but for opposite evidence, see DeGraff, 2005).

Finally, learning to read requires forming new associations, such as between letters and speech sounds and between orthographic, phonological, and semantic word representations. Evidence suggests that paired associative learning tasks, especially in a visual–verbal domain, explain reading performance on top of the other well-known reading subskills, such as phonological skills (Hulme, Goetz, Gooch, Adams, & Snowling, 2007; Wimmer, Mayringer, & Landerl, 1998), even after controlling for IQ (Ho, Chan, Tsang, Lee, & Chung, 2006; Windfuhr & Snowling, 2001). Associative learning is typically assessed by repeated presentation of visual–auditory item pairs with corrective feedback, similarly in both CA and PP domains. In adults, moderate to strong correlations with a range of r =.51–.76 between CA and PP domains in associative learning tasks have been obtained (Fratti, Bowden & Cook, 2017; Paul et al., 2005). To our knowledge, there are no studies with children comparing CA and PP domains in paired associative learning.

Purpose of the Study

The aim of this study is to develop and validate digital assessment methods (serious game assessments, SGA) for reading and reading-related skills suitable for a transparent orthography. These measures are integrated in a serious game (GraphoLearn) to serve early identification of reading difficulties and specifically to individually target the intervention according to the assessment results. The assessment tasks are designed so that they are simple enough to be used independently and are concise and child-friendly enough to maintain the children's interest, even those with low-end attentional and scholastic capabilities. An important feature of our SGA method is that immediate feedback on accuracy for each response is provided to direct the attention towards learning objectives (Black & Wiliam, 1998) and to serve learning opportunities in a style of dynamic assessment.

The first analyses on this assessment battery showed that with a set of reading and spelling SGA measures, the children with a risk for reading difficulties could be identified with very high specificity and sensitivity (91% and 95%, respectively; Authors, 2020). The aim of the present study is to explore and develop this SGA method further by investigating the validity of the measures more thoroughly as well as analysing the skills known to be related to reading that have previously not been broadly assessed in the context of CA (i.e., vocabulary, PA, RAN, STM, phonological memory, and associative learning). The specific research questions are (1) What is the validity of SGA measures compared to the PP assessment methods? and (2) To what extent do the SGA tasks assess reading accuracy and fluency?

We focus our investigation on the three aspects of validity, concurrent, construct and criterion validity. Concurrent validity informs to which extent two tasks designed to measure the same skill correlate with each other. In our study the goal is to obtain high correlation between a novel SGA measure and the established corresponding PP measure. The construct validity refers to the phenomenon that several tasks measure partly the same underlying skill or ability. In this case, these tasks load on the same factor reflecting the underlying skill. The goal in our study is that both SGA and PP versions of a task load on the same underlying factor and that the identified factors fit on theoretical understanding of reading skill. For example, theoretical understanding is that reading accuracy and fluency are somewhat separate skills, as well as memory and vocabulary. Criterion validity informs whether a sufficient degree of variance in a skill is captured by either a single measure or set of measures. Here the goal is to obtain a very high level of variance explained in PP measureswith combined SGA measures.

METHODS

Participants

The participants of this study consisted of 735 Finnish speaking monolinguals first- to fourth-grade children (aged 6 to 11, females 54%) from 17 different elementary schools in the central Finland area. Written consent to participate in the study was obtained from the children as well as their guardians. There were no differences in the parental education level or family gross earnings between the grades and schools. Therefore, the participating children represent Finnish L1 children attending typical public schools in Finland. Finnish schools are basically inclusive; thus, it is supposed that the normal variation of reading skills is represented in this sample. The description of the procedure of the data collection is provided elsewhere (Authors, 2020).

Measures

Paper and Pencil (PP) Measures

Reading was measured with three widely used tasks: lists of words (Häyrinen, Serenius-Sirve, & Korkman, 2013) and pseudowords (adapted from Salmi, Eklund, Järvisalo, & Aro, 2011) with increasing difficulty were read aloud. The third task was a Finnish adaptation of the Woodcock-Johnson sentence reading fluency test (Woodcock, McGrew, & Mather, 2001), in which the child has to verify written sentences (e.g. Strawberries are blue) as true or false. The child was instructed to read as quickly and accurately as possible. The percentage of correctly read or selected items and reading fluency (i.e., the number of items read correctly in one minute) were used as outcome scores. A more detailed description of the reading measures is reported elsewhere (Authors, 2020).

Reading-related skills were measured with standardised tests when available. We measured word spelling (Häyrinen et al., 2013), pseudoword spelling (created for this project, described below), letter knowledge (naming the Finnish letters), PA (phonological processing subtask of the *Developmental Neuropsychological Assessment NEPSY II*; Korkman, Kirk & Kemp, 2007), phonological memory (phonological memory subtask of NEPSY; Korkman, Kirk, & Kemp, 1998), STM (digit span forward subtest from the Wechsler Intelligence Scale for Children; Wechsler,

2010), associative learning (name learning subtask of NEPSY II; Korkman et al., 2007), RAN (letters and objects subtests from Ahonen, Tuovinen & Leppäsaari, 2003), and receptive vocabulary (shortened version of the Peabody Picture Vocabulary test PPVT; Dunn, Dunn, Bulheller, & Häcker, 1965; Lerkkanen et al., 2010). The descriptions of the standardised tasks are available from the manuals of the tests. The only measure with no such information is pseudoword spelling, which was created for the purposes of this study. In this task, the items with increasing difficulty (n = 10; from four to 15 letters, e.g., *siul, tuupposristekka*) were matched to the SGA items in terms of syllable and word structures as well as the number of letters. The child was asked to spell the pseudowords based on recorded auditory stimuli. In all PP reading-related measures accuracy (either the percentage or the number of correct answers) and in RAN fluency (time) was also used as an outcome measure.

Serious Game Assessment (SGA) Measures

Word Reading. A child listened to a word and selected a corresponding written word from four alternatives comprising words and pseudowords. The alternatives were phonologically highly similar to the target word (e.g., *luuri* 'handset' vs. pseudowords *luuki, luri, ruuli*). In the case of an incorrect selection, the correct response was shown on the screen. The length of the words increased gradually from bisyllabic words (e.g., *myrsky* 'storm') up to 6-syllable inflected words (the maximum length was 16 letters, e.g., *tuomaristossakin* 'also in the jury'). The task was discontinued after four errors within six trials. The percentage of correct selections (max = 40) and the response fluency (i.e., the amount of correct selections per minute) were used as the outcome scores. The test–retest correlation was .74.

Pseudoword Reading. A task similar to the word reading was presented with pseudowords (30 target items). The task proceeded from monosyllabic (e.g., *sien*) to multisyllabic pseudowords (the maximum of four syllables and 13 letters, e.g., *souraannuttaa*). The test–retest correlation was .66.

Sentence Reading Comprehension (SRC). As in PP mode, a Finnish adaptation of the Woodcock-Johnson sentence reading fluency (Woodcock et al., 2001) was used. The sentences were presented individually on the screen accompanied by true (green) and false (red) response options. Before continuing to the next sentence, a correct response was shown in case the child selected an incorrect response. The test–retest correlation was .89. The children who completed the SGA task before the two-minute time limit with less than 60% accuracy were considered guessers and excluded from the analyses (n = 29). Exploration of these children's data showed that most of them were first graders whose reading performance in PP measures was poor – i.e., they were guessing most likely because they could not read the sentence.

Word Spelling. A set of sublexical written items, letters and/or syllables, were presented on the screen (Figure 1). A word was presented via an audio track, and the task was to construct this word by selecting the right letter and/or syllables from the items (including distractors) presented on the screen and putting them in the correct order. The task proceeded to the next word once the target word was constructed correctly. The length of the target words increased progressively (e.g., from *ei* 'no' to *geenimanipuloitu* 'gene manipulated'). The percentage of correctly constructed target words at the first attempt (of a maximum of 20 words) and the response fluency (i.e., the amount of correctly constructed target words per minute) were used as the outcome scores. The test–retest correlation was .60.

Figure 1. A Screen Capture from the SGA Word Spelling Task.

Pseudoword Spelling. A task similar to the word spelling task was used for assessing pseudoword spelling (24 target items). The task was to construct pseudowords that increased in length from two to 15 letters (e.g., *ri* to *laannusvastikko*). The test–retest correlation was .69.

Letter Knowledge. The sound (representing a phoneme) for a letter was presented via an audio track, and the task was to select the corresponding letter among all 23 Finnish letters presented in a fixed random order on the tablet screen. In case of an incorrect selection, the correct letter was shown before proceeding to the next item. The number of correct selections (maximum = 23) was used as the outcome scores. Due to the ceiling effect, no retest correlation was calculated for this task.

Phonological Awareness. Three different words (e.g., *puku* 'suit', *puhe* 'speech', *suvi* 'summer') were presented aurally in an odd-one-out paradigm. The task was to identify the word that started with a different sound (i.e., syllable) compared to the other two. Each word was paired with a button on the tablet's screen, and the odd one out was selected by pressing the corresponding button. A practice trial with corrective feedback was provided to ensure that the child understood the task. The number of correct responses (maximum = 20) was used as the outcome scores. The test–retest correlation was .46.

Phonological Memory. The child was asked to detect which of the three pseudowords auditorily presented was different from the other two by pressing a corresponding button representing the first, second, or third item. The sets increased in length gradually. The deviating item had the same length and syllable structure as the other two, but distinct phonological content (e.g., /*heinostus/ /**vuonostus**/ /heinostus/).* The number of correct responses (maximum = 18) was used as the outcome scores. The test–retest correlation was .44.

Short-Term Memory. A set of coloured buttons (red, yellow, blue, green, black, white) was presented on the tablet screen. The child heard a list of colours and was instructed to select the colours in the same order he or she heard them. Two consecutive trials consisted of the same number of colours to recall. The colours in each set increased gradually from two to six, and the task was discontinued after failing to recall both trials with the same number of colours. The number of correct responses (maximum = 10) was used as the outcome scores. The test–retest correlation was .30.

Associative Learning. The child's task was to learn associations between five arbitrary visual symbols and spoken syllables. First, the monosyllabic sounds (*/ma/, /ha/, /sa/, /ka/, /ra/*) paired with corresponding visual symbols were introduced. Then, a monosyllabic sound was

presented via an audio track, and the child's task was to select the corresponding visual symbol from four options. Following an incorrect selection, the correct response was presented, and the child had to select it before continuing to the next trial. The task consisted of 25 trials in total. During the task, each sound-visual symbol pair was presented altogether five times. The score was the number of correct responses. The test–retest correlation was .75.

Rapid Automatised Naming. Two separate tasks, one for letters (*U, I, K, S, T*) and one for objects (*star, fence, hand, worm, button*), were presented in 10-item rows on the screen. Each set of items was replicated eight times in a random order but so that the items were not repeated successively. The naming performance was recorded with a computer application developed by our research team and was analysed by an open-source speech recognition program (Bolaños, 2012). Because the accuracy scores of the RAN task are typically high and do not discriminate between good and poor readers (e.g., Snyder & Downey, 1995), only the naming time for the whole list of items was used as the outcome score for each task.

Receptive Vocabulary. A new task in the style of the PPVT (Dunn et al., 1965) was developed. Four alternative pictures were presented on the screen, and the child was instructed to pick the alternative that matched the word he or she heard. The number of correct responses (maximum = 30) was used as the outcome scores. The test–retest correlation was .67.

Data analysis

The concurrent validity was analysed with inspecting paired correlations of corresponding PP and SGA tasks. The construct validity of SGA was analysed by an explorative factor analysis with the Oblimin rotation, allowing correlation of factors by analysing if the SGA measures were loading to the same factor as their supposed PP counterparts and therefore measured the same skills. To approach the frequently applied 10 observation per variable -ratio (Costello & Osborne, 2005), measures with low paired correlations or pairs loading inconsistently to factors were omitted. Finally, to analyse the criterion validity of the SGA measures explaining reading, stepwise regression analyses were conducted to find the best SGA predictors for the composite measures of reading fluency and accuracy as well as to show their explanatory power. In addition, we investigated whether and to which extent PP measures would add explanatory power above SGA measures.

RESULTS

The descriptive statistics on the background variables of participants, SGA measures, and PP measures for each school grade are provided in Tables 1 and 2. The results of comparing the performance between schools and school grades showed that there was no difference between the different schools and grades regarding either gender or socioeconomic status, and the performance level increased from grade to grade in each of the PP and SGA tasks. As was expected based on previous results from transparent orthographies, the differences were minor in most of the accuracy measures because of the ceiling effect.

Table 1. Descriptive Statistics of Paper and Pencil Measures in Grades 1-4.

	Grade 1 Mean (SD)	Grade 2 Mean (SD)	Grade 3 Mean (SD)	Grade 4 Mean (SD)	Effect Size	Post Hoc
Word reading accuracy %[a]	91.43 (9.97)	96.06 (4.18)	96.86 (3.20)	97.61 (2.52)		1 < 2, 3 < 4
Word reading fluency (words correct/min)	22.69 (10.67)	34.12 (8.36)	40.49 (8.11)	44.67 (7.38)	.416	1 < 2 < 3 < 4
PSW reading accuracy %[a]	77.92 (19.66)	89.11 (11.37)	84.31 (12.29)	88.42 (10.56)		1 < 2 < 3 < 4
PSW reading fluency (words correct/min)	17.95 (12.31)	30.17 (12.45)	23.53 (8.76)	28.49 (10.62)	.181	1 < 3 < 2, 4
SRC fluency (sentences correct/min)	6.73 (3.91)	12.07 (4.31)	16.40 (4.49)	20.37 (4.85)	.564	1 < 2 < 3 < 4
Word spelling accuracy %[a]	69.62 (25.92)	90.31 (13.31)	87.55 (10.33)	92.41 (6.30)		1 < 2 < 3 < 4
PSW spelling accuracy %[a]	59.89 (29.03)	79.40 (19.99)	89.94 (13.46)	92.72 (10.10)		1 < 2 < 3, 4
Letter knowledge (max 23) [a]	21.97 (1.25)	22.56 (.735)	22.77 (.572)	22.95 (.23)		1 < 2 < 3 < 4
Phonological awareness (max 53)	39.23 (5.90)	42.65 (5.16)	45.42 (3.54)	47.01 (3.29)	.262	1 < 2 < 3 < 4
Phonological memory (max 16)	9.35 (2.20)	9.94 (1.89)	9.93 (1.92)	10.18 (2.17)	.021	1 < 2, 3 < 4
Short term memory (max 16)	5.64 (1.20)	6.06 (1.24)	6.51 (1.26)	6.76 (1.37)	.102	1 < 2 < 3, 4
Associative learning (max 24)	10.45 (4.30)	12.40 (4.26)	13.41 (4.64)	14.36 (4.18)	.100	1 < 2 < 3 < 4
RAN objects duration (seconds)	63.67 (17.73)	57.66 (15.67)	52.49 (8.75)	47.49 (10.16)	.182	4 < 3 < 2 < 1
RAN letters duration (seconds)	41.03 (11.46)	32.41 6.93)	28.29 (5.76)	25.84 (5.34)	.375	4 < 3 < 2 < 1
Receptive vocabulary (max 30)	9.34 (4.75)	11.50 (5.99)	13.81 (5.63)	17.58 (5.65)	.228	1 < 2 < 3 < 4

Note. In tasks with a time limit, the number of correct responses is presented with accuracy percentage. In tasks without a time limit the number of correct responses is presented. PSW = pseudoword, SRC = Sentence reading comprehension, RAN = Rapid Automatised Naming

[a]Tests without effect sizes analyzed with non-parametric tests due to skewed distributions, all the tests significant at .01 level

Table 2. Descriptive Statistics of SGA Measures in Grades 1-4.

	Grade 1 Mean (SD)	Grade 2 Mean (SD)	Grade 3 Mean (SD)	Grade 4 Mean (SD)	Effect Size	Post Hoc
Word reading accuracy %	65.71 (14.71)	73.82 (10.78)	78.70 (11.14)	82.48 (8.48)	.228	1 < 2 < 3 < 4
Word reading fluency (correct/min)	8.90 (3.55)	12.03 (2.93)	13.44 (3.33)	15.35 (3.11)	.341	1 < 2 < 3 < 4
PSW reading accuracy	64.53 (17.28)	74.53 (10.10)	78.69 (9.32)	82.51 (8.92)	.236	1 < 2 < 3 < 4
PSW reading fluency (correct/min)	7.77 (3.10)	10.70 (2.49)	11.98 (2.80)	13.22 (2.68)	.343	1 < 2 < 3 < 4
SRC fluency (correct/min)	10.34 (5.62)	16.02 (5.15)	19.17 (4.83)	23.27 (5.13)	.400	1 < 2 < 3 < 4
Word spelling accuracy %	64.96 (17.24)	76.31 (13.36)	80.91 (12.04)	84.17 (9.56)	.198	1 < 2 < 3 < 4
PSW spelling accuracy %	60.95 (16.41)	70.45 (12.60)	76.13 (11.63)	78.44 (11.02)	.194	1 < 2 < 3, 4
Letter knowledge (max 23)[a]	20.73 (1.99)	20.92 (2.15)	21.23 (1.61)	21.35 (1.83)		1 < 3, 4; 2 < 4
Phonological awareness (max 20)[a]	13.46 (4.57)	14.91 (4.13)	16.38 (3.49)	17.25 (3.38)		1 < 2 < 3 < 4
Phonological memory (max 18)	10.27 (4.16)	11.12 (4.19)	12.44 (3.52)	13.80 (2.55)	.122	1 < 2 < 3 < 4
Short term memory (max 10)	6.01 (2.29)	6.51 (1.81)	7.35 (1.47)	7.73 (1.42)	.111	1, 2 < 3 < 4
Associative learning (max 25)	11.74 (4.29)	11.57 (4.71)	11.53 (5.08)	11.88 (5.13)		
RAN objects duration (seconds)	50.35 (12.46)	45.56 (9.52)	41.36 (7.24)	37.38 (7.24)	.227	4 < 3 < 2 < 1
RAN letters duration (seconds)	34.32 (9.88)	27.02 (5.88)	23.67 (6.08)	22.52 (5.11)	.217	4, 3 < 2 < 1
Receptive vocabulary (max 30)	19.09 (3.29)	20.34 (3.45)	21.38 (3.13)	22.88 (2.97)	.155	1 < 2 < 3 < 4

Note 1. In tasks with a time-limit or with a differing maximum score between school grades the outcome of correct answers is presented as accuracy percentage. In tasks without a time limit, the number of correct answers is reported and used as an outcome.
Note 2. [a]Tests without effect sizes analyzed with non-parametric tests due to skewed distributions, all the tests significant at .01 level.
Note 3. Main effect of grade was significant at the level < .01 except in associative learning < .05 level.
Note 4. SGA = Serious Game Assessment, PSW = pseudoword, SRC = Sentence reading comprehension, RAN = Rapid Automatised Naming.

Concurrent validity

Paired correlations between PP and SGA versions are presented in Table 3. Very strong (>.80) or strong (>. 60–.79) correlations were obtained for all reading fluency measures (word, pseudoword, and SRC), except for word reading fluency in Grade 3 (r = .59). Strong correlations were also obtained for spelling in Grade 1 and RAN Objects at Grade 4 and RAN Letters at Grade 1 and 3. Moderate (.40–.59) correlations were obtained for the rest of the RANs, spelling in Grades 2–4, PA in Grades 1–3, STM in Grade 2, and vocabulary in Grades 2–4. Weak (.20–.39) or very weak (< .19) correlations were obtained for letter knowledge, PA in Grade 4, phonological memory (Grades 1–4), and associative learning (Grades 1–4). These results indicate a high or moderate correspondence between PP and SGA in measures of reading fluency and RAN across the grades, whereas for reading accuracy, spelling and PA measures the correlations are high to moderate only in early grades. The rest of the reading-related skill measures, such as phonological memory, STM, and associative learning, seem to correspond with each other only to a moderate degree throughout the grades.

Consruct validity

For obtaining acceptable number of variables and an observation ratio, only subset of measures was selected into the factor analysis. First, measures with low concurrent validity (associative learning, phonological memory, letter knowledge) were omitted. In addition, word reading accuracy, PA and phonological memory measures did not consistently load on the same factor at each Grade and were therefore dropped from the final model. After these removals, five theoretical constructs of fluency, accuracy, RAN, short-term memory and vocabulary

Table 3. Paired Correlations Between Paper and Pencil and SGA Measures in Grades 1–4.

Variable	Grade 1	Grade 2	Grade 3	Grade 4
Word reading accuracy	.42	.41	.34	.31
Word reading fluency[1]	.81	.70	.59	.60
PSW reading accuracy	.60	.47	.50	.39
PSW reading fluency[1]	.76	.66	.69	.67
SRC fluency	.86	.74	.77	.70
Word spelling accuracy[2]	.73	.53	.44	.46
PSW spelling accuracy[2]	.66	.48	.42	.25
Letter knowledge[2]	.29	.13[3]	.09[3]	.20[3]
Phonological awareness	.55	.56	.45	.26
Phonological memory	.24	.29	.07[3]	.22
Short term memory	.35	.44	.22	.38
Associative learning	.24	.36	.24	.37
RAN objects	.58	.59	.58	.72
RAN letters	.70	.53	.61	.58
Receptive vocabulary	.36	.46	.51	.52

Note. SGA = Serious Game Assessment, PSW = Pseudoword, SRC = Sentence reading comprehension, RAN = Rapid automatised naming

[1] Reading accuracy included in fluency measures due to the ceiling effect.

[2] Analyzed with non-parametric methods (Spearman) due to the skewed distribution.

[3] n.s., while all other correlations were significant at the <.01 level.

Table 4. Summary of Exploratory Factor Analysis for Reading and Related Measures in Grades 1-4 Using Oblimin Rotation.

	Grade 1					Grade 2					Grade 3					Grade 4				
Measures	Acc	Flu	RAN	Voc	Mem	Acc	Flu	RAN	Voc	Mem	Acc	Flu	RAN	Voc	Mem	Acc	Flu	RAN	Voc	Mem
PP PSW reading %	.88					.74					-.61						.58			
SGA PSW reading %	.76					.58					-.59					-.70				
PP word spelling %	.86					.88					-.53	.40				-.69				
SGA word spelling %	.82					.68					-.83					-.47				
PP PSW spelling %	.82					.80					-.58					-.55				
SGA PSW spelling %	.66					.62					-.80					-.71				
PP word reading fluency		1.04					.86					.65					.79			
SGA word reading fluency		.87					.82					.59					.73			
PP PSW reading fluency		.84					.79					.72					.79			
SGA PSW reading fluency		.81					.84					.87					.78			
PP SRC fluency		.86					.69					.69					.77			
SGA SRC fluency		.68					.80					.86					.79			
PP RAN objects			.83					.71					.75					.73		
SGA RAN objects			.75					.71					.75					.76		
PP RAN letters			.64					.75					.75					.84		
SGA RAN letters			.81					.83					.82					.72		
PP Receptive vocabulary				.82					.82					.81					.84	
SGA Receptive vocabulary				.79					.84					.83					.82	
PP STM					.69					.74					.65					.77
SGA STM					.77					.73					.74					.84
Eigenvalues	1.47	9.72	1.72	1.23	1.03	2.05	7.48	1.83	1.43	.982	1.34	7.01	2.60	1.43	1.07	1.62	6.91	2.16	1.50	1.09
% of variance	7.32	48.61	8.61	6.16	5.13	10.25	37.05	9.16	7.13	4.91	6.71	35.07	13.00	7.17	5.36	8.08	34.56	10.80	7.47	5.46

Note. Acc = Accuracy, Flu = Fluency, RAN = Rapid automatised naming, Mem = Memory, Voc = Vocabulary, PP = Paper and pencil task, SGA = Serious Game Assessment, PSW = Pseudoword, SRC = Sentence reading comprehension, STM = Short-term memory

remained, determining the number of factors to constant five. Table 4 provides the factor loadings for each grade separately. The factor analyses produced a very stable factor structure highly similar across grades. In each grade, the largest factor was Fluency and smallest factor was Memory, while the order of the rest of the factors varied. Overall, the factor loading of Accuracy for each task seems to decrease from first to fourth grade, which is apparently caused by accuracy measures approaching the ceiling. In other factors, the loading remained high in all grades.

Criterion validity

First, we computed a composite measure of reading fluency from PP word, pseudoword and SRC tasks (Cronbach's alphas were 0.85, 0.83, 0.81, and 0.80 for the Grades 1-4, respectively). This composite measure was then analysed in a hierarchical regression analysis with a stepwise selection of independent variables (Table 5). To study how much SGA measures can explain the dependent variable, all SGA measures were added at the first step. Then, to study what extent the remaining PP measures help in explaining additional variance in the dependent variable, these measures were added at the second step. For fluency the SGA measures of SRC, pseudoword reading, word reading, and RAN letters were the best predictors, explaining 83%, 76%, 71%, and 68% of reading fluency from first to fourth grade, respectively. The PP measures increased the explanation rate only marginally (three percent at best). Overall the results indicate that the SGA reading tasks are sufficient for predicting reading fluency and that only marginal benefit would be obtained by including SGA spelling and other reading-related tasks (RAN, letter knowledge and receptive vocabulary).

For the accuracy construct, we computed a composite from pseudoword reading and spelling, and word spelling measures (Cronbach's alphas were 0.90, 0.84, 0.77, and 0.59 for the Grades 1-4, respectively). The composite measure was first inverted and log-transformed to obtain normality and similar regression analysis was run as was the case for the reading fluency composite (Table 6). The composite was predicted with all SGA measures by 66%, 57%, 58%, 51% from first to fourth grade, respectively. The overall lower explanation ratio relative to reading fluency may reflect in part the ceiling effects, and on the other hand, the lower concurrent validity of the accuracy measures relative to fluency measures. The PP measures of reading related skills explained additional variance of 3%, 8 %, 6% and 6 %, respectively, mostly via the PP Word reading accuracy measure. In evaluating of other significant SGA predictors apart of the reading measures, at least spelling measures are needed for assessing accuracy skill. PA, Letter knowledge and RAN Letters also explained some variance but only at specific Grades so their inclusion in the assessment battery may not be justified.

Table 5. Stepwise Regression Analyses on Measures Predicting Reading Fluency in Grades 1-4.

Grade 1	SGA Measures	R^2	$R^2\Delta$	*p*
	SGA SRC fluency	.718	.718	<.001
	SGA PSW reading fluency	.802	.084	<.001
	SGA PSW spelling	.816	.014	<.001
	SGA RAN letters	.823	.007	.014
	SGA Word reading fluency	.829	.006	.028
	PP additions	R^2	$R^2\Delta$	*p*
	PP Letter knowledge	.835	.006	.022
	PP RAN letters	.840	.005	.030
Grade 2	SGA Measures	R^2	$R^2\Delta$	*p*
	SGA SRC fluency	.619	.619	<.001
	SGA Word reading fluency	.705	.086	<.001
	SGA PSW reading fluency	.733	.028	<.001
	SGA PSW accuracy	.747	.014	.002
	SGA RAN letters	.756	.009	.012
	SGA Vocabulary	.762	.007	.029
	PP additions	R^2	$R^2\Delta$	*p*
	PP RAN letters	.769	.007	.027
Grade 3	SGA Measures	R^2	$R^2\Delta$	*p*
	SGA SRC fluency	.526	.526	<.001
	SGA PSW reading fluency	.669	.144	<.001
	SGA RAN letters	.696	.027	<.001
	SGA Word reading accuracy	.713	.017	.002
	PP additions	R^2	$R^2\Delta$	*p*
	PP Word writing accuracy	.733	.019	.001
	PP RAN Letters	.746	.013	.004
	PP Vocabulary	.754	.008	.019
Grade 4	SGA Measures	R^2	$R^2\Delta$	*p*
	SGA SRC fluency	.456	.456	<.001
	SGA PSW reading fluency	.593	.137	<.001
	SGA RAN letters	.628	.034	<.001
	SGA PSW reading accuracy	.669	.041	<.001
	SGA Word reading fluency	.680	.011	.020
	Model 2: PP additions	R^2	$R^2\Delta$	p
	PP RAN Objects	.691	.011	.018

Note. SGA = Serious Game Assessment, PP = Paper and pencil task, SRC = Sentence reading comprehension, PSW = Pseudoword, RAN = Rapid automatised naming, STM = Short term memory

Table 6. Stepwise Regression Analyses on Measures Predicting Reading Accuracy in Grades 1-4.

Grade 1	SGA Measures	R^2	$R^2\Delta$	*p*
	SGA PSW reading acc.	.515	.515	<.001
	SGA Word spelling acc.	.586	.071	<.001
	SGA SRC fluency	.636	.050	<.001
	SGA PSW reading fluency	.646	.010	.041
	SGA RAN letters	.658	.012	.025
	PP additions	R^2	$R^2\Delta$	*p*
	PP Phon. awareness	.682	.023	.001
	PP Vocabulary	.692	.011	.027
Grade 2	SGA Measures	R^2	$R^2\Delta$	*p*
	SGA PSW reading acc.	.403	.403	<.001
	SGA Phon. awareness	.479	.076	<.001
	SGA PSW reading fluency	.529	.050	<.001
	SGA Word spelling	.565	.036	<.001
	PP additions	R^2	$R^2\Delta$	*p*
	PP Word reading acc.	.635	.069	<.001
	PP Letter knowledge	.645	.010	.030

Grade 3	SGA Measures	R^2	$R^2\Delta$	*p*
	SGA PSW reading fluency	.342	.342	<.001
	SGA PSW spelling acc.	.518	.176	<.001
	SGA Letter knowledge	.554	.036	<.001
	SGA SRC fluency	.578	.025	.002
	PP additions	R^2	$R^2\Delta$	*p*
	PP Word reading accuracy	.620	.042	<.001
	PP Phon. awareness	.639	.019	.004
Grade 4	SGA Measures	R^2	$R^2\Delta$	*p*
	SGA PSW reading fluency	.337	.337	<.001
	SGA Word reading accuracy	.434	.098	<.001
	SGA PSW spelling accuracy	.487	.053	<.001
	SGA Word spelling accuracy	.510	.023	.006
	PP additions	R^2	$R^2\Delta$	p
	PP Word reading accuracy	.540	.030	.001
	PP Phon. memory	.557	.016	.016
	PP Phon. awareness	.569	.012	.034

Note. SGA = Serious Game Assessment, PP = Paper and pencil task, SRC = Sentence reading comprehension, PSW = Pseudoword, RAN = Rapid automatised naming, STM = Short term memory

DISCUSSION

The aim of this study was to develop and validate SGA measures for reading, spelling, and reading-related skills (PA, RAN, STM, phonological memory, receptive vocabulary, and associative learning) in the orthographically transparent Finnish language. Our results with children from the first four school grades demonstrated that the developed SGA measures of reading and spelling are valid methods for assessing reading fluency and accuracy, as reading measures alone were able to capture 68–83 % of the variance of reading fluency in each grade, and reading and spelling measures 51–66 % of variance in reading accuracy. Other reading related skills could explain only a marginal amount variance on top of these measures and in an inconsistent manner across different grades. Despite the clear benefits of SGA as a fast, cost-effective way of assessing reading skills, the measures and the methodology still have some further steps to take before SGA can fully replace the traditional paper and pencil methods of assessment. Next, we will discuss our results and the suggested further steps in more detail.

The Validity of the SGA Measures

The results showed that in general the concurrent validity of the SGA reading fluency measures (word/pseudoword reading and SRC) was high, as evidenced by the strong correlations (.59 – .86) between the paired SGA and PP tasks, while the correlation for most of the accuracy measures were high only at first grade but deteriorated towards higher grades. However, the rest of the reading related skills showed much weaker correlations with their PP counterparts – a finding that seems to emerge across studies (DeGraff, 2005; Merrell & Tymms, 2007; Protopapas & Skaloumbakas, 2007; Protopapas et al., 2008). There may be several reasons for the weak pairwise correlations in terms of reading-related measures between SGA and PP measures. First, the low correlations may be due to statistical issues caused by the ceiling effects found in the accuracy measures. Second, as our aim was not to develop SGA measures that are identical to PP measures, some of the low correlations may derive from the different demands of the tasks in SGA that will be discussed later in more detail.

We studied construct validity by investigating whether paired SGA and PP tasks measure the same underlying latent trait in a factor analysis. Indeed, the corresponding SGA and PP tasks loaded on the same stable factors (accuracy, fluency, RAN, vocabulary and STM) across the grades and vocabulary from the second grade onwards. However, the reading-related measures of phonological memory, PA, and associative learning failed to show consistent construct validity. Based on these findings, we conclude that the SGA measures for reading fluency, accuracy, RAN, vocabulary, and short-term memory measure the same skills as their PP counterparts across the grades. Despite notable methodological differences between studies, our results are in line with the previous validation study in opaque English orthography, where similar factors of decoding, fluency, and vocabulary were identified in Grades 1 and 3 (DeGraff, 2005). The loadings for the Accuracy factor showed a decreasing trend from first to fourth grade, which was apparently caused by most of the children learning to read and spell accurately; therefore, the ceiling effect was achieved (Aro & Wimmer, 2003; Seymour et al., 2003). Meanwhile, the loadings for the Fluency, Rapid Naming, Vocabulary and STM factors remained high in all grades, being indicative of the high discriminative capabilities of these

measures over development (Eklund et al., 2015; Landerl & Wimmer, 2008; Verhoeven, van Leeuwe, & Vermeer, 2011; Gathercole, Service, Hitch, Adams, & Martin, 1999; Nithart et al., 2011; Siegel & Ryan, 1989).

Noting the limits of cross-sectional data, a comparison between the grade levels allowed us to consider some points related to the development of the skills assessed and come to some conclusions of the theoretical validity of the measures. As expected, all the raw scores of the tasks showed a trend for increasing skills from one grade to another unless the ceiling effect was reached (accuracy measures; see also Auphan et al., 2019 for similar results).

Why did some reading-related measures (e.g., PA, phonological memory, STM, associative learning) fail the test of concurrent and construct validity in correlational and factor analyses, respectively? Some explanations may derive from the methodological limitations of the measures mentioned above (e.g., the ceiling effects or the small number of items limiting the variation) or differences in the cognitive requirements between SGA and PP tasks (e.g., oral production, motor dexterity, visual search, or strategies used for fluent performance). These explanations cannot be reliably affirmed in this study, but some remarks on the possible reasons are warranted. First, the measures for PA did not reach the ceiling, yet the paired correlations dropped in the third and fourth grades. This may be because the standardised PP task used in this study extended to more complex phonological skills in Grades 3–4 relative to Grades 1–2, whereas the SGA task remained the same through the grades. Moreover, the SGA version required only the identification of a relatively large (easy) psycholinguistic segment, namely differing initial syllable, whereas the scope of the PP tasks was more diverse, requiring a child to pay attention to the middle parts of the words and phonemes in addition to syllables. Thus, SGA PA tasks clearly need to be developed further to increase the content validity of the measure.

The SGA tasks assessing phonological and associative learning did not correlate well with their PP counterparts, which may be partly due to the different demands of the tasks. PP version of the phonological memory assessment required vocal reproduction, while in the SGA, the response was simply made by selection. Further, as the items in the PP tasks were very difficult to pronounce, one could easily make an error in spite of fairly good memory representation of the item. Finally, the associative learning tasks both relied on learning with corrective feedback, yet in the PP domain, the task was to learn names for faces, but in the SGA, the task was to learn meaningless syllables for unfamiliar visual symbols. No prior studies explicate the correlation between these two learning domains of verbal–visual learning, yet one may assume learning names for faces may be based on at least partially different neuronal networks due to their special socio-emotional function.

In sum, the validity of SGA reading measures was good, forming a strong basis for game-based reading assessment for school-aged children. However, the converging findings on the lower validity of reading-related skills (DeGraff, 2005; Merrell & Tymms, 2007; Protopapas & Skaloumbakas, 2007; Protopapas et al., 2008) clearly suggest that more work needs to be done to develop a SGA system capable of reliably and efficiently assessing not only reading skills but also related cognitive skills, such as phonological processing.

To What Extent Do the Developed Serious Game Assessment Tasks Assess Reading Skills?

We studied the criterion validity of the SGA measures, that is to what extent they are able to explain PP reading skills. In the regression analyses, the large majority (68–83%) of the variance of reading fluency could be assessed with corresponding SGA reading tasks. Such a high level of explanatory power is comparable to those reported by a previous study comparing CA and PP assessment in the English language (DeGraff, 2005). No clear developmental trend between school grades was apparent in how well reading fluency could be assessed with the SGA tasks, indicating that SGA can be used to evaluate reading skills from the first grade onwards. It should be noted though that even with the assessment task with the best overall explanatory power, sentence reading comprehension (SRC), its power seemed to drop between third and fourth grade. This may be due to the developmental shift in older children when reading becomes more affected by reading processes other than those captured by the SRC task (Rasinski, Rikli, & Johnston, 2009). Taken together, these results indicate that first- to fourth-grade children's reading fluency can be reliably assessed by three short SGA tasks, namely word reading, pseudoword reading, and SRC tasks administrated for a whole class at the same time by using inexpensive tablet devices. Since no manual work for scoring is needed and results are immediately available, this makes the SGA assessment method highly time- and cost-efficient.

Also reading and spelling accuracy could be explained well (ca. 51-66 %) by the SGA reading and spelling measures. However, the explanatory power was clearly highest at first grade, suggesting that the practical applicability of the present SGA accuracy tasks may be limited to first grade assessments. It must be noted though that assessing reading accuracy is mostly needed in early grades in transparent orthographies before accuracy approaches the ceiling in later grades (Seymour et al., 2003). According to our results, assessing also reading accuracy, not only fluency, would require including pseudoword reading and word spelling tasks into the SGA battery to compensate the ceiling effect of word reading accuracy.

We were also interested in the correlates of reading (spelling, PA, RAN, STM, phonological memory, receptive vocabulary, and associative learning) explaining reading fluency in transparent orthography. Of these non-reading tasks, the RAN task seemed to work best. This result is in line with the consensus on the strong correlation between RAN and reading fluency (e.g., Kirby et al., 2010; Moll et al., 2014; Norton & Wolf, 2012). Even though the execution of SGA RAN as oral production working on speech recognition technology sets limitations for its practical utilisation in a classroom context, it casts promises to the further use of SGA RAN in the field of assessment of reading and related skills, not least for the automatised analysis of the speech data. With that said, however, adding the non-reading tasks in analyses increased the variance explained of reading fluency only marginally, RAN 1.5–3.8%, and the other non-reading tasks (e.g., letter naming accuracy) less than 2% each. This result is in accordance with the previous conclusions by Hammill and colleagues, who argued that the skills they called "nonprint abilities" (e.g., phonology, grammar, RAN) have a rather minor role in predicting and instructing reading in school-aged children compared to print-based skills – i.e., "actual reading" (Hammill, 2004; Hammill, Mather, Allen, & Roberts, 2002). Even though we agree with their conclusion, we add that our results do not trivialise the use of these non-reading measures as indicators of reading difficulties before reading

instruction (i.e., when print-based assessment tools are not applicable; see de Jong & van der Leij, 2003; Lervåg & Hulme, 2009; Puolakanaho et al., 2007; Wolf, Bowers, & Biddle, 2000). Undoubtedly, non-reading measures are also of great use as tools for clinical work in analysing the background of reading skills more thoroughly (Hammill, 2004).

Limitations

This study focused on developing SGA technology of reading and related skills for an unselected sample of L1 children, not assessing the possible reasons behind low or high performance in the studied skills. Therefore, we did not assess children's general cognitive ability or the presence of neurological or developmental disorders. As a consequence, the present technology should be used only for screening and should not be used alone for diagnosing reading problems. Even though the generalisation of this study may be limited for some student groups (e.g., L2 or children with special needs), children of all reading levels were included in the study. This study thus provides important indications for the use of SGA for identification of children with a risk for reading difficulties in a transparent orthography, as also shown in another study investigating the specificity and sensitivity of these measures (Authors, 2020).

One of the drawbacks of SGA is the need to use multiple-choice tasks in assessment, which increases the temptation to guess the answers. However, features of the programme, such as corrective feedback provided for every item and the delay in presenting the next item after an incorrect answer, will probably reduce this tendency. Indeed, when the performance of children with low accuracy rates was analysed in more detail, it was shown that most of the children prone to guessing were actually poor readers who probably compensated for their inability to read with guessing. Finally, the research design was cross-sectional, therefore providing no information on how predictive the specific SGA measures are for future reading development.

Based on previous research regarding the significant moderator effect of adult interaction in computer-assisted instruction (not assessment) (McTigue, Solheim, Zimmer, & Uppstad, 2020), we would be cautious in administering the SGA without any adult supervision. However, based on the results of this study, the group administration with one or two supervisors seems to be sufficient for valid results, which is a great benefit compared to individual, professional-administered assessment.

CONCLUSIONS

The previous findings on computerised assessments, and here the extended findings of our study illustrate the applicability of serious game assessment tasks to assess reading and related skills (Authors, 2020; DeGraff, 2005; Merrel & Tymms, 2007; Protopapas et al., 2008; Puolakanaho & Latvala, 2017). These kinds of applications provide considerable benefits: there are no restrictions as to who can administer validated assessments, the assessments will be always equal since the assessment tasks in SGA will provide the very same instructions and feedback for those taking the assessments, due to the possibility of quick group assessments and exact data as well as results being immediately available valuable working time of

practitioners will be saved. Adopting CA/SGA further increases the time efficiency of the assessment (see Merrell & Tymms, 2007; DeGraff, 2005).

In conclusion, our results support the use of serious game assessments for reading skills as a tool for detecting a risk for reading problems in transparent orthographies among elementary school children. Ready to be integrated in an educational serious game, this method serves as a tool for adaptive intervention as explicated next.

IMPLICATIONS FOR APPLICATION

Having validated the assessment methods, we can integrate the SGA methods into the digital intervention method, the serious game GraphoLearn, providing the needed extra tool for teachers for early identification and support of struggling readers. GraphoLearn (called Ekapeli in Finland) is already used by tens of thousands of children in Finland. GraphoLearn is also available in other languages for research purposes (see e.g., Borleffs et al., 2018). However, thus far, GraphoLearn has not included a validated reading assessment module. Therefore, based on the evidence of the current study, the next step is to develop language specific assessment content to other GraphoLearn versions that would form reliable, usable, and time-efficient SGA methods for everyone.

Furthermore, by directly applying validated digital assessment data, adaptive assessment systems integrated in serious games could provide appropriate feedback to the learner as well as enable an individually adaptive learning environment for learners with individual needs (Authors, 2020; Jamshidifarsani, Garbaya, Lim, Blazevic, & Ritchie, 2019), which seems to benefit especially those with lower initial skills (Hooshyar, Yousefi, & Lim, 2018). As the serious game platform, GraphoLearn, used here is already widely used in many orthographies, the results of this study may provide a basis for the use of SGA also in other language versions at least in regard to identifying reading fluency problems, a character that seems to be a universal feature for dyslexia (Ziegler, Perry, Ma-Wyatt, Ladner, & Schulte-Körne, 2003).
Recent advances show that dynamic assessment principles can be successfully implemented in digital and game-based formats (Foldnes, Uppstad, Grønneberg, & Thomson, 2024; Glatz et al., 2023; Maassen, Glatz, Borleffs, Martínez, & de Groot, 2025). Such approaches provide a scalable way to evaluate learning potential and responsiveness, complementing earlier face-to-face models (Petersen, Allen, & Spencer, 2016; Cho et al., 2014). Building on these results, the aim of our work is to integrate early identification with dynamic assessment and support in a digital environment that benefits both children at risk and their educators.

ENDNOTES

1. The GraphoLearn is a serious game developed at the University of Jyväskylä and Niilo Mäki Institute (NMI; a practice-orientated research and development non-governmental, non-profit organization focusing on child and youth learning disabilities and support) since 2004. Prior 2017 the game was called GraphoGame. In 2017 the IP's of the game including the name GraphoGame were transferred to a private company Grapho Group Ltd, therefore now GraphoGame is an application solely owned by Grapho Group Ltd that distributes GraphoGame applications, including their sales, marketing and

advertising. Independent to the private company, JYU and NMI continue autonomous, open and ethically sound research and development of the serious game calling the game GraphoLearn. In Finland, GraphoLearn (for the Finnish language called "Ekapeli", and the Finnish-Swedish language called "Spell-Ett") is available for everyone for free. However, outside Finland, GraphoLearn is used for research purposes only and it is not a commercially available product.

REFERENCES

Ahonen, T., Tuovinen, S., & Leppäsaari, T. (2003). *Nopean sarjallisen nimeämisen testi [The test of rapid serial naming].* ER-Paino OY: Haukkarannan koulun julkaisut, Niilo Mäki Instituutti, Lievestuore.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics, 24*, 621–635. https://doi.org/10.1017/s0142716403000316

Auphan, P., Ecalle, J., & Magnan, A. (2019). Computer-based assessment of reading ability and subtypes of readers with reading comprehension difficulties: A study in French children from G2 to G9. *European Journal of Psychology of Education, 34*, 641–663. https://doi.org/10.1007/s10212-018-0396-7

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*, 370–407. https://doi.org/10.3102/0091732x14554179

Björn, P. M., Aro, M. T., Koponen, T. K., Fuchs, L. S., & Fuchs, D. H. (2016). The many faces of special education within RTI frameworks in the United States and Finland. *Learning Disability Quarterly, 39*, 58–66. https://doi.org/ 10.1177/0731948715594787

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Blok, H., Oostdam, R., Otter, M. E., & Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: A review. *Review of Educational Research, 72*(1), 101–130. https://doi.org/10.3102/00346543072001101

Bolanos, D. (2012, December). The bavieca open-source speech recognition toolkit. In 2012 IEEE Spoken Language Technology Workshop (SLT) (pp. 354-359). IEEE.

Borleffs, E., Glatz, T. K., Daulay, D. A., Richardson, U., Zwarts, F., & Maassen, B. A. (2018). GraphoGame SI: The development of a technology-enhanced literacy learning tool for Standard Indonesian. *European Journal of Psychology of Education, 33*, 595–613. https://doi.org/10.1007/s10212-017-0354-9

Brandenburg, J., Klesczewski, J., Fischbach, A., Schuchardt, K., Büttner, G., & Hasselhorn, M. (2015). Working memory in children with learning disabilities in reading versus spelling: Searching for overlapping and specific cognitive factors. *Journal of Learning Disabilities, 48*(6), 622–634. https://doi.org/10.1177/0022219414521665

Caravolas, M. (2004). Spelling development in alphabetic writing systems: A cross-linguistic perspective. *European Psychologist, 9*(1), 3–14. https://doi.org/10.1027/1016-9040.9.1.3

Carson, K., Boustead, T., & Gillon, G. (2014). Predicting reading outcomes in the classroom using a computer-based phonological awareness screening and monitoring assessment (Com-PASMA). *International Journal of Speech-Language Pathology, 16*, 552–561. https://doi.org/10.3109/17549507.2013.855261

Cho, E., Compton, D. L., Fuchs, D., Fuchs, L. S., & Bouton, B. (2014). Examining the predictive validity of a dynamic assessment of decoding to forecast response to tier 2 intervention. *Journal of Learning Disabilities, 47*, 409–423. https://doi.org/10.1177/0022219412466703

Clemens, N. H., Hagan-Burke, S., Luo, W., Cerda, C., Blakely, A., Frosch, J. … & Jones, M. (2015). The Predictive Validity of a Computer-Adaptive Assessment of Kindergarten and First-Grade Reading Skills. *School Psychology Review, 44*(1), 76–97. https://doi.org/10.17105/spr44-1.76-97

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation, 10*(1), 7. https://doi.org/10.7275/jyj1-4868

de Jong, P. F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology, 91*(3), 450–476. https://doi.org/10.1037/0022-0663.91.3.450

de Jong, P. F., & van der Leij, A. (2003). Developmental changes in the manifestation of a phonological deficit in dyslexic children learning to read a regular orthography. *Journal of Educational Psychology, 95*(1), 22–40. https://doi.org/10.1037/0022-0663.95.1.22

DeGraff, A. J. (2005). *Monitoring Growth in Early Reading Skills: Validation of a Computer Adaptive Test*. Retrieved from http://purl.flvc.org/fsu/fd/FSU_migr_etd-0788

Dunn, L. M., Dunn, L. M., Bulheller, S., & Häcker, H. (1965). *Peabody picture vocabulary test*. Circle Pines, MN: American Guidance Service.

Eklund, G., Sundqvist, C., Lindell, M., & Toppinen, H. (2021). A study of Finnish primary school teachers' experiences of their role and competences by implementing the three-tiered support. *European Journal of Special Needs Education, 36*(5), 729-742. https://doi.org/10.1080/08856257.2020.1790885

Eklund, K., Torppa, M., Aro, M., Leppänen, P. H., & Lyytinen, H. (2015). Literacy skill development of children with familial risk for dyslexia through grades 2, 3, and 8. *Journal of Educational Psychology, 107*(1), 126–140. https://doi.org/10.1037/a0037121

Fletcher, J. M., Francis, D. J., Foorman, B. R., & Schatschneider, C. (2021). Early detection of dyslexia risk: Development of brief, teacher-administered screens. *Learning Disability Quarterly, 44*(3), 145-157. https://doi.org/10.1177/0731948720931870

Foldnes, N., Uppstad, P. H., Grønneberg, S., & Thomson, J. M. (2024). School entry detection of struggling readers using gameplay data and machine learning. *Frontiers in Education, 9*, 1487694. https://doi.org/10.3389/feduc.2024.1487694

Fratti, S., Bowden, S. C., & Cook, M. J. (2017). Reliability and validity of the CogState computerized battery in patients with seizure disorders and healthy young adults: Comparison with standard neuropsychological tests. *The Clinical Neuropsychologist, 31*(3), 569–586. https://doi.org/10.1080/13854046.2016.1256435

Fuchs, D., & Fuchs, L. S. (2005). Responsiveness-to-intervention: A blueprint for practitioners, policymakers, and parents. *Teaching Exceptional Children, 38*(1), 57–61. https://doi.org/10.1177/004005990503800112

Furnes, B., & Samuelsson, S. (2010). Predicting reading and spelling difficulties in transparent and opaque orthographies: A comparison between Scandinavian and US/Australian children. *Dyslexia, 16*(2), 119–142. https://doi.org/10.1002/dys.401

Furnes, B., & Samuelsson, S. (2011). Phonological awareness and rapid automatized naming predicting early development in reading and spelling: Results from a cross-linguistic longitudinal study. *Learning and Individual Differences, 21*(1), 85–95. https://doi.org/10.1016/j.lindif.2010.10.005

Gathercole, S. E., Service, E., Hitch, G. J., Adams, A. M., & Martin, A. J. (1999). Phonological short-term memory and vocabulary development: Further evidence on the nature of the relationship. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 13*(1), 65–77. https://doi.org/10.1002/(sici)1099-0720(199902)13:1<65::aid-acp548>3.0.co;2-o

Georgiou, G. K., Torppa, M., Landerl, K., Desrochers, A., Manolitsis, G., de Jong, P. F., & Parrila, R. (2020). Reading and spelling development across languages varying in orthographic consistency: Do their paths cross? *Child Development, 91*(2), e266–e279. https://doi.org/10.1111/cdev.13218

Glatz, T., Tops, W., Borleffs, E., Richardson, U., Maurits, N., Desoete, A., & Maassen, B. (2023). Dynamic assessment of the effectiveness of digital game-based literacy training in beginning readers: a cluster randomised controlled trial. *PeerJ, 11*, e15499. https://doi.org/10.7717/peerj.15499

GraphoLearn. (n.d.). Partners. Retrieved from https://info.grapholearn.com/partners/

Hammill, D. D. (2004). What we know about correlates of reading. *Exceptional Children, 70*, 453–468. https://doi.org/10.1177/001440290407000405

Hammill, D. D., Mather, N., Allen, E. A. & Roberts, R. (2002). Using semantics, grammar, phonology, and rapid naming tasks to predict word identification. *Journal of Learning Disabilities, 35*, 121–136. https://doi.org/10.1177/002221940203500204

Heikkilä, R., Korpivaara P., Kettunen, A., Shenouda Khalil, K., Ryssy, J., Stylman, E., Westerholm, J., Hautala, J., & Niskakoski, M. (2023). *AKI – Alakoulun DigiLukiseula: luku- ja kirjoitustaidon sähköiset tuen tarpeen tunnistamisen välineet luokille 1–6. [DigiLukiseula - Digital Reading and Spelling Screening Tool for Elementary School]* Niilo Mäki Instituutti. Available in: https://digilukiseula.nmi.fi/alakoulun-digilukiseula/

Ho, C. S.-H., Chan, D. W., Tsang, S.-M., Lee, S.-H., & Chung, K. K. H. (2006). Word learning deficit among Chinese dyslexic children. *Journal of Child Language, 33*, 145–161. https://doi.org/10.1017/s0305000905007154

Hooshyar, D., Yousefi, M., & Lim, H. (2018). A procedural content generation-based framework for educational games: Toward a tailored data-driven game for developing early English reading skills. *Journal of Educational Computing Research, 56*, 293–310. https://doi.org/10.1177/0735633117706909

Hulme, C., Goetz, K., Gooch, D., Adams, J., & Snowling, M. J. (2007). Paired-associate learning, phoneme awareness, and learning to read. *Journal of Experimental Child Psychology, 96*(2), 150–166. https://doi.org/10.1016/j.jecp.2006.09.002

Häyrinen, T., Serenius-Sirve S., & Korkman, M. (2013). *Lukilasse 2. Lukemisen, kirjoittamisen ja laskemisen seulontatesti 1.–6. vuosiluokille. [A screening test for reading, spelling, and arithmetic for grades 1-6].* Helsinki: Hogrefe Psykologien Kustannus Oy.

Jamshidifarsani, H., Garbaya, S., Lim, T., Blazevic, P., & Ritchie, J. M. (2019). Technology-based reading intervention programs for elementary grades: An analytical review. *Computers & Education, 128*, 427–451. https://doi.org/10.1016/j.compedu.2018.10.003

Jiménez, J. E., García, E., & Balade, J. (2024). Advancing Dyslexia Assessment in Children through Computerized Testing. *J. Vis. Exp, 210*, e67031. https://doi.org/10.3791/67031

Juul, H., Poulsen, M., & Elbro, C. (2014). Separating speed from accuracy in beginning reading development. *Journal of Educational Psychology, 106*(4), 1096–1106. https://doi.org/10.1037/a0037100

Kingston, N. M. (2008). Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education, 22*(1), 22–37. https://doi.org/10.1080/08957340802558326

Kirby, J. R., Georgiou, G. K., Martinussen, R., & Parrila, R. (2010). Naming speed and reading: From prediction to instruction. *Reading Research Quarterly, 45*(3), 341–362. https://doi.org/10.1598/rrq.45.3.4

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY-II*. San Antonio, TX: Pearson.

Korkman, M., Kirk, U., & Kemp, S. L. (1998). *NEPSY: Lasten neuropsykologinen tutkimus [The neuropsychological assessment of children].* Helsinki: Psykologien kustannus.

Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (2019). Construct equivalence of pisa reading comprehension measured with paper-based and computer-based assessments. *Educational Measurement: Issues and Practice, 38*(3), 97–111. https://doi.org/10.1111/emip.12280

Landerl, K., Freudenthaler, H. H., Heene, M., De Jong, P. F., Desrochers, A., Manolitsis, G., ... & Georgiou, G. K. (2019). Phonological awareness and rapid automatized naming as longitudinal predictors of reading in five alphabetic orthographies with varying degrees of consistency. *Scientific Studies of Reading, 23*(3), 220–234. https://doi.org/10.1080/10888438.2018.1510936

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology, 100*(1), 150–161. https://doi.org/10.1037/0022-0663.100.1.150

Leppänen, U., Niemi, P., Aunola, K., & Nurmi, J. E. (2006). Development of reading and spelling Finnish from preschool to grade 1 and grade 2. *Scientific Studies of Reading, 10*(1), 3–30. https://doi.org/10.1207/s1532799xssr1001_2

Lerkkanen, M.-K., Poikkeus A.-M., Ahonen, T., Siekkinen, M., Niemi, P., & Nurmi, J.-E. (2010). Luku- ja kirjoitustaidon kehitys sekä motivaatio esi- ja alkuopetusvuosina. [The development of reading and spelling skills and motivation during early school grades]. *Kasvatus, 41*(2), 116–128.

Lervåg, A., & Hulme, C. (2009). Rapid automatized naming (RAN) taps a mechanism that places constraints on the development of early reading fluency. *Psychological Science, 20*(8), 1040–1048. https://doi.org/10.1111/j.1467-9280.2009.02405.x

Lervåg, A., & Hulme, C. (2010). Predicting the growth of early spelling skills: Are there heterogeneous developmental trajectories? *Scientific Studies of Reading, 14*(6), 485-513. https://doi.org/10.1080/10888431003623488

Ma, W. A., Richie-Halford, A., Burkhardt, A. K., Kanopka, K., Chou, C., Domingue, B. W., & Yeatman, J. D. (2025). ROAR-CAT: Rapid Online Assessment of Reading ability with Computerized Adaptive Testing. *Behavior Research Methods*, *57*(1), 56. https://doi.org/10.3758/s13428-024-02578-y

Maassen, B. A., Glatz, T., Borleffs, E., Martínez, C., & de Groot, B. J. (2025). Digital game-based learning for dynamic assessment and early intervention targeting reading difficulties: Cross-linguistic studies of GraphoLearn. *Clinical Linguistics & Phonetics*, *39*(6-8), 576-601. https://doi.org/10.1080/02699206.2025.2452979

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

Melby-Lervåg, M., Lyster, S. A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin, 138*(2), 322–352. https://doi.org/10.1037/a0026744

Merrell, C., & Tymms, P. (2007). Identifying reading problems with computer-adaptive assessments. *Journal of Computer Assisted Learning, 23*(1), 27–35. https://doi.org/10.1111/j.1365-2729.2007.00196.x

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. https://doi.org/10.1016/j.learninstruc.2013.09.003

Nergård-Nilssen, T., & Friborg, O. (2022). The Dyslexia Marker Test for Children: Development and Validation of a New Test. *Assessment for Effective Intervention*, *48*(1), 23-33. https://doi.org/10.1177/15345084211063533

Niskakoski, M., Määttä, S., Korpivaara, P., & Westerholm, J. (2020). *Yläkoulun DigiLukiseula: digitaalinen luku- ja kirjoitustaidon arviointimenetelmä 7. ja 8. -luokkalaisille. [DigiLukiseula - Digital Reading and Spelling Screening Tool for Higschool]*. Available in: https://digilukiseula.nmi.fi/ylakoulun-digilukiseula/

Nithart, C., Demont, E., Metz-Lutz, M. N., Majerus, S., Poncelet, M., & Leybaert, J. (2011). Early contribution of phonological awareness and later influence of phonological memory throughout reading acquisition. *Journal of Research in Reading, 34*(3), 346–363. https://doi.org/10.1111/j.1467-9817.2009.01427.x

Norton, E. S., & Wolf, M. (2012). Rapid automatized naming (RAN) and reading fluency: Implications for understanding and treatment of reading disabilities. *Annual Review of Psychology, 63*, 427–452. https://doi.org/10.1146/annurev-psych-120710-100431

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology, 98*(3), 554–566. https://doi.org/10.1037/0022-0663.98.3.554

Paananen, M., Pöyliö, H., Määttä, S., Hautala, J., Eklund, K., Kinnunen, M., Westerholm, J., & Holopainen, L. (2019). *DigiLukiseula: digitaalinen luku- ja kirjoitustaidon arviointimenetelmä nuorille ja aikuisille. [DigiLukiseula - Digital Reading and Spelling Screening Tool for young people and adults]*. Available in: https://digilukiseula.nmi.fi/digilukiseula/

Paleczek, L., Seifert, S., & Schöfl, M. (2021). Comparing digital to print assessment of receptive vocabulary with GraWo-KiGa in Austrian kindergarten. *British Journal of Educational Technology, 52*(6), 2145-2161 https://doi.org/10.1111/bjet.13163

Parrila, R., Kirby, J. R., & McQuarrie, L. (2004). Articulation rate, naming speed, verbal short-term memory, and phonological awareness: Longitudinal predictors of early reading development? *Scientific Studies of Reading, 8*(1), 3–26. https://doi.org/10.1207/s1532799xssr0801_2

Paul, R. H., Lawrence, J., Williams, L. M., Richard, C. C., Cooper, N., & Gordon, E. (2005). Preliminary validity of "integneuroTM": A new computerized battery of neurocognitive tests. *International Journal of Neuroscience, 115*(11), 1549–1567. https://doi.org/10.1080/00207450590957890

Petersen, D. B., Allen, M. M., & Spencer, T. D. (2016). Predicting reading difficulty in first grade using dynamic assessment of decoding in early kindergarten: A large-scale longitudinal study. *Journal of Learning Disabilities, 49*(2), 200–215. https://doi.org/10.1177/0022219414538518

Protopapas, A., & Skaloumbakas, C. (2007). Traditional and computer-based screening and diagnosis of reading disabilities in Greek. *Journal of Learning Disabilities, 40*(1), 15–36. https://doi.org/10.1177/00222194070400010201

Protopapas, A., Skaloumbakas, C., & Bali, P. (2008). Validation of unsupervised computer-based screening for reading disability in the Greek elementary Grades 3 and 4. *Learning Disabilities: A Contemporary Journal, 6*(1), 45–69.

Puolakanaho, A., Ahonen, T., Aro, M., Eklund, K., Leppänen, P. H., Poikkeus, A. M., ... & Lyytinen, H. (2007). Very early phonological and language skills: Estimating individual risk of reading disability. *Journal of Child Psychology and Psychiatry, 48*, 923–931. https://doi.org/10.1111/j.1469-7610.2007.01763.x

Puolakanaho, A. & Latvala, J. M. (2017). Embedding preschool assessment methods into digital learning games to predict early reading skills. *Human Technology, 13*, 216–236. https://doi.org/10.17011/ht/urn.201711104212

Rasinski, T., Rikli, A., & Johnston, S. (2009). Reading fluency: More than automaticity? More than a concern for the primary grades?. *Literacy Research and Instruction, 48*(4), 350–361. https://doi.org/10.1080/19388070802468715

Richardson, U., & Lyytinen, H. (2014). The GraphoGame method: The theoretical and methodological background of the technology-enhanced learning environment for learning to read. *Human Technology, 10*(1), 39–60. https://doi.org/10.17011/ht/urn.201405281859

Sainsbury, M., & Benton, T. (2011). Designing a formative e-assessment: Latent class analysis of early reading skills. *British Journal of Educational Technology, 42*(3), 500–514. https://doi.org/10.1111/j.1467-8535.2009.01044.x

Salmi, P., Eklund, K., Järvisalo, E., & Aro, M. (2011). LukiMat-Oppimisen arviointi: Lukemisen ja kirjoittamisen tuen tarpeen tunnistamisen välineet 2. luokalle. Käyttäjän opas. Available in: http://www.lukimat.fi/lukimat-oppimisen-arviointi/materiaalit/tuen-tarpeen-tunnistaminen/2lk/lukeminen/kayttajan-opas

Seifert, S., & Paleczek, L. (2021). Digitally Assessing Text Comprehension in Grades 3-4: Test Development and Validation. *Electronic Journal of e-Learning, 19*(5), 336–348. https://doi.org/10.34190/ejel.19.5.2467

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

Siegel, L. S., & Ryan, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled children. *Child Development*, 973–980. https://doi.org/10.2307/1131037

Singleton, C., Thomas, K., & Horne, J. (2000). Computer-based cognitive assessment and the development of reading. *Journal of Research in Reading, 23*(2), 158–180. https://doi.org/10.1111/1467-9817.00112

Snyder, L. S., & Downey, D. M. (1995). Serial rapid naming skills in children with reading disabilities. *Annals of Dyslexia, 45*, 29–49. https://doi.org/10.1007/bf02648211

Swanson, H. L., Zheng, X., & Jerman, O. (2009). Working memory, short-term memory, and reading disabilities: A selective meta-analysis of the literature. *Journal of Learning Disabilities, 42*(3), 260–287. https://doi.org/10.1177/0022219409331958

Torppa, M., Georgiou, G. K., Niemi, P., Lerkkanen, M-K., & Poikkeus, A-M. (2017). The precursors of double dissociation between reading and spelling in a transparent orthography. *Annals of Dyslexia, 67*, 42–62. https://doi.org/10.1007/s11881-016-0131-5

Torppa, M., Lyytinen, P., Erskine, J., Eklund, K., & Lyytinen, H. (2010). Language development, literacy skills, and predictive connections to reading in Finnish children with and without familial risk for dyslexia. *Journal of Learning Disabilities, 43*(4), 308–321. https://doi.org/10.1177/0022219410369096

Torppa, M., Parrila, R., Niemi, P., Lerkkanen, M. K., Poikkeus, A. M., & Nurmi, J. E. (2013). The double deficit hypothesis in the transparent Finnish orthography: A longitudinal study from kindergarten to Grade 2. *Reading and Writing, 26*(8), 1353–1380. https://doi.org/10.1007/s11145-012-9423-2

Vaessen, A., & Blomert, L. (2010). Long-term cognitive dynamics of fluent reading development. *Journal of Experimental Child Psychology, 105*(3), 213–231. https://doi.org/10.1016/j.jecp.2009.11.005

Vaessen, A., Gerretsen, P., & Blomert, L. (2009). Naming problems do not reflect a second independent core deficit in dyslexia: Double deficits explored. *Journal of Experimental Child Psychology, 103*(2), 202–221. https://doi.org/10.1016/j.jecp.2008.12.004

Van den Bos, K. P., Zijlstra, B. J., & Spelberg, H. C. (2002). Life-span data on continuous-naming speeds of numbers, letters, colors, and pictured objects, and word-reading speed. *Scientific Studies of Reading, 6*(1), 25–49. https://doi.org/10.1207/s1532799xssr0601_02

Virinkoski, R., Lerkkanen, M. K., Holopainen, L., Eklund, K., & Aro, M. (2018). Teachers' ability to identify children at early risk for reading difficulties in Grade 1. *Early Childhood Education Journal, 46*(5), 497–509. https://doi.org/10.1007/s10643-017-0883-5

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5–24. https://doi.org/10.1177/0013164407305592

Wechsler D. (2010). *Wechsler intelligence scale for children—Fourth Edition Finnish*. Helsinki: Psykologien Kustannus Oy.

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading, 15*(1), 8–25. https://doi.org/10.1080/10888438.2011.536125

Wimmer, H., & Schurz, M. (2010). Dyslexia in regular orthographies: manifestation and causation. *Dyslexia, 16*(4), 283–299. https://doi.org/10.1002/dys.411

Wimmer, H., Mayringer, H., & Landerl, K. (1998). Poor reading: A deficit in skill-automatization or a phonological deficit? *Scientific Studies of Reading, 2*(4), 321–340. https://doi.org/10.1207/s1532799xssr0204_2

Windfuhr, K. L., & Snowling, M. J. (2001). The relationship between paired associate learning and phonological skills in normally developing readers. *Journal of Experimental Child Psychology, 80*(2), 160–173. https://doi.org/10.1006/jecp.2000.2625

Wolf, M., Bowers, P. G., & Biddle, K. (2000). Naming-speed processes, timing, and reading: A conceptual review. *Journal of Learning Disabilities, 33*(4), 387–407. https://doi.org/10.1177/002221940003300409

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of cognitive* abilities (pp. 371–401). Riverside Pub.

Yeatman, J. D., Tran, J. E., Burkhardt, A. K., Ma, W. A., Mitchell, J. L., Yablonski, M., ... & Richie-Halford, A. (2024). Development and validation of a rapid and precise online sentence reading efficiency assessment. *Frontiers in Education* (Vol. 9, p. 1494431). https://doi.org/10.3389/feduc.2024.1494431

Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Körne, G. (2003). Developmental dyslexia in different languages: Language-specific or universal? *Journal of Experimental Child Psychology*, *86*(3), 169–193. https://doi.org/10.1016/s0022-0965(03)00139-5

Zygouris, N. C., Vlachos, F., Styliaras, S. K., Tziallas, G. D., & Avramidis, E. (2025). Validation of the Askisi-Lexia neuropsychological web-based screener: A neuropsychological battery for screening cognitive and phonological skills of children with dyslexia. *Applied Neuropsychology: Child*, 1–17. https://doi.org/10.1080/21622965.2025.2461192

Authors' Note

The authors thank Katja Korhonen for working as a research coordinator in the project, the numerous research assistants taking part of the data gathering at schools, and the teachers and children taking part in our study. We also thank Eva Malessa for her valuable comments on the manuscript. We would also like to thank the funders of this project. This research was funded by the Academy of Finland's Future Knowledge and Skills funding program for the project "Technology-enhanced environment for supporting reading development in all learners (ReadAll)" with grants 274050 and 274190 for the years 2014–2017. In addition, this research has been funded by Foundation Botnar, and working of J. H. for years 2018–2023 was supported by the grant 319911 from Academy of Finland; working of R. H. for years 2016–2018 was supported by the grant 277340 from Academy of Finland.

All correspondence should be addressed to
Riikka Heikkilä
Niilo Mäki Institute
PL 29, 40101 Jyväskylä, Finland
riikka.heikkila@nmi.fi

Human Technology
ISSN 1795-6889
https://ht.csr-pub.eu