frontiers in Education

ORIGINAL RESEARCH
published: 29 July 2020
doi: 10.3389/feduc.2020.00132

An Evaluation of the Efficacy of GraphoGame Rime for Promoting English Phonics Knowledge in Poor Readers

*Henna Ahmed[1], Angela Wilson[1], Natasha Mead[1], Hannah Noble[1], Ulla Richardson[2], Mary A. Wolpert[3] and Usha Goswami[1]**

[1] Centre for Neuroscience in Education, Department of Psychology, University of Cambridge, Cambridge, United Kingdom, [2] Centre for Applied Language Studies, Faculty of Humanities and Social Sciences, University of Jyväskylä, Jyväskylä, Finland, [3] Faculty of Education, University of Cambridge, Cambridge, United Kingdom

Here, we report further analysis of data drawn from a Randomized Controlled Trial (RCT) run in the United Kingdom designed to evaluate the efficacy of an adaptive software game to aid the learning of English phonics, GraphoGame Rime. We evaluate the efficacy of GraphoGame Rime for the "top half" of players in the RCT, children aged 6 to 7 years who played above the group mean play progress point (95 children). We also analyze three sub-groupings of this cohort. The GraphoGame family of games in different languages was originally designed to support children at family risk of dyslexia, hence we analyzed data for the subgroup of the GraphoGame Rime children who were struggling in school and had Individual Education Plans (IEPs). Secondly, we analyzed data from the younger children in the RCT, born in the Spring and Summer months, as international studies of GraphoGame have found the strongest effects during the first year of reading tuition and our participants were in their second year of reading tuition. Finally, we analyzed GraphoGame Rime data from players in schools rated as "requiring improvement." Schools that are found to be "requiring improvement" in the United Kingdom are encouraged to use additional teaching strategies to achieve better outcomes. GraphoGame Rime is relatively cheap to acquire and easy to implement, hence if it offers significant gains over "business-as-usual" this would be a valulable additional strategy for such schools. We find that GraphoGame Rime is more effective than "business-as-usual" in developing knowledge of English phonics for all of the groupings analyzed. We conclude that the supplementary use of GraphoGame Rime in addition to ongoing classroom literacy instruction can benefit children in learning phonic decoding and spelling skills.

Keywords: rhyme, phonics, reading software, spelling, phonological awareness

OPEN ACCESS

Edited by:
Angela Jocelyn Fawcett,
Swansea University, United Kingdom

Reviewed by:
Angeliki Mouzaki,
University of Crete, Greece
Eliane Segers,
Radboud University Nijmegen,
Netherlands

****Correspondence:***
Usha Goswami
ucg10@cam.ac.uk

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *26 February 2020*
Accepted: *30 June 2020*
Published: *29 July 2020*

Citation:
Ahmed H, Wilson A, Mead N, Noble H, Richardson U, Wolpert MA and Goswami U (2020) An Evaluation of the Efficacy of GraphoGame Rime for Promoting English Phonics Knowledge in Poor Readers. Front. Educ. 5:132. doi: 10.3389/feduc.2020.00132

INTRODUCTION

Educational technology is widely assumed to show great promise regarding cost-effective learning in classrooms, with its proponents expecting gains in many areas of the curriculum (Beddington et al., 2008). Educational technology can also provide equality of opportunity, as in theory learners in very different settings can all access the same optimized curriculum delivered under optimal conditions, for example learners in rural Africa (Ojanen et al., 2015). A pertinent example regarding equality of opportunity is computer-assisted reading instruction (CARI) technology, typically aimed at improving word recognition and phonic skills in young learners. CARIs offer one way of providing learners with repeated practice in key component skills for reading, such as acquiring and automatizing grapheme–phoneme correspondences (GPCs), in a game-like environment (Richardson and Lyytinen, 2014). However, a series of meta-analyses have indicated both that educational technology in general and CARIs in particular have a small to moderate effect on the improvement of reading skills (Blok et al., 2002; Andrews et al., 2007; Livingstone, 2012; Cheung and Slavin, 2012, 2013; Archer et al., 2014). The earliest meta-analysis by Blok et al. (2002) surveyed 42 studies using CARIs with beginning readers, and found an overall small effect size of $d = 0.19$. Blok et al. (2002) noted that the field was dogged by poor-quality studies, but that CARIs for English-speaking learners tended to generate larger effect sizes. However, Cheung and Slavin (2013) reviewed 20 studies including around 7,000 children in which English-speaking struggling readers used CARIs, and reported another small overall effect size, of $d = 0.14$. In a second meta-analysis examining all students in classrooms from kindergarten to grade 12, rather than only struggling readers, Cheung and Slavin (2012) reported a similar small effect size, of $d = 0.16$. Accordingly, it has been argued that CARIs have failed to deliver on their early promise. On the other hand, as gaming technologies improve, there is an opportunity for CARIs to improve also. As Cheung and Slavin (2013) noted, the nature of the software may determine the effectiveness of the technology.

Early CARIs were not always adaptive nor motivating for children (Deci and Ryan, 2008). Deci and Ryan (2008) argued that effective educational games needed to create a sense of relatedness to enhance motivation, and that it was important to give the child playing the game a sense of competence and autonomy. Adaptive gaming software can enhance motivation and feelings of competence, as children can move on to more difficult learning challenges once they master a particular educational level. Another important aspect regarding effective educational technology is the amount of repetition that is provided (van Gorp et al., 2014). More recent CARI technology recognizes that learning fundamental skills important for progress in literacy, such as grapheme–phoneme conversion, requires sufficient amounts of repetition to enable automatization. More recent CARIs provide the amount of practice and repetition that each individual learner needs on an individualized basis, via their inbuilt adaptive algorithms. The adaptive algorithms are designed to increase motivation and avert any boredom associated with repetition, as children only practice what they do not learn quickly. The effects of motivating digital games that include adaptive training on the development of children's reading skills have been investigated in a few prior experimental studies.

For example, two recent studies in Dutch utilized strong experimental designs. van de Ven et al. (2017) recruited 8-year-old Dutch children ($N = 60$) with mild learning disabilities, who then participated in nine sessions of 15 min each during which they played an adaptive game called Letter Prince. The game taught skills such as grapheme–phoneme conversion and semantic categorization of words via an adventure game that included several motivational elements. Children were tested at three points during the study on standardized literacy measures. van de Ven et al. (2017) used a staggered design, whereby half the group were randomly allocated to play Letter Prince between test sessions 1 and 2, and the other half played Letter Prince between test sessions 2 and 3. Playing Letter Prince was found to significantly improve children's pseudoword and text-reading fluency, but not their (self-reported) reading motivation. As another example, van Gorp et al. (2016) investigated the effectiveness of a game called Reading Race for 8-year-old Dutch children ($N = 62$), who were selected because of their poor word-decoding skills (below the 25th percentile). Their study used a pretest and post-test retention design with a waiting list control group. The game aimed to improve reading efficiency by giving players tasks based on real and then also pseudowords, including decoding the words and making semantic categorizations. Gaming elements were used for motivation, for example players could progress from driving a submarine to a rocket, and gaming was adaptive, to encourage the player to produce faster and faster responses. Playing Reading Race for a total of 5 h over 5 weeks significantly increased the children's word decoding efficiency, and these benefits were retained 5 weeks after the intervention had ended.

These Dutch games have only been developed for one language. An adaptive research-driven CARI technology that has been developed for many languages is GraphoGame, a family of CARIs developed by a team in Finland. GraphoGame was first devised for the Finnish language, and has now been adapted for over 20 languages including non-alphabetic languages like Chinese (Lyytinen et al., 2009; Borleffs et al., 2017; Li et al., 2017). While many individual experimental studies have reported significant beneficial effects from playing GraphoGame (e.g., Brem et al., 2010, German; Saine et al., 2011, Finnish; Kyle et al., 2013, English), a recent meta-analysis of 19 GraphoGame studies in a range of languages concluded that GraphoGame was only effective in certain educational contexts, with effect sizes for word reading ranging from −1.07 to 1.58 (McTigue et al., 2019). According to McTigue et al.'s (2019) hypotheses, the contexts expected to be relevant to efficacy included the complexity of the orthography being learned (for example, the consistency of the GPCs), the duration of the intervention, and the level of supportive adult interaction. However, the meta-analysis only found significant effects for one hypothesized factor, the level of supportive adult interaction. GraphoGame studies with high levels of adult interaction showed an average positive effect size of 0.48.

While it is encouraging that GraphoGame can achieve relatively high effect sizes, the aim of CARIs is that they should be effective without requiring constant attention from the classroom teacher or other educational instructor. Accordingly, in the current study we report further analysis of data relevant to the efficacy of the English version of GraphoGame, GraphoGame Rime (hereafter GG Rime). The data were originally collected as part of a Randomized Controlled Trial (RCT) of GG Rime carried out in the United Kingdom. The children in this RCT typically played GG Rime solo, without adult encouragement and feedback. Accordingly, the level of supportive adult interaction was low. Therefore, the data enable a relatively pure test of the value of GraphoGame CARIs as educational technologies that enable equality of opportunity and cost-effective learning without a high level of adult supportive interaction.

The present study was funded by the Education and Neuroscience scheme, a collaboration between the United Kingdom Education Endowment Foundation (EEF) and United Kingdom Wellcome Trust that was set up to enable RCTs to test promising educational interventions based on educational neuroscience. The scheme allocated independent evaluators to selected projects, and the independent evaluator selected the children for the RCT, allocated them to the participant groups, and selected the efficacy measures. The independent evaluators also post-tested participating children on their selected measures, to assess efficacy blind to who had received the intervention. The full evaluation of GG Rime is available at https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/graphogame-rime/.

The trial involved 398 Year 2 (age 6–7 years) pupils in 15 United Kingdom primary schools, all of whom entered the trial because they had failed the United Kingdom government phonics screening check, a statutory assessment of phonic decoding ability taken by all Year 1 (age 5–6 years) pupils in England at the end of the first school year (hereafter Phonics Check). As the children who failed the Phonics Check were in their second year of reading instruction during the current study, and had not been progressing as expected under universal tuition, we believe that they would be equivalent to Tier 2 children in the United States. Accordingly, both the "business-as-usual" control group and the GG Rime group in the RCT would be classified as Tier 2 children. In the United Kingdom children who fail the Phonics Check are given extra literacy support. The GG Rime group received individualized computerized instruction, typically being left to play GG Rime solo in a corner of the classroom or in the school library. The "business-as-usual" control group received phonics tuition from a classroom teacher or teaching assistant, either in small groups or one-on-one, using a range of literacy materials, for example Reading Recovery. The GG Rime group received the computerized training instead of this phonics tuition, typically at the same time during the school day, most usually during the Literacy Hour that is a daily event in United Kingdom primary schools. Formal reading instruction in Year 1 in the United Kingdom has to be based on synthetic phonics. Accordingly, an intervention like GG Rime cannot be given during the first year of schooling, as it is not based on synthetic phonics. All children in the current trial had received reading instruction using synthetic phonics during Year 1, but had not progressed well in their reading, as they failed the Phonics Check.

Half of the 398 Year 2 children in the trial were randomly assigned to the GG Rime intervention by the independent evaluators, the National Foundation for Educational Research (NFER). The other half were assigned to "business-as-usual" literacy instruction as described above. Overall 361 children provided final data. The primary outcome measures were chosen by the independent evaluator and comprised the New Group Reading Test (NGRT) and the Single Word Spelling Test (SWST, both GL Assessment). These tests were administered by independent test administrators provided by NFER who were blind to group status, within a month of the intervention ending. The EEF report (Worth et al., 2018) concluded that the improvements made by the children playing GG Rime solo as assessed by the NGRT and SWST were equivalent to the improvements made by children receiving direct teaching via "business-as-usual." It also concluded that the teachers and teaching assistants involved found the GG Rime intervention easy to set up and to implement for their children. The report noted that teachers, senior leaders and pupils considered GG Rime highly engaging, motivational and enjoyable.

In the current study, we provide further analyses of the GG Rime intervention data. These analyses are not conducted blind to the intervention groupings, but they add value to the RCT as they evaluate efficacy for children who played the game consistently enough to progress through at least half of the gaming streams. McTigue et al. (2019) noted that in their meta-analyses, GraphoGame was least effective when it was used in strict accordance with its prescriptions, that is when it was played in solitude by young children, nevertheless this was typically the case in the current study. In the current study, the children were playing GG Rime by themselves, indeed most typically they were put in a corner of the classroom with a computer or in the school library with the other GG Rime children from their class, while the teacher and teaching assistants focused on interacting with the other children in the classroom who were participating in Literacy Hour. GG Rime automatically collects attainment and playing time data. Perhaps unsurprisingly given that they were left by themselves, these data showed that many children in the RCT were not actually playing the game when logged in. The game takes players through 25 streams of phonic knowledge, and the mean progression point reached by the whole GG Rime group (N = 195) was Stream 16, level 5, just over half-way through the game (range Stream 2, level 2 to stream 25 level 7, the final level). Playing time was very variable, with some children in the RCT playing for as little as 133 min (just over 2 h) in the entire school term. This appears to suggest that they were using their time alone on the computer to do other activities. Significant gains in phonics learning reported in small-scale experimental studies of GG Rime were accrued after playing daily for 12 weeks for 10–15 min a day (Kyle et al., 2013). Accordingly, it may be hypothesized that children in the RCT who only played through a few streams did not receive sufficient solitary exposure to the game to affect their phonic learning.

To get nearer to the level of playing density typical of smaller-scale experimental studies, we decided to analyze only the outcome data from the "top half" of GG Rime players in the RCT, the 95 children who played the game beyond the mean play progress point for the whole intervention cohort. We then compared this "top half" of players to the entire "business-as-usual" group, who were all receiving direct literacy tuition from a teacher or teaching assistant using a range of different phonics interventions. We did consider taking a random half of the control group for analysis purposes, but we decided it was fairer to utilize the entire control group, as the schools were using such a wide range of different phonics schemes and as this group did not have the opportunity to do something other than literacy during literacy hour. Regarding the GG Rime players, our reasoning was that this "top half" of players may have received sufficient independent and solitary exposure to the game to learn English phonics. As shown in **Table 1**, they had spent on average 8.5 h playing the game.

In our previous small-scale experimental studies of the efficacy of GG Rime (Kyle et al., 2013; Bhide et al., 2013; Patel et al., 2018), the experimenters ensured that the children in the intervention group received sufficient exposure to the game. For example, Kyle et al.'s (2013) study compared children playing GG Rime to children playing an alternative English language version of GraphoGame, GraphoGame Phoneme. Small groups of children played either GG Rime or GG Phoneme daily for 12 weeks under supervision, or formed an untreated control group. The GG Rime group showed medium effect sizes for reading (0.66, 0.53) and large effect sizes for spelling and phonic decoding (0.91, 1.43). The GG Phoneme group showed small effect sizes for reading (0.22, 0.43) and spelling (0.45) and a medium effect size for phonic decoding (0.60). A reading intervention is usually considered effective if effect sizes are greater than 0.13–0.23 (Torgesen et al., 2001). An effect size of 0.2 is considered small, an effect size of 0.5 is considered medium, and an effect size of 0.8 is considered large. While Kyle et al. (2013) reported large effect sizes for spelling and phonic decoding, the experimenter was present throughout the gaming periods, thereby guaranteeing time on task and also providing general encouragement to the players. This may have been another factor contributing to the large effect sizes found by Kyle et al. (2013) (see McTigue et al., 2019).

TABLE 1 | Group characteristics expressed as mean and (SD) for GG Rime children who played above the group mean progress point and the "Business as usual" control group.

	GG Rime	Control
N	95	196
Age (years; months)	6;7 (3.2)	6;7 (3.4)
Phonics Check Y1	19.6 (8.6)	18.9 (9.3)
NGRT pretest raw	9.0 (4.9)	8.9 (5.2)
TOWRE word pretest raw	20.83 (10.8)	18.9 (11.0)
TOWRE nonword pretest raw	9.2 (5.0)	9.2 (5.3)
Playing time in minutes	509.0 (145)	–
Playing days	34.9 (9.8)	–
Level reached in game (play progress in streams and levels)	Stream 21, level 5 (4.3 streams, equating to 43% of the way through a level)	–

Here, we also report on three further sub-groupings of the "top half" of GG Rime children in the RCT that are important for deciding which educational contexts may reap the most benefit from CARI technologies. The GraphoGame family of phonics games in different languages was originally designed to support children at family risk of dyslexia (Lyytinen et al., 2007), providing repeated practice of letter-sound correspondences (which children at risk for reading difficulties are slow to acquire in all languages). Accordingly, we also analyzed data for the subgroup of the GG Rime children who were struggling in school and had Individual Education Plans (IEPs, *N* = 15). IEPs are documents developed under United Kingdom law which set out individually designed education plans for children with special educational needs. As international studies of GraphoGame have found the strongest effects when the games are administered during the first year of reading tuition, we also analyzed data for the sub-group of children born in the Spring and Summer months (*N* = 51), the younger half of our cohort. As noted, we could not administer GG Rime during Year 1, the first year of formal reading tuition in the United Kingdom, because United Kingdom law mandates a "synthetic phonics" approach and GG Rime does not use synthetic phonics. However, children born in the Spring and Summer months are younger than other Year 2 children, and hence closer in age to those children who might be expected (on international comparisons) to benefit most from GG Rime. Children born in the Spring and Summer months are also potentially disadvantaged by being much younger when they have to sit the Phonics Check test in the United Kingdom. Sitting the test when younger may mean that developmental immaturity rather than risk for intrinsic reading difficulties may contribute to lower attainment. Finally, we analyzed data for the sub-group of children attending schools found to be "requiring improvement" by OFSTED (*N* = 27), the United Kingdom Office for Standards in Education. OFSTED is the body that inspects schools in England. Schools that are found to be "requiring improvement" are encouraged to use additional teaching strategies to those currently employed to achieve better outcomes for their students. If a gaming App that is relatively cheap to acquire and easy to implement offers significant gains over "business-as-usual" in the classroom for such schools, then this information is of educational and practical value to teachers and headteachers.

MATERIALS AND METHODS

Design

The RCT (Worth et al., 2018) was a two-armed pupil-randomized controlled trial, carried out over two consecutive school years with a relatively large number of pupils (398 randomized; 361 in the final analysis). The final analysis included primary outcome data for pupils from all 15 schools involved in the trial. Less than ten per cent of participating pupils had missing data and NFER considered that this attrition was likely to be unbiased (page 5, Worth et al., 2018). Training, technical support and some delivery support (e.g., fixing school firewall problems) for GG Rime was provided by the current authors. In addition to the primary outcome measures (NGRT, SWST), a process evaluation used case-study visits, telephone interviews and analysis of data on pupils' usage of the game to capture the perceptions and experiences of participating teaching staff and pupils (Worth et al., 2018). Finally, a decoding test widely used in experimental research was administered to the cohort by ourselves, the Test of Word Reading Efficiency (TOWRE, Torgesen et al., 1999). The TOWRE consists of two subtests measuring speeded decoding of words (SWE, Sight Word Efficiency) and nonwords (PDE, Phonetic Decoding Efficiency). All children gave their assent prior to testing, and the study was reviewed by the Psychology Research Ethics Committee of the University of Cambridge.

Participants

Three hundred and ninety eight Year 2 children aged between 6 and 7 years old participated in the study, all of whom were eligible for inclusion because they had failed the Phonics Check at the end of Year 1 (scoring 31 or less, a cut-off decided by NFER). This threshold was chosen to target the program at struggling readers and to ensure that a consistent selection threshold was applied across all the schools involved. As noted previously, we believe that all the children in our study would be classified as Tier 2 children in the United States. At the end of the study, following some movement and drop-out, data from 361 children were available for our analyses. Pretest data for the GG Rime children who had played beyond the mean progress point for the game (N = 95) and all control children (N = 196) are shown in **Table 1**. As not all children had data for all the outcome measures analyzed here, we report the relevant N for the NGRT, SWS, and TOWRE measures in the footnotes for each table. Missing outcome data ranged from 2 to 8 participants depending on task for the GG Rime group and from 6 to 20 children for the control group. As these outcome data were collected in part by NFER, the reasons for missing data are unknown, however, the most likely reason is the child's absence from school on the test day. Allocation of the children to either the GG Rime or control groups was carried out by NFER to ensure full randomization.

Assessments

Reading

The children completed two standardized assessments of reading during the Autumn and Summer terms of the 3-term United Kingdom school year. The first was the New Group Reading Test Level IB, an untimed multiple choice test with three sub-sections, Phonics, Sentence Comprehension, and Passage Comprehension. In the Phonics section (15 items), children find the word which rhymes with a target from a multiple choice selection, for example selecting 'ocean' to rhyme with 'motion' (both irregular spellings), or complete word endings and word beginnings by ticking the stem which best completes a target word, for example selecting 'al' to complete the stem 're' (to make 'real,' thereby artificially dividing a vowel digraph). In the sentence completion section (18 items), children read sentences and then choose the word which best fits a gap in the sentence ('She put the book – [under] her bed'). In the passage comprehension section (10 items), children read a passage independently and then answer multiple choice questions. There are five alternative choices for every item in each section of the NGRT, hence chance responding is 20% (8.6 items). Children are not required to read aloud when completing the NGRT as it is a multiple choice test. Pretesting with the NGRT was carried out by the authors prior to group assignment during the Autumn term, and post-testing was carried out by the independent assessors during the Summer term. The second standardized measure was the TOWRE, we used both the real word subscale (SWE) and the nonword subscale (PDE). Children were required to read aloud from a list of items graded in difficulty as many words or nonwords as they could in 45 s, as quickly and as accurately as possible. This test was administered by the authors at both pretest (prior to group assignment during the Autumn term), post-test (immediately after the independent assessors had finished their post-tests of the intervention, that is finished administering the NGRT and SWS in each school in the Summer term), and delayed post-test (3 months after the intervention, hence at the beginning of the school year following the 6-week summer vacation). The purpose of the delayed post-test was to assess whether any gains in reading persisted over the school summer holidays.

Spelling

The SWST was administered at post-test only by the independent assessors from NFER, during the Summer term. This was a spelling to dictation task which was untimed. The items begin with relatively simple words like 'on,' 'it,' and 'up,' and progress to more difficult words like 'shout' and 'team.' The test was administered according to the instruction manual.

GraphoGame Rime

GG Rime was administered during the Spring term of the United Kingdom school year, which runs from mid-January to mid-April. GG Rime is now a gaming App available from a Finnish educational technology company, Grapho Group Oy. GG Rime provides highly repetitive and individualized intervention aimed at developing phonics skills in young learners. The game is based on the intrasyllabic unit of the rime, the vowel sound and any subsequent consonants (e.g., st – AMP; cl – OCK), thought to be an important psycholinguistic unit for English-speaking children (Treiman et al., 1995). The player hears auditory targets consisting of either sounds or words and has to match these auditory targets to visual targets (letters and sequences of letters) displayed on the screen of a computer, tablet or mobile phone. The letters and letter sequences are displayed as part of different games played by the child's avatar, for example catching pirate cannonballs by clicking on them. Children progress through a series of graduated game streams (total streams 25), each of which has multiple levels (ranging from 5 to 9 levels). To keep motivation levels high, children are rewarded with tokens at the end of each level within a stream, which they save up and then spend in a "shop." The shop sells kit for their avatar. There are also word formation games to encourage spelling skills, in which children are presented with boxes containing letters or onset and rime patterns (onsets correspond to any phonemes before the vowel in a syllable) and are asked to put them into the correct order to spell target words (e.g., c – at). As the game is adaptive, the exact letters and letter sequences practiced by different players will vary depending on speed of progression through the game. Overall the game teaches GPCs, but using methods based on rhyme families.

GG Rime uses a success criterion of at least 80% for each level before children are able to move onto the next level. If a child fails to achieve 80% accuracy on a level, they are given individualized extra training levels in which the computer automatically selects targets that the child knew and contrasts them with targets that the child did not know. The words for GG Rime were recorded by a female speaker who had a British accent. The teaching sequence in GG Rime is based primarily on orthographic rime units. Children are introduced to single letter-sound correspondences (e.g., C, A, T, N), which are then blended into orthographic rime units (-at, -an), and then into CVC words (c-at, c-an). For example, in Stream 1 a small set of seven single phonemes and graphemes are introduced (C, S, A, T, P, I, N), and the children are told "Let's put these sounds together to make rime units." The children are then told "Now let's put another sound in front of the rime units you have just played with," and CVC words like *cat* and *tin* are created by showing blending of $c + at$ and $t + in$. The children are also reinforced on the GPCs in these CVC words ("The sounds in *tin* are *t*, *i*, *n*"). Subsequently, orthographic rimes that are not also real words are created, like *op* and *ap*, enabling creation of CVC words like *top* and *cap*. So the primary teaching sequence is to show a child some GPCs ("sounds"), to blend these GPCs into rimes, to blend onsets onto these rimes to create words (the term "onset" was not used in the game, onsets were called "sounds"), and then to segment the words back into GPCs.

The use of rhyme families enables GG Rime to highlight the higher-level statistical consistencies in the English orthography that are present when GPCs are considered in the context of the orthographic rime unit. The rhyme family format means that in GG Rime, GPC information is always linked to oral rhyming patterns (hence rhyme awareness is trained at the same time as phoneme awareness). Rhyme families are not taught exhaustively, rather 4–8 members of a particular family are introduced, and the child is then left to infer for herself that words with analogous orthographic rimes that might be subsequently encountered during classroom reading and spelling activities would be similar. The streams in the game begin with CVC items from the most consistent and most dense rime phonological neighborhoods of English (De Cara and Goswami, 2002), taking into account word frequency and orthographic consistency. Later streams introduce CCVC and CVCC words (e.g., 'bring,' 'sting,' Stream 7; 'best,' 'quest,' Streams 8–10).

Procedure

The RCT compared outcomes for pupils who were intended to spend 10–15 min each day for 12 weeks of the school Spring term playing GG Rime on a computer during literacy lessons in a quiet corner of the classroom with pupils from the same classes who received "business-as-usual" direct literacy tuition during these lessons. The risk of contamination was deemed small by NFER as the intervention could be used by intervention pupils with minimal risk of being used by control pupils in the same class. The intervention being tested was use of a game, so teachers' training about the intervention was, on its own, deemed unlikely by NFER to have an influence on control pupils' learning.

Fidelity to the Program

Fidelity to the GG Rime intervention program was measured by the Finnish GraphoGame team, who provided detailed logs including the time spent by each participant in playing GraphoGame and their progress through the game. The log feature enables individualized assessment of learning, and is intended to allow the teacher to identify streams in the game which are causing difficulty and to decide whether to provide extra (game-based or non-game) reinforcement. This feature also provides one index of whether a pupil using a computer in a quiet corner of the classroom is in fact playing GG Rime or whether they are logged into the computer but are not in fact playing the game and instead doing something else.

RESULTS

Descriptive Statistics

Following Worth et al. (2018), raw scores were used for all outcome analyses. Also following Worth et al. (2018), we used independent samples *t*-tests, two-tailed and uncorrected. This was done to enable a direct comparison between the tables in the NFER report and the current paper. Hedges g was used to compute effect sizes for the *t*-tests, as this method corrects for unequal group sizes. In addition, we compared relative progress by group taking into account children's pretest performance via repeated measures ANOVAs. We used 2×2 (Group [GG Rime, "business-as-usual"] × Test [pretest, post-test]) repeated measures ANOVAs for the NGRT, the TOWRE SWE and TOWRE PDE (word and nonword scales), in order to compare progress by group immediately after the intervention ended. We used 2×3 (Group [GG Rime, "business-as-usual"] × Test Session [pretest, post-test, delayed post-test]) repeated measures ANOVAs for the TOWRE data, where we additionally had delayed post-test scores available. The 2 × 3 ANOVAs enabled us to compare retention of learning by each group over the school summer vacation after pretest performance had been taken into account. All statistical analyses were carried out using IBM SPSS Statistics version 25.

GG Rime Players in the "Top Half" of the Sample

Inspection of **Table 1** shows that both groups of children were performing at chance levels on the NGRT at pre-test, scoring on average nine items (chance = 8.6 items, GG Rime group, *t*[1,94] = 0.7; Control group, *t*[1,196] = 0.8). The means and standard deviations for the outcome measures at post-test and delayed post-test are presented in **Table 2**. Inspection of **Table 2** reveals that the GG Rime group showed larger absolute

TABLE 2 | Outcome data (raw scores) for GG Rime children playing above the group mean play progress point and the "Business as usual" control group.

	GG Rime	Control
NGRT post-test[1]	14.18 (7.9)	13.97 (7.7)
SWST[2]	17.77 (5.9)	16.80 (7.0)
TOWRE word post-test[3]	31.9 (14.2)	30.0 (14.8)
TOWRE nonword post-test[4]	13.5 (7.3)	12.2 (6.9)
TOWRE word delayed post-test[5]	34.9 (15.3)	33.5 (17.1)
TOWRE nonword delayed post-test[6]	16.7 (8.1)	16.2 (8.8)
Phonics Check Y2	32.18 (7.3)	31.46 (8.8)

[1]GG Rime = 93 children, control 176 children; [2]GG Rime = 93 children, control = 179 children; [3]GG Rime = 95 children, control = 193 children; [4]GG Rime = 94 children, control = 193 children; [5]GG Rime = 88 children, control = 186 children; [6]GG Rime = 88 children, control = 185 children.

TABLE 3 | Outcome data (raw scores) for GG Rime children who had IEPs and the "Business as usual" control group who had IEPs.

	GG Rime	Control
NGRT pretest	7.47 (2.0)	7.28 (3.8)
NGRT post-test[1]	8.87 (4.6)	9.65 (4.6)
SWST[2]	14.13 (6.9)	9.59 (6.5)*
TOWRE word pretest	14.3 (6.7)	10.4 (8.1)
TOWRE word post-test[3]	23.1 (11.8)	16.7 (13.1)
TOWRE nonword pretest	6.1 (3.4)	5.8 (6.0)
TOWRE nonword post-test[3]	7.6 (5.0)	6.8 (6.3)
TOWRE word delayed post-test[4]	25.9 (16.4)	18.0 (16.3)
TOWRE nonword delayed post-test	9.9 (5.6)	9.4 (7.5)

*[1]GG Rime = 15 children, control 26 children; [2]GG Rime = 15 children, control = 27 children; [3]GG Rime = 15 children, control = 28 children; [4]GG Rime = 14 children, control = 27 children. *$p < 0.05$.*

scores on all outcome measures at post-test. At post-test, group performance on the NGRT was now significantly above chance, GG Rime group, $t(1{,}93) = 6.8$, $p < 0.001$; Control group, $t(1{,}175) = 9.2$, $p < 0.001$. None of the comparisons by group were significant when t-tests were computed, however, t-tests do not take account of pretest differences in performance.

In order to compare progress by group on the TOWRE with progress on the NGRT, three 2×2 (Group × Test) repeated measures ANOVAs were run, taking the raw scores in each case (NGRT, TOWRE SWE, TOWRE PDE) as the dependent variable. The critical interaction between Group and Test only reached significance for the TOWRE PDE measure, $F(1{,}285) = 5.0$, $p = 0.027$, $\eta^2_p = 0.017$. The NGRT and TOWRE SWE data did not show the critical interaction between Group and Test, but did each show a significant main effect of Test, NGRT $F(1{,}267) = 170.5$, $p < 0.001$, $\eta^2_p = 0.390$; TOWRE SWE $F(1{,}285) = 462.7$, $p < 0.001$, $\eta^2_p = 0.619$, showing that real word reading progressed to an equivalent extent for both groups. Accordingly, while both groups progressed equally for real word reading, playing GG Rime led to significantly more progress than "business-as-usual" regarding phonic decoding, with a small effect size. In order to analyze whether progress on the TOWRE measures was maintained over the school summer holidays, two 2×3 (Group × Test Session) repeated measures ANOVAs were then run, taking the raw scores on the TOWRE word subscale (SWE) or the nonword subscale (PDE) at pretest, post-test and delayed post-test, respectively, as the dependent variable. The critical interaction between Group and Test Session did not reach significance for either analysis, SWE $F(2{,}542) = 0.2$; PDE $F(2{,}540) = 1.5$, respectively. Hence, the TOWRE data show that the "top half" of children who played GG Rime showed significantly enhanced phonic decoding skills at the end of the RCT compared to "business-as-usual." These gains were not maintained over the school summer holiday, however.

Children With Individual Education Plans

Children receiving the GG Rime intervention who played above the group mean playing point and who also had IEPs in place ($N = 15$) were then compared to the control group children who had IEPs ($N = 29$) for the same outcome measures. The results are shown in **Table 3**. Again, the table shows that the children who played GG Rime in general showed better performance in the outcome measures than the children receiving "business-as-usual." The only measure to reach significance by group was the spelling measure (SWST), $t(40) = 2.13$, $p < 0.05$ (Hedges' $g = 0.68$). As the spelling test was administered blind to participant grouping, this significant enhancement from GG Rime can be considered reliable. Repeated measures ANOVAs, three 2×2 (Group × Test) and two 2 × 3 (Group × Test Session), were also run for the NGRT, TOWRE SWE, and TOWRE PDE data, respectively. The critical Group × Test and Group × Test Session interactions were not significant for any of these analyses. Hence for children with IEPs in place, playing GG Rime only significantly enhanced spelling skills in comparison to business-as-usual, and this result showed a medium effect size.

Children With Spring or Summer Births

Children born in the Spring or Summer months in the United Kingdom have frequently spent less time in school than those born in Autumn and Winter, and are developmentally less mature. Such children in the "top half" of players ($N = 51$) were compared to the control group children who had been born in the Spring or Summer months ($N = 112$). This comparison is of interest as the GraphoGame family of games across languages are intended to be used in the earliest phase of schooling, supplementing initial reading instruction, and so the younger children in our cohort may be more likely to accrue benefit from playing GG Rime. The results are shown in **Table 4**. Again, the table shows that the children who played GG Rime in general made more progress in the outcome measures than the children receiving "business-as-usual." For the t-tests, none of the outcome measures differed significantly at pretest, but at post-test the GG Rime group showed significantly better performance for the TOWRE word measure, $t(150) = 2.05$, $p < 0.05$ (Hedges' $g = 0.36$), with a small effect size. Further, this advantage was maintained at delayed post-test, $t(150) = 2.04$, $p < 0.05$ (Hedges' $g = 0.36$), suggesting that gains in real word reading persisted over the school summer holidays. The GG Rime players also showed significantly better performance in the delayed nonword post-test, $t(150) = 2.02$, $p < 0.05$, Hedges' $g = 0.36$, again with a

TABLE 4 | Outcome data (raw scores) for GG Rime children born in the Spring and Summer months and the "Business as usual" control group born in these months.

	GG Rime	Control
NGRT pretest	9.06 (5.1)	8.65 (5.2)
NGRT post-test[1]	14.30 (8.6)	12.97 (7.6)
SWST[2]	18.16 (6.1)	15.92 (7.0)$^{+}$
TOWRE word pretest[3]	21.1 (11.6)	17.1 (10.5)*
TOWRE word post-test[4]	32.8 (14.2)	27.5 (15.0)*
TOWRE nonword pretest[3]	9.2 (5.6)	8.5 (5.0)
TOWRE nonword post-test	13.8 (7.9)	11.3 (6.5)*
TOWRE word delayed post-test[5]	37.0 (16.4)	30.5 (17.4)*
TOWRE nonword delayed post-test[6]	18.5 (8.7)	15.2 (9.0)*

*[1] GG Rime = 50 children, control 98 children; [2] GG Rime = 50 children, control = 100 children; [3] GG Rime = 50 children, control = 112 children; [4] GG Rime = 51 children, control = 110 children; [5] GG Rime = 46 children, control = 108 children; [6] GG Rime = 46 children, control = 107 children. *$p < 0.05$, $^{+}p < 0.06$.*

small effect size. They thus appeared to maintain their gains in nonword reading over the school summer break.

In order to compare progress by group on the TOWRE with progress in the NGRT once pretest performance was taken into account, three 2 × 2 (Group × Test) repeated measures ANOVAs were run, taking the raw scores in each case (NGRT, TOWRE SWE, TOWRE PDE) as the dependent variable. The critical interaction between Group and Test only reached significance for the TOWRE PDE measure, $F(1,158) = 5.6$, $p = 0.020$, $\eta_p^2 = 0.034$. Accordingly, playing GG Rime led to significantly greater gains in phonic decoding skills immediately following the end of the intervention. The NGRT and TOWRE SWE data did not show significant interactions between Group and Test, but did each show a significant main effect of Test, NGRT $F(1,146) = 73.9$, $p < 0.001$, $\eta_p^2 = 0.336$; TOWRE SWE $F(1,158) = 251.6$, $p < 0.001$, $\eta_p^2 = 0.614$. Hence when pretest performance was taken into account, both groups improved in real word reading to a similar extent. In order to analyze whether progress on the TOWRE measures was maintained over the school summer holidays, two 2 × 3 (Group × Test Session) repeated measures ANOVAs were then run, taking the raw scores on the TOWRE word subscale (SWE) or the nonword subscale (PDE) at pretest, post-test and delayed post-test, respectively, as the dependent variable. The critical interaction between Group and Test Session was significant for the nonword analysis, $F(2,300) = 4.6$, $p = 0.011$, $\eta_p^2 = 0.029$, but did not reach significance for the word analysis, $F(2,302) = 2.3$. The data show that for younger children with Spring and Summer birthdays, playing GG Rime did significantly enhance phonic decoding skills at the end of the intervention, and further that these gains were maintained over the school summer holidays. Performance in the SWST was also higher at post-test for the GG Rime children, and this group difference approached significance, $t(148) = 1.93$, $p = 0.055$ (Hedges' $g = 0.33$), showing a small effect size.

Children in Schools Rated as Requiring Improvement by OFSTED

Some of the primary schools who participated in the RCT had been rated as "requiring improvement" by OFSTED. As the GG Rime App is relatively cheap (∼$10 per head at the time of writing), it could be a very useful addition to classroom literacy tuition for such schools. Accordingly, we also analyzed the data for children who played beyond the mean playing point in GG Rime who were attending such schools ($N = 27$), and compared them to the "business-as-usual" control children in the same schools ($N = 81$). The results are shown in **Table 5**. Once more, the data show that the children who played GG Rime made more progress in the outcome measures than the children receiving "business-as-usual." None of the outcome measures differed significantly at pretest, post-test or delayed post-test on the t-test comparisons. However, the t-tests do not take pretest performance into account.

TABLE 5 | Outcome data (raw scores) for GG Rime children who were attending schools rated by OFSTED as "requiring improvement" and the "Business as usual" control group attending these schools.

	GG Rime	Control
NGRT pretest	8.78 (5.4)	8.81 (5.4)
NGRT post-test[1]	13.22 (7.9)	12.58 (7.5)
SWST[2]	17.89 (6.3)	14.88 (7.2)$^{+}$
TOWRE word pretest[3]	19.0 (12.0)	17.3 (10.9)
TOWRE word post-test[4]	32.1 (14.8)	27.5 (14.8)
TOWRE nonword pretest[3]	9.0 (6.0)	9.1 (5.2)
TOWRE nonword post-test [5]	14.0 (7.8)	11.6 (6.9)
TOWRE word delayed post-test[6]	36.8 (18.2)	30.7 (16.9)
TOWRE nonword delayed post-test[7]	19.00 (10.3)	15.71 (9.0)

[1] GG Rime = 27 children, control 72 children; [2] GG Rime = 27 children, control = 74 children; [3] GG Rime = 26 children, control = 81 children; [4] GG Rime = 27 children, control = 80 children; [5] GG Rime = 27 children, control = 81 children; [6] GG Rime = 25 children, control 77 children; [7] GG Rime = 25 children, control 76 children. $^{+}p < 0.06$.

In order to compare progress by group on the TOWRE with progress in the NGRT, three 2 × 2 (Group × Test) repeated measures ANOVAs were again run, taking the raw scores in each case (NGRT, TOWRE SWE, TOWRE PDE) as the dependent variable. The critical interaction between Group and Test reached significance for both the TOWRE SWE measure, $F(1,104) = 4.2$, $p = 0.044$, $\eta_p^2 = 0.039$, and for the TOWRE PDE measure, $F(1,104) = 7.2$, $p = 0.008$, $\eta_p^2 = 0.065$. Accordingly, playing GG Rime led to significantly greater gains in both real word reading skills and also phonic decoding skills immediately after the intervention. The NGRT data did not show a significant interaction between Group and Test, but did show a significant main effect of Test, $F(1,97) = 42.1$, $p < 0.001$, $\eta_p^2 = 0.302$. Thus while both groups improved equally on real word reading as assessed by the multiple choice NGRT test, playing GG Rime led to significantly more progress than "business-as-usual" for the TOWRE measures that required reading aloud, for both real word reading and phonic decoding.

In order to analyze whether progress on the TOWRE measures was maintained over the school summer holidays, two 2 × 3 (Group × Test Session) repeated measures ANOVAs were then run, taking the raw scores on the TOWRE word subscale (SWE) or the nonword subscale (PDE) at pretest, post-test and delayed post-test, respectively, as the dependent variable.

The critical interaction between Group and Test Session was significant for the nonword analysis, $F(2,198) = 4.2$, $p = 0.017$, $\eta_p^2 = 0.040$, but did not reach significance for the SWE analysis, $F(2,200) = 2.4$. For children in schools rated by OFSTED as requiring improvement, therefore, playing GG Rime significantly enhanced phonic decoding skills over "business-as-usual" at the end of the intervention, and in addition these gains were maintained over the school summer holidays. The GG Rime group also showed better performance than the control group for the SWST and this group difference approached significance, $t(99) = 1.92$, $p = 0.057$ (Hedges' $g = 0.43$), showing a small effect size.

DISCUSSION

The present study provides further contextual information regarding the potential educational value of supplementing initial literacy teaching about phonics with educational technology via gaming Apps such as GraphoGame. High-quality educational technology can enable cost-effective learning in both school and home settings, and if it is available online, can also provide equality of opportunity, as learners in very different settings can access the same curriculum delivered under optimal conditions (Beddington et al., 2008; Ojanen et al., 2015). However, educational technologies in general have been reported to exert only small to moderate effects on the improvement of educational skills. For example, reviews of the efficacy of CARIs regarding improving the literacy skills of beginning or struggling readers have reported modest effect sizes (Blok et al., 2002, $d = 0.19$; Cheung and Slavin, 2013, $d = 0.14$). One reason for this was suggested to be the quality of many of the early CARI technologies. Cheung and Slavin (2013) specifically noted that the nature of the early software could be related to the apparent ineffectiveness of the technology. A second factor identified by previous reviews concerns the lack of training and support for the teachers who were tasked with delivering the technology (Archer et al., 2014). Indeed, Archer et al. (2014) noted that when teachers received diligent training and support, effect sizes regarding CARI for reading comprehension improved from small to medium. A further factor has been suggested to be the language of instruction. McTigue et al. (2019) conducted a meta-analysis in which they expected the complexity of the orthography being learned (for example, the consistency of the GPCs) to be critical, however, they did not find support for this hypothesis. By contrast, Blok et al. (2002) found that studies investigating educational technologies for beginning readers appeared to generate larger effect sizes when English was the target language. Nevertheless, in Cheung and Slavin's (2013) meta-analysis all the studies reviewed involved learning to read in English, yet only a small overall effect size was found ($d = 0.14$). Another factor that may be relevant is whether the intervention is delivered by experimental researchers or by classroom teachers. Small-scale research studies in which researchers deliver the intervention have tended to produce larger effect sizes (e.g., Kyle et al., 2013; $d = 0.66$ for reading). However, the meta-analysis by Archer et al. (2014) reported no difference in effect sizes according to whether teachers or researchers delivered the intervention. A related factor to who delivers the intervention is the level of supportive adult interaction. The meta-analysis by McTigue et al. (2019) reported a significant effect in this regard, finding larger effect sizes for GraphoGame when there was more adult interaction (average $d = 0.48$). On the other hand, the aim of CARIs is to reduce the need for adult interaction, in order to provide equality of opportunity for learners in different educational settings. Given the inconsistencies in the current CARI literature, it remains possible that high quality software would generate better effect sizes, and that more recently developed CARIs may prove to be more effective than available meta-analyses would suggest.

The recently developed CARI GraphoGame provides a case in point. GraphoGame is an educational technology developed on the basis of systematic research over the past two decades, and is now available in over 20 languages in research formats. The Finnish and English games are also available as downloadable Apps (from GraphoGroup Oy). GraphoGame is a software intervention designed to provide adaptive practice of letter-sound or character-sound correspondences, and research-based versions of GraphoGame are available both in alphabetic languages like German and Finnish, and in character-based scripts like Chinese PinYin (Richardson and Lyytinen, 2014; Borleffs et al., 2017; Li et al., 2017). GraphoGame was extended to the English orthography by utilizing rhyme analogy theory (Goswami, 1993), resulting in GraphoGame Rime (Kyle et al., 2013). Small-scale experimental studies of GG Rime have shown encouraging effects, however, a recent RCT conducted in the United Kingdom reported that playing GG Rime offered no extra benefits over "business-as-usual" for struggling second grade readers[1]. This is still a valuable finding, as it shows that GG Rime is a cost-effective alternative to providing the extra teacher-led instruction that is usually given to children in the United Kingdom who have failed the Phonics Check. However, in the current re-analyses, we provide evidence that significant extra benefits in phonics knowledge accrue from playing GG Rime. These extra benefits were found for those children in the RCT who played to at least the mean progress point for the entire intervention group.

The rationale behind analyzing this "top half" of the GG Rime intervention children was to get nearer to the level of playing density typical of smaller-scale experimental studies of GG Rime such as Bhide et al. (2013) and Kyle et al. (2013). The gaming logs that are automatically collected by the GG Rime software showed that some children in the intervention were only playing GG Rime for a few hours, rather than the exposure time of over 12 h that was the initial goal for the RCT. Hence the intervention group were being compared to "business-as-usual" children who received much more literacy intervention, as they were being taught directly by classroom teachers and teaching assistants during Literacy Hour, while the GG Rime children played on a computer. We reasoned that the "top half" of GG Rime players may have received sufficient independent

[1] https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/graphogame-rime/

and solitary exposure to the game to learn English phonics in a cost-effective manner.

The main outcome measure demonstrating phonic gains in our re-analyses was the TOWRE nonword decoding test (PDE subscale). The TOWRE is a standard measure for assessing phonic learning in experimental studies of reading acquisition, as it measures the specificity and accuracy of children's phonic decoding skills for individual items. Efficacy in the original RCT was assessed using the NGRT, a multiple choice test that is not able to measure children's phonic decoding skills in a comparable fashion. Indeed, for children who are struggling with reading at the end of Year 1, it is unlikely that irregularly spelled items like 'ocean' (the rhyming match for 'motion' among the Phonics section multiple choice options) are within the child's reading capacity. The Phonics section of the NGRT contains many such items, for example 'seize' must be rhymed with 'cheese,' and 'stressed' must be rhymed with 'guest.' For children who are struggling with reading at the end of Year 1, it is unlikely that items like 'ocean' and 'seize' are within the child's reading capacity (indeed, group NGRT performance was at chance levels for both groups at pre-test, and so could simply reflect guessing). By contrast, significant gains on the TOWRE nonword measure compared to "business-as-usual" were shown by the "top half" of GG Rime players immediately after the intervention ended, and by all sub-groupings of GG Rime players analyzed here except for the children with IEPs. For both the Spring/Summer birth cohort and for the children attending schools rated as requiring improvement by OFSTED, the significant gains in phonic decoding were maintained over the school summer holidays. Providing significant and cost-effective benefits to children attending schools rated as requiring improvement by OFSTED is a particularly important educational outcome of the current study.

Although the TOWRE was administered by the current authors and not by assessors blind to participant groupings, it is a speeded test. Arguably it is therefore less liable to unconscious tester bias, as the child only gets 45 s in which to recode items to sound. However, the data reported here should be considered as exploratory only, since further RCTs focusing on just the sub-groups considered here, along with blind administration of the TOWRE, are required to support these findings. Nevertheless, the outcome data from the TOWRE nonword measure in general converged with the spelling outcome data (SWST) administered by the independent NFER assessors (see **Tables 2–5**). For the children with IEPs, the data showed that playing GG Rime led to spelling attainment that was significantly better than "business-as-usual" (medium effect size). As the spelling test was not administered by the authors, the convergence of the spelling data with the TOWRE nonword reading data support the view that the children's gains as measured by the TOWRE do not reflect unconscious experimenter bias. Again, however, without further RCTs, these positive outcomes should be considered exploratory.

Finally, it is worth noting that the training experienced by the children who played GG Rime above the mean playing point was still comparatively brief (around 8.5 h), and so most children did not play for long enough to complete the game. The App is intended to be played on a daily basis for around 10 min a day during the first year of schooling in order to enable maximum learning gains to accrue. Previous small-scale experimental studies of GG Rime (Kyle et al., 2013; Bhide et al., 2013; Patel et al., 2018) reported mean playing times of around 11 h for GG Rime. Therefore, greater gains than those observed here could be expected if the original aims of the RCT regarding playing time had been met. For example, Kyle et al. (2013) reported that 11 h of training with GG Rime led to medium and large effect sizes for reading (0.66) and spelling (0.91), and to large effect sizes for phonological awareness (phoneme, 1.27, rime 1.0) and nonword reading (1.43). Improvements in standard score per hour of playing the game were 0.69 SS per hour (Kyle et al., 2013). In comparison, the gains reported for more personnel-intensive non-technological training programs, such as the phonological linkage program of Hatcher et al. (1994), are 0.31 SS per hour (see Hatcher, 2003).

As noted, the large effect sizes found in the GG Rime experimental studies were likely in part dependent on an experimenter being present throughout the game periods (see Kyle et al., 2013; Bhide et al., 2013; Patel et al., 2018). This guaranteed time on task and also enabled the adult to provide general encouragement (McTigue et al., 2019). Other gaming Apps, for example the CogMed working memory training App, also appear to be more effective if the child receives one-to-one support from their training aide (Holmes et al., 2010). Nevertheless, it is notable that the significant gains in phonic decoding reported here were achieved with the children playing solo, without high levels of adult support. Accordingly, the current data are also of interest regarding the larger debate around whether enthusiasts for educational technology have overpromised on the potential gains to be made for individual learners (Livingstone, 2012). The data reported here show that solo gaming with GG Rime can be very beneficial for children, particularly for younger children or children who are in schools that have been rated as requiring improvement regarding the educational provision offered. An important proviso is that the children in the "top half" of the GG Rime players may have stayed on task solo because they had better executive function skills than the children in the bottom half of players, those children who did not play for as long and who made little progress through the game. Children with lower self-regulatory skills may need adult supervision to benefit from the game. In order to assess the role of self-regulation in progressing through the game, executive function skills would need to be measured.

It is also worth reiterating that the Worth et al. (2018) report concluded that playing GG Rime led to *equivalent* reading outcomes to "business as usual" literacy tuition with the classroom instructor. Hence GG Rime is a potentially cost-effective alternative mode of tuition regarding phonic knowledge that delivers equivalent gains to normal classroom practice in the United Kingdom when embedded in a broader literacy curriculum. It also provides equality of opportunity, as the learners studied here were all accessing the same optimized curriculum delivered under equal conditions (see Ojanen et al., 2015). Regarding the overall meta-analysis conducted by McTigue et al. (2019, see their **Figure 3**), the GG Rime RCT sits right in the middle of the forest plot regarding efficacy, with an average effect size of 0. The effect sizes in the other GG Rime studies reviewed by McTigue et al. (2019) were all positive (Kyle et al., 2013; Bhide et al., 2013; Patel et al., 2018). Accordingly, GG Rime does appear to be an effective CARI for promoting phonic decoding skills when learning English GPCs, despite the inconsistent nature of the English orthography.

The current study has a number of important limitations. Firstly, the sample size was smaller than in the main RCT. Sample size should be increased in future studies. Secondly, progress through the game is only one index of sufficient exposure for learning, and the results could have been different if a different measure had been selected. Thirdly, the TOWRE was administered by the current authors rather than by independent assessors. This may have inflated GG Rime children's performance. In addition, our decision to compare the "top half" of GG Rime children to the entire cohort of controls may have introduced inadvertent bias, as although the prior attainment of these children did not differ compared to that of the entire cohort of controls (see **Table 1**), they may have had certain other characteristics that were not apparent in the pretesting that underpin the differences reported here. In addition, our participants were children who played GG Rime during the *second* year of reading tuition at school and not the first year of tuition. The children had all experienced synthetic phonics teaching during their first year of reading instruction. This was an unavoidable feature of the RCT due to current United Kingdom government policy. Finally, the total intervention received was limited (an average of 8.5 h spent playing the game, across 12 weeks), and so very few participants were able to complete all 25 streams in the game and hence to learn the majority of the phonic "rules" of English as intended by the game's designers. In future work, it would be optimal to ensure that children play daily until they have played their way through the entire game, and then assess their phonic learning via an RCT. This would enable the strongest experimental test of the efficacy of playing GG Rime.

CONCLUSION

The current study suggests that young learners of the English orthography show significant benefits in learning both phonic decoding skills and spelling skills from the supplementary use of GG Rime in addition to ongoing classroom literacy instruction. For some learners, these significant gains were maintained over the school summer holiday, being retained for at least 3 months. Accordingly, CARIs such as GraphoGame can be valuable for supporting the teaching of different spelling systems to young children in different educational contexts (see GraphoGame data from Finnish, German and French studies, Lyytinen et al., 2009; Brem et al., 2010; Saine et al., 2011; Ruiz et al., 2017). CARIs offer evidence-based technological tools that are cost-effective, and support classroom teachers by providing individualized instruction (Beddington et al., 2008; Connor et al., 2009). The RCT data analyzed here suggest that such individualized instruction regarding practice in one of the component skills of reading, phonics learning, may be of particular benefit to younger learners in the United Kingdom born in the Spring and Summer months, and to children attending schools rated by OFSTED as "requiring improvement." These learning benefits were accrued despite playing GG Rime without high levels of adult interaction, and despite the ambiguous nature of GPCs in the English orthography. Nevertheless, we would endorse the conclusions of McTigue et al. (2019) that literacy is a social undertaking, and so CARIs ideally need to be nested within authentic and social reading activities in schools or in homes for their full potential for learners to be realized.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Cambridge Psychology Research Ethics Committee. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

HA and AW contributed to the investigation, data curation, and formal analysis. HA wrote the original draft of manuscript. NM and HN contributed to the investigation and data curation. UR contributed to the software and resources. MW and UG contributed to the conceptualization and funding acquisition. UG contributed to the methodology, supervision, project administration, and writing original draft and revision. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the Education Endowment Foundation/Wellcome Trust Education and Neuroscience scheme (grant number RG 78836).

ACKNOWLEDGMENTS

We thank the developing team of the "GraphoGame" at the University of Jyväskylä, and the Finnish support team for the RCT, especially Iivo Kapanen and Mika Halttunen. We also thank NFER and the children, families and schools who participated in the study.

REFERENCES

Andrews, R., Freeman, A., Hou, D., McGuinn, N., Robinson, A., and Zhu, J. (2007). The effectiveness of information and communication technology on the learning of written English for 5- to 16-year-olds. *Br. J. Educ. Technol.* 38, 325–336. doi: 10.1111/j.1467-8535.2006.00628.x

Archer, K., Savage, R., Sanghera-Sidhu, S., Wood, E., Gottardo, A., and Chen, V. (2014). Examining the effectiveness of technology use in classrooms: a tertiary meta-analysis. *Comp. Educ.* 78, 140–149. doi: 10.1016/j.compedu/2014.06.001

Beddington, J., Cooper, C. L., Field, J., Goswami, U., Huppert, F. A., Jenkins, R., et al. (2008). The mental wealth of nations. *Nature* 455, 1057–1060. doi: 10.1038/4551057a

Bhide, A., Power, A. J., and Goswami, U. (2013). A rhythmic musical intervention for poor readers: a comparison of efficacy with a letter-based intervention. *Mind Brain Educ.* 7, 113–123. doi: 10.1111/mbe.12016

Blok, H., Oostdam, R., Otter, M. E., and Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: a review. *Rev. Educ. Res.* 72, 101–130. doi: 10.3102/00346543072001101

Borleffs, E., Glatz, T. K., Daulay, D. A., Richardson, U., Zwarts, F., and Maassen, B. A. M. (2017). GraphoGame SI: the development of a technology-enhanced literacy learning tool for Standard Indonesian. *Eur. J. Psychol. Educ.* 33, 595–613. doi: 10.1007/s10212-017-0354-9

Brem, S., Bach, S., Kucian, K., Guttorm, T. K., Martin, E., Lyytinen, H., et al. (2010). Brain sensitivity to print emerges when children learn letter-speech sound correspondences. *Proc. Natl. Acad. Sci. U.S.A.* 107, 7939–7944. doi: 10.1073/pnas.0904402107

Cheung, A. C., and Slavin, R. E. (2012). ow features of educational technology applications affect student reading outcomes: a meta-analysis. *Educ. Res. Rev.* 7, 198–215. doi: 10.1016/j.edurev.2012.05.002

Cheung, A. C., and Slavin, R. E. (2013). Effects of educational technology applications on reading outcomes for struggling readers: a best-evidence synthesis. *Read. Res. Q.* 48, 277–299. doi: 10.1002/rrq.050

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., et al. (2009). Individualizing student instruction precisely: effects of child x instruction interactions on First Graders' Literacy Development. *Child Dev.* 80, 77–100. doi: 10.1111/j.1467-8624.2008.01247.x

De Cara, B., and Goswami, U. (2002). Similarity relations among spoken words: the special status of rimes in English. *Behav. Res. Methods* 34, 416–423. doi: 10.3758/bf03195470

Deci, E. L., and Ryan, R. M. (2008). Facilitating optimal motivation and psychological wellbeing across life's domains. *Can. Psychol.* 49, 14–23. doi: 10.1037/0708-5591.49.1.14

Goswami, U. (1993). Toward an interactive analogy model of reading development: decoding vowel graphemes in beginning reading. *J. Exp. Child Psychol.* 56, 443–475. doi: 10.1006/jecp.1993.1044

Hatcher, P. J. (2003). Reading intervention: a 'conventional' and successful approach to helping dyslexic children acquire literacy. *Dyslexia* 9, 140–145. doi: 10.1002/dys.254

Hatcher, P. J., Hulme, C., and Ellis, A. W. (1994). Ameliorating early reading failure by integrating the teaching of reading and phonological skills: the phonological linkage hypothesis. *Child Dev.* 65, 41–57. doi: 10.2307/1131364

Holmes, J., Gathercole, S. E., Place, M., Dunning, D. L., Hilton, K. A., and Elliott, J. G. (2010). Working memory deficits can be overcome: impacts of training and medication on working memory in children with ADHD. *Appl. Cogn. Psychol.* 24, 827–836. doi: 10.1002/acp.1589

Kyle, F., Kujala, J., Richardson, U., Lyytinen, H., and Goswami, U. (2013). Assessing the effectiveness of two theoretically-motivated computer-assisted reading interventions, GG Rime and GG Phoneme. *Read. Res. Q.* 48, 61–76. doi: 10.1002/rrq.038

Li, Y., Li, H., De, X., Sheng, X., Richardson, U., and Lyytinen, H. (2017). An evidence-based research on facilitating students' development of individualize learning by game-based learning-pinyin graphogame as an example. *China Educ. Technol.* 364, 95–101.

Livingstone, S. (2012). Critical reflections on the benefits of ICT in education. *Oxford Rev. Educ.* 38, 9–24. doi: 10.1080/03054985.2011.577938

Lyytinen, H., Erskine, J., Kujala, J., Ojanen, E., and Richardson, U. (2009). In search of a science-based application: a learning tool for reading acquisition. *Scand. J. Psychol.* 50, 668–675. doi: 10.1111/j.1467-9450.2009.00791.x

Lyytinen, H., Ronimus, M., Alanko, A., Poikkeus, A., and Taanila, M. (2007). Early identification of dyslexia and the use of computer game-based practice to support reading acquisition. *Nordic Psychol.* 59, 109–126. doi: 10.1027/1901-2276.59.2.109

McTigue, E. M., Solheim, O. J., Zimmer, W. K., and Uppstad, P. H. (2019). Critically reviewing GraphoGame across the world: recommendations and cautions for research and implementation of computer-assisted instruction for word-reading acquisition. *Read. Res. Q.* 55, 45–73. doi: 10.1002/rrq.256

Ojanen, E., Ronimus, M., Ahonen, T., Chansa-Kabali, T., February, P., Jere-Folotiya, J., et al. (2015). GraphoGame – a catalyst for multi-level promotion of literacy in diverse contexts. *Front. Psychol.* 6:671. doi: 10.3389/fpsyg.2015.00671

Patel, P., Torppa, M., Aro, M., Richardson, U., and Lyytinen, H. (2018). GraphoLearn India: the effectiveness of a computer-assisted reading intervention in supporting struggling readers of english. *Front. Psychol.* 9:1045. doi: 10.3389/fpsyg.2018.01045

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

Ruiz, J., Lassault, J., Sprenger-Charolles, L., Richardson, U., Lyytinen, H., and Ziegler, J. (2017). *GraphoGame: un Outil Numérique Pour Enfants en Difficultés D'apprentissage de la Lecture*. Paris, ANAE:Actualtites de neuropsychologie de l'enfant, 333–343.

Saine, N. L., Lerkkanen, M.-K., Ahonen, T., Tolvanen, A., and Lyytinen, H. (2011). Computer-assisted remedial reading intervention for school beginners at risk for reading disability. *Child Dev.* 82, 1013–1028. doi: 10.1111/j.1467-8624.2011.01580.x

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., and Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities. *J. Learn. Disabil.* 34, 33–58. doi: 10.1177/002221940103400104

Torgesen, J. K., Wagner, R. K., and Rashotte, C. A. (1999). *Test of Word Reading Efficiency*. Austin, TX.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., and Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *J. Exp. Psychol.* 124, 107–136. doi: 10.1037/0096-3445.124.2.107

van de Ven, M., de Leeuw, L., van Weerdenburg, M., and Steenbeek-Planting, E. G. (2017). Early reading intervention by means of a multicomponent reading game. *J. Comput. Assist. Learn.* 33, 320–333. doi: 10.1111/jcal.12181

van Gorp, K., Segers, E., and Verhoeven, L. (2014). Repeated reading intervention effects in kindergartners with partial letter knowledge. *Int. J. Disabil. Dev. Educ.* 61, 225–229. doi: 10.1080/1034912X.2014.932572

van Gorp, K., Segers, E., and Verhoeven, L. (2016). Enhancing decoding efficiency in poor readers via a word identification game. *Read. Res. Q.* 52, 105–123. doi: 10.1002/rrq.156

Worth, J., Nelson, J., Harland, J., Bernardinelli, D., and Styles, B. (2018). *GraphoGame Rime: Evaluation Report and Executive Summary*. London: Wellcome Trust.

Conflict of Interest: The University of Cambridge has licensed GraphoGame Rime to Grapho Group Oy for commercial distribution, and UG will be one beneficiary of monies received by the University.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.