Title: The effectiveness of adaptive versus non-adaptive learning with digital educational games

Abstract: Digital educational games are typically used for the training of academic skills. Given the heterogeneity in children's prior knowledge and learning pace in classrooms, especially educational games with integrated adaptivity are promising. Adaptive educational games are considered superior to non-adaptive educational games, because they constantly assess children's performance, and accordingly adapt the difficulty of the tasks corresponding to the children's individual level. However, empirical evidence with regard to the effectivity of adaptive compared to non-adaptive educational games to date is limited. A study was conducted with 191 children from the 3rd year of Kindergarten who were enrolled in one of three conditions, i.e. playing an adaptive version of the Reading Game (RG), a non-adaptive version of the RG or training with pen-and-paper exercises. In all three conditions, children trained emergent reading (i.e. phonological awareness and letter knowledge) once a week for 30 minutes over a period of five weeks. Children's performance on cognitive (i.e. phonological awareness, letter knowledge, reading fluency) and non-cognitive (i.e. motivation, self-concept) factors was assessed. Results revealed a significant improvement in phonological awareness and letter knowledge in all conditions. However, no differences between the conditions were observed with respect to children's improvement on phonological awareness and letter knowledge or on their posttest scores for reading fluency. In addition, with regard to motivation and self-concept, again, no differences in these non-cognitive factors were observed across conditions.

Keywords: digital game-based learning, adaptive learning, emergent reading, preschool, intervention

1 Theoretical background

1.1 Educational games to support learning

Very early in education, children need to learn important skills such as reading and math. Persistent training of these academic skills is required in order to prevent children from later learning difficulties as it is proved that early reading and math skills are predictive for children's future academic achievement (Melby-Lervåg, Lyster, & Hulme, 2012; Sasanguie, Göbel, Moll, Smets, & Reynvoet, 2013). One opportunity to establish practice in the classroom is the use of digital educational games which are specifically designed to contribute to educational purposes (Jamshidifarsani, Garbaya, Lim, Blazevic, & Ritchie, 2019; Szűcs & Myers, 2017). Specific characteristics such as immediate feedback, meaningful and engaging contexts, interaction, modality and adaptive practice are promising features in order to enable learning (Wouters & Van Oostendorp, 2013). Because digital educational games are not only informative but also entertaining, they are assumed to impact learners' cognitive as well as non-cognitive learning outcomes (All, Nuñez Castellar, & Van Looy, 2015; Wouters, van Nimwegen, van Oostendorp, & van Der Spek, 2013).

For a long time, it has been uncritically assumed that digital game-based learning (DGBL) automatically results in benefits for learning (Feldon, 2010). Empirical studies have shown positive effects of a DGBL training on children's math and reading achievement in normally developing children as well as in children with learning difficulties (Lämsä, Hämäläinen, Aro, Koskimaa, & Äyrämö, 2018; Clark, Tanner-Smith, & Killingsworth, 2016). However, several systematic reviews (e.g., Backlund & Hendrix, 2013; Vandercruysse, Vandewaetere, & Clarebout, 2012) and meta-analyses (e.g., Byun & Young, 2018; Girard, Ecalle, & Magnan, 2013) have revealed that the empirical evidence regarding the effectiveness of digital educational games is less clear.

More and more authors emphasize the ambiguity in the results on the effectiveness of DGBL. As a counter movement, an increasing number of studies have started focusing on the effectiveness of specific characteristics of educational games. Previously, studies typically adhered to a media comparison approach evaluating cognitive and non-cognitive effects by comparing a game condition to a non-game condition. In the non-game condition, often, traditional activities were used for comparison which did not necessarily follow the same objectives and equivalent design as the games. This makes it difficult to understand why one game leads to better effects than the other (Sampayo-Vargas, Cope, He, & Byrne, 2013). However, relying on the well-known media-debate (Clark et al., 2016), it is argued that not the medium *as such* determines the learning effects of DGBL. Therefore, it might be more promising to follow a value-added approach in which a particular characteristic of the game is examined (e.g., feedback, adaptivity, narrative etc.), rather than the medium (i.e. game-based training). Designs adapting a value-added approach typically compare the effects of two versions of the same game which only differ with regard to one game characteristic. Consequently, the mechanisms by which games induce learning and motivate players to persist in gameplay can be investigated.

1.2 Adaptive learning

One promising game feature is the implementation of adaptivity in educational games (Wouters & Van Oostendorp, 2013). In primary education, teachers are faced with huge differences in terms of learners' prior knowledge (Tomlinson et al., 2003). Through individualized learning support one can account for these differences. However, individualized learning is not easy to establish in a traditional classroom due to the time and knowledge teachers need to develop individualized programs (Dowker, 2017). In this regard, adaptive digital educational games provide a response to this challenge as they allow adaptive learning (Shute & Towle, 2010). Papousek and Pelanek (2005, p. 1) wrote "The goal of adaptive educational systems is to make learning more effective and engaging by tailoring the behavior of the system to a particular student. The adaptive behavior is based on learner models which estimate the knowledge of students (and potentially other characteristics like their affective state)”. There are several adaptive learning environments that take into account cognitive needs. This implies that children are for example provided with exercises on their individual level.

The framework of Vandewaetere and Clarebout (2014) distinguishes static and dynamic approaches to learner modeling and their combination. In the *static approach*, the learner model is built before a learner enters the learning environment. Adaptation is based on pre-task measurements of learner characteristics and takes place before the instruction starts. A *dynamic approach* of modeling means that the models are able to track changes in the learners’ individual characteristics which can in turn result in an update of the model (Vandewaetere & Clarebout, 2014, for an example: Klinkenberg, Straatemeier, & Van Der Maas, 2011). In recent years, *combination models* emerged in which the static and dynamic learner models are combined. Accordingly, the learner model does not only use the prior information about learners (i.e. static) but also keeps track of the learning characteristics based on the learner-item interaction (i.e. dynamic) and accordingly provides individualized practice (e.g., Ketamo, 2003; Shute, Ke, Wang, 2017).

1.3 The effectiveness of adaptive educational games

According to All and colleagues (2015), DGBL may influence children’s cognitive, non-cognitive and efficiency outcomes. Regarding cognitive outcomes, many researchers argued that positive effects may be obtained because adaptive games challenge and support children by providing exercises that are not too difficult or too easy for them (Papoušek & Pelánek, 2015). Hooshyar, Yousefi, and Lim (2018) fostered early English reading skills in

150 children. Participants were randomly assigned to an adaptive condition or a non-adaptive condition. Results demonstrated a larger increase in performance on phonological awareness and letter knowledge in children playing the adaptive game compared to children playing the non-adaptive one. Holmes, Gathercole, and Dunning (2009) trained working memory in 42 children and reported positive outcomes for verbal working memory, visuo-spatial working memory and following instruction tasks in favor of the adaptive condition compared to the non-adaptive condition, while for intelligence and academic attainment the learning effects were the same in both conditions. By contrast, other studies did not observe differences in learning gain between the adaptive and non-adaptive condition (Orvis, Horn, & Belanich, 2008; Plass, Homer, Pawar, Brenner, & MacNamara, 2018).

Regarding non-cognitive outcomes, motivation is an important aspect of learning which can be influenced by an adaptive game by selecting the suitable difficulty of tasks (Andrade, Ramalho, Gomes, & Corruble, 2006; Jansen et al., 2013; Papousek & Pelanek, 2015). In this context, the Self-Determination Theory (SDT, Deci & Ryan, 2004) is a well-established motivational framework. The SDT describes three important needs (i.e., need for competence, autonomy, relatedness) that should be addressed to keep learners motivated (Deci & Ryan, 2004). Need for competence refers to the need that learners must feel competent during training. It is argued that, to support children’s need for competence, educational games should set challenging but manageable goals (van Roy & Zaman, 2017). Because this goal is better accomplished in adaptive educational games, these games could enhance the learner’s self-concept and intrinsic motivation (Nurmi & Aunola, 2005). Despite this reasoning, no evidence was found for increased self-concept and motivation in the adaptive condition compared to the non-adaptive condition (Orvis et al., 2008, Van Oostendorp, van der Spek, & Linssen, 2014). For instance, Sampayo-Vargas and colleagues (2013) trained secondary school students on Spanish cognates in an adaptive, non-adaptive or control condition. Engagement, perceived competence and preferred level of difficulty were measured, but no differences between the three conditions were observed.

With respect to efficiency, All and colleagues (2015, p. 36) defined this measure as "The time required to teach the target group certain content. This is a relative worth, compared to other instructional methods.". Adaptive educational games can contribute in accounting for children's different learning paces: quick learners are provided with less exercises and consequently, they need less time to achieve certain learning goals, whereas slower learners are presented with more exercises and will need more time to learn the same content (Tomlinson et al., 2003; van Oostendorp et al., 2014). As such, individualized learning can be achieved resulting in some children showing greater learning efficiency than others (All et al., 2015). Van Oostendorp and colleagues (2014) assigned students to an adaptive condition or a non-adaptive condition of a game training triage. Results revealed that participants in the adaptive condition completed more tasks in the same time span compared to the students in the control condition and were therefore more efficient (van Oostendorp et al, 2014). In another study, students were assigned to either an adaptive game-based e-learning or a non-adaptive e-learning environment for training geometry, physics and chemistry. The results revealed that the students achieved higher scores in less time in the adaptive system compared to the non-adaptive system (Ali & Sah, 2018).

1.4 The present study

The available literature does not provide a clear pattern with respect to the effects of adaptive educational games. First, only a small number of studies have investigated the effects of adaptive games by contrasting two versions that only differ with regard to the adaptive adjustment mechanism used (see also Miljanovic & Bradbury, 2018). Second, several studies have compared the adaptive and non-adaptive condition with a passive control condition, or do not include any control condition which hampers a clear understanding of the effectiveness of adaptive educational games (All, Nunez Castellar, & Van Looy, 2016). Third, the effectiveness of the intervention is mostly measured in terms of cognitive outcomes, but rarely in terms of non-cognitive or efficiency outcomes.

In the current study, we investigated the effects of an intensive intervention in which third grade kindergarten children (5-6 years old) practiced emergent reading skills with an adaptive or non-adaptive version of a Reading Game (RG). In our study, only the adaptivity was manipulated, while the medium (i.e. game) and the content/time that was spent on training remained constant. The adaptive version of the RG was based on a dynamic learner model which updated and changed the parameters depending on children's accuracy during gameplay. Both game conditions were further contrasted with a condition in which children trained the same content with pen-and-paper exercises (i.e. active control condition). To examine the effectiveness of the medium (i.e. game), the non-adaptive condition was compared with the active control group.

The effectiveness of the intervention was investigated on cognitive and non-cognitive learning outcomes. This is in part inspired on the theoretical framework we described that made clear that adapting a learning environment to the ability of a learner might result in better learning gains (Sampayo-Vargas et al., 2013) and increased motivation (Deci and Ryan 2004; van Roy & Zaman, 2017). In addition, previous studies in the domain of DGBL also have argued that children's self-concept might change as a result of training (Choi, 2015; Jansen et al., 2013; Miller & Robertson, 2011). With respect to the cognitive outcomes, near (i.e., transfer to closely related skills) and far (i.e., transfer to remote skills) transfer were measured. More specifically, phonological awareness and letter knowledge were the near transfer skills that were evaluated and reading fluency the far transfer skill. With respect to non-cognitive outcomes, previous research has confirmed that even young children show low interest for particular subjects and that there is a decrease over time in interest for particular subjects (such as reading) and self-concept (Nurmi & Aunola, 2005). Therefore, motivation was measured with the Task-Value Scale for Children (TVS-C, Nurmi & Aunola, 2005) as it is has been demonstrated that task motivation is a similar concept as ‘intrinsic motivation’ (Deci & Ryan, 2004) used in previous research on learning motivation (Nurmi & Aunola, 2005). The following research questions were addressed: (1) How does the use of pen-and-paper exercises, non-adaptive gameplay and adaptive gameplay affect children’s cognitive factors (i.e. phonological awareness, letter knowledge, reading fluency) in emergent reading?; and (2) How does the use of pen-and-paper exercises, non-adaptive gameplay and adaptive gameplay affect children’s non-cognitive factors (i.e., interest in reading, self-concept) in emergent reading?.

2 Methodology

2.1 Sample

In total, 191 children (114 boys, *M* (*SD*) age = 5,88 years (.37)) of the third grade of Kindergarten from six different elementary schools took part in this study. Due to the implementation in a realistic context, such as a class group, full randomization of subjects was not possible. Classes (*N*=9) were assigned as an intact group to the experimental condition and the control condition (i.e. cluster randomized trial). Children from the experimental condition were randomly assigned to either the adaptive condition or the non-adaptive condition (i.e. individual randomized trial). Each condition consisted of a similar number of participants, more specifically 64 children were assigned to the non-adaptive condition, 62 children to the adaptive condition and 65 children to the control condition. Participants assigned to the adaptive or non-adaptive condition were not aware of the condition they belonged to. Children were excluded from the study if their parents did not accept the active informed consent, if they missed two or more RG-interventions or when they were not able to complete all pretests and/or posttests. Prior to the study, approval of the social and societal ethics committee (SMEC) was confirmed to conduct this study (G-2018 03 1194).

2.2 Reading Game

The educational game RG was used for the training of emergent reading. During Kindergarten, children train phonological awareness and letter knowledge to be prepared for the start of formal reading education in the first grade. Reading is a complex cognitive process for which phonological awareness and letter knowledge are domain-specific cognitive skills shown to be predictive for individual differences in children's (later) reading development (Chu, Van Marle, & Geary, 2016; Melby-Lervåg, et al., 2012). Phonological awareness refers to the "awareness of and access to the sound structure of oral language" (Wagner et al., 1997,

p. 469). Letter knowledge refers to letter names and the correspondence between letter names and letter sounds (Melby-Lervåg et al., 2012). The RG consists of 65 levels in total, and within each level children can choose among three or four subgames which train the same content. When a subgame is accomplished successfully (i.e. earn at least 1 star), the next level is unlocked (see Figure 1). For example, children train phonological awareness a.o. by hearing a word in parts, and are instructed to synthesize the words and to find the corresponding picture. Letter knowledge is trained by tasks involving particular letters and graphemes. One example of a task is when children hear a letter and have to choose the grapheme among three other graphemes (see Figure 2).

[Insert: *Figure 1*. Screenshot of the learners' dashboard.]

[Insert: *Figure 2.* Screenshots of subgames in the RG.]

To establish adaptivity in the RG, the Elo-rating system was used which combines good predictive accuracy with simplicity (Park, Joo, Cornillie, van der Maas, & Van den Noortgate, 2019; Pelánek et al., 2017). This approach enables learning environments in which the difficulty of the item is constantly matched to the ability level of the learner (Wauters, Desmet, & Van den Noortgate, 2012). Therefore, item response theory (IRT, Van der Linden & Hambleton, 1997) was used to model and estimate the level difficulty. IRT is a psychometric approach that emphasizes the fact that the probability of a discrete outcome, such as the correctness of a response to an item, is a function of characteristics of the item and qualities of the person. This approach has already been successfully applied in the DGBL environment 'Math garden' which trains arithmetic skills in primary school children (Klinkenberg et al., 2011).

2.3 Description of the conditions

Children assigned to the *non-adaptive condition* played a non-adaptive version of the RG. Children assigned to the *adaptive condition* played an adaptive version of the RG. Both versions contain the same order of levels, as it is important that children train all letters of the alphabet and do not skip some. The non-adaptive and adaptive version only differ in the difficulty adjustment mechanism used. In what follows, two characteristics related to difficulty adjustment, on which the non-adaptive and the adaptive version of the RG differ, are described.

A first characteristic is related to the number of exercises provided in each subgame. In the non-adaptive game, children are in each subgame presented with 20 exercises. In contrast, in the adaptive game, the number of exercises children are presented changes according to their performance during playing. More specifically, the number of exercises will get larger when the user ability estimate is lower than the item difficulty (and vice versa). After each response given, the ability estimate is updated. The adaptivity will make sure that strong performing children will go faster through the levels because they get fewer exercises (minimal 10) in a subgame. In contrast low performing children will get more exercises each subgame (maximal 30), and will go slower through the levels.

A second characteristic is related to the applied pass level. In the non-adaptive game, regardless of children's performance in the subgame, they always proceed to the next level after having completed the exercises of the current subgame. As such, children are always at least rewarded with one star if they complete the exercises of a subgame. Children earn one star when they complete the exercises within the subgame regardless of their performance; two stars are collected if at least 75 percent of the exercises were solved correctly; three stars are won if all exercises were solved correctly. With regard to the adaptive game, children pass to the next level if they solve at least 65 percent of the exercises correctly. Then, children earn one star and can continue with the next level. If not, children were instructed to replay a subgame within the same level. Two stars are collected if at least 75 percent of the exercises were solved correctly. Three stars are won if all exercises were solved correctly. Figure 1 shows the overview display for the children where they could follow their own level proficiency.

In the *active control condition*, children took pen-and-paper tasks with content similar to that of the RG (i.e. training phonological awareness and letter knowledge), but without being embedded in a digital game, and thus, no immediate feedback, no attractive narrative and no adaptivity. Teachers were instructed to not use tablet or computer games during the intervention period. In advance, they received all pen-and-paper exercises and the accompanying instructions for the training sessions (see an example of an exercise in Figure 3). The tasks were identical as in the game and consisted of the same words, letters and graphemes. More specifically, in the first two weeks, children trained auditory blending of letters to words, the meaning of words and recognizing individual letters. In the following weeks, the children trained intensively letter knowledge by means of auditory and visual memory/blending tasks. From week 4, the learning content was not limited anymore to individual letters, but children were also presented with words consisting of the learnt letters.

[Insert: *Figure 3.* Exercise in the active control condition (left panel) and in the RG (right panel). Instructions: “Look to the letter image, then find the same grapheme which is presented by the letter image among the three graphemes.”]

2.4 Procedure

Over a period of five weeks, each experimental condition received in total five game training session of 30 minutes. For each intervention, a researcher came to the classroom to distribute the tablets and headphones provided for the research. The children from the control condition received paper-and-pen tasks, which were developed by the researchers (see section 2.3). The researchers provided a task schedule for the teachers of the control condition, and similar as in the experimental conditions, the children practiced five times 30 minutes over a period of five weeks in a non-gamified context.

2.5 Materials

The materials that were used to measure the cognitive and non-cognitive factors are described below.

Phonological awareness. Phonological awareness was measured before and after the intervention with two different instruments measuring auditory blending skills and auditory memory skills, which are both part of a standardized test battery to measure emergent literacy (Aarnoutse, Beernink, & Verhagen, 2016). To measure auditory blending, children heard separate letter sounds of a word and had to choose the corresponding picture out of four different pictures, e.g. [b-e-d]. Then, the child had to answer with marking the picture of a "bed". The original instrument consists of 20 items. In this study, the first ten items were used as a pretest, and the remaining ten items were used as a posttest. To measure auditory memory, respondents were presented with a pair of illustrations which the teacher called by name. The instrument consisted of eight items in total. For the first four items, children were asked which of the words, that were read aloud, started with a certain sound. For the next four items, they had to indicate the illustration that represented the word ending with the given sound. The reliability of the auditory blending instrument and the auditory memory instrument have Cronbach's alpha's of respectively .89 and .90 (Verhagen, Aarnoutse, & Van Leeuwe, 2009).

Letter knowledge. Two different letter knowledge tests were used (Aarnoutse et al., 2016). The first instrument, consisting of eight items was used in both pre- and posttest (Letterkennistoets 1, Aarnoutse et al., 2010). The children were presented with caterpillars containing 8 different graphemes in the different segments of each caterpillar. Children were required to listen to the letter the teacher prompted and to search and colour the corresponding grapheme in the caterpillar. The second instrument was an extended version of the same test with 21 items (Letterkennistoets 2, Aarnoutse et al., 2010). The instruments were reliable with Cronbach's alpha's of respectively .81 and .90 (Verhagen et al., 2009). This extended letter knowledge instrument was only assessed as a posttest because we were afraid that children would score at ceiling level in the short version of this instrument. Besides, the extended letter knowledge test measures new letters that children might have learned during the training.

Reading fluency. In order to evaluate whether children transferred the trained knowledge to reading new words, reading fluency was assessed after the training. The Three-Minutes-Test (TMT; Moelands, Kamphuis, & Rymenans, 2003) contains of reading cards with different word types of vowel and consonant combinations that must be read aloud during 1 min. We used one reading card to verify if respondents were able to read simple consonant-vowel-consonant words at the end of the intervention period. The number of correctly read words in 1 minute was registered. The reliability of the test is high (Cronbach's alpha for grade 1 to grade 6 between .86 and .90, Moelands et al., 2003).

To measure the non-cognitive factors, data were gathered using previous validated instruments which were translated to Dutch and adapted to the topic of the study (i.e. emergent reading) by the researchers. The questionnaire data of the participants was collected after the training in a one-to-one setting with the experimenter and the participant.

Interest in reading. The Task-Value Scale for Children was conducted to assess children's interest in reading (TVS-C, Nurmi & Aunola, 2005; Viljaranta et al., 2014, 2017).

The instrument was slightly adapted in function of the tasks with which children were confronted during the study. Children from the experimental conditions were asked three questions about their motivation for reading. For the first item, it was referred to images which were also presented in the RG ('How much do you like letter and reading tasks?'). The second and third item referred to how much children liked letter and reading tasks in general ('How much do you enjoy doing exercises with letters and words in the classroom?'; 'How much do you enjoy doing exercises with letters and words at home?'). Children from the control condition were asked the same questions, but the experimenter referred for the first item to exercises they were presented with during the pen-and-paper exercises. The experimenter read aloud the questions, and explained the meaning of each of the five faces (1=I do not like it at all/I dislike doing those tasks; 5= I like it very much/I really enjoy doing these tasks). The children were asked to point the face which corresponded best with their interest. In this study, the reliability of the questionnaire was low (Cronbach's alpha = .44), which was surprising because previous studies reported reliable Cronbach alpha's of at least .61 in Kindergarten children (a.o., Nurmi & Aunola, 2005; Viljaranta et al., 2017). The three items of the interest in reading measure will be analyzed separately as the first item refers explicitly to the tasks children were presented with in the experimental or control condition, while the second and third item refer to their general interest in reading.

Self-concept. Children's self-concept regarding their ability for reading was measured with the Self-Concept of Ability instrument (Aunola, Leskinen, Onatsu-Arviommi, & Nurmi, 2002; Nicholls, 1978, Viljaranta et al., 2017). Children were presented with a vertical row of 20 faces in a line from the top to the bottom of the page. The children were told that these faces represented children of the same age and that the one at the top of the page represented the child who was best at reading, who was in the middle and so on down to the poorest performer. The following instructions were used: 'Think how good or poor you are at reading and letter tasks. how do you feel? Are you, in your class, the most skilful, somewhere up here, or are you down here, the least skilled in literacy, or are you somewhere here at the middle? Where are you?'. The children responded by pointing to one of the faces. The instrument was complemented with pictures of the tasks the children performed during the study, a pen-and-paper exercise in the active control group and an illustration of an exercise from the RG in the experimental groups.

To examine whether the learning outcomes were moderated by individual differences, we investigated possible effects of prior knowledge, home language and socio-economic status (SES). Prior knowledge was operationalized by using the pretest results on the letter knowledge task. Using a parents questionnaire, information was collected about which language children speak at home as it has been showed that oral language is an important predictor of later reading competence (Chung, 2015). Parents could choose from four possible answers (only Dutch; mostly Dutch and sometimes another language; sometimes Dutch but mostly another language; no Dutch, but always another language). Finally, SES was indexed by asking the mother's highest degree of education (indicating one of the following 6 possible answers: no degree, primary education, lower secondary education, higher secondary education, higher education, university degree) through the same parents questionnaire. We operationalized SES like this because previous studies have shown that parental education (and mothers in particular) is one of the best indicators of a family's SES (Chung, 2015).

3 Analyses

Prior to analysing the data of the pre- and posttests, it was investigated how children proceeded through the game in both conditions. The log data were checked before and during the intervention in order to control whether the adaptivity algorithm worked properly. After the intervention, the log data were used to determine the number of levels the children ran through. We observed that children in the adaptive condition on average reached the 36th level (*M*=36.13; *SD*=15.93) while children in the non-adaptive condition on average reached the 31st level (*M*=31.80; *SD*=20.63). Although we instructed the children always to advance to the next level, we observed that children sometimes played subgames in previously unlocked levels. In the adaptive condition, children played on average 4 previously unlocked levels (*M*=4.35; *SD*=6.17), while in the non-adaptive condition, children played on average 3 previously unlocked levels (*M*=2.98; *SD*=4.80). However, the differences between the adaptive and non-adaptive condition were not significant with regard to the on average reached level and the on average played previously unlocked levels.

Next, one-way Analyses of Variances (ANOVA) on all pretest measures were conducted in order to confirm that there were no differences in the results of the children between the three conditions prior to the intervention. No differences were observed between the three conditions for auditory blending ($F(2,188)=0.22$, $p=.80$), auditory memory ($F(2,188)=0.99$, $p=.38$) and letter knowledge ($F(2,190)=1.00$, $p=.37$).

For assessing the effect of the intervention on the cognitive factors, three Repeated-Measures ANOVAs were conducted with moment (pretest; posttest) as within-subjects factor and condition (non-adaptive condition; adaptive condition; active control condition) as between-subjects factor on children's scores on the cognitive tests (i.e. phonological awareness, letter knowledge). For the extended letter knowledge task and the reading fluency task, no pretest data was available, therefore two one-way ANOVAs were conducted with condition as factor and the cognitive measure as dependent variable. In addition, it was examined whether prior knowledge, home language and SES moderated the effect of the intervention, Analyses of Covariance were carried out.

With regard to the results of the intervention on non-cognitive factors, four one-way ANOVAs (i.e. on the scores on the three items of interest in reading and the score on self-concept) were conducted with condition as a between-subject factor.

4 Results

4.1 Effects on cognitive factors

For auditory memory skills, a main effect of moment was present, $F(1,182)=15.02$, $p=.00$, $\eta_p^2=.08$, while an interaction effect of moment by condition was not observed. This indicates that all children improved on the test. On the auditory blending skills, a main effect of moment was present, $F(1,182)=5.25$, $p=.02$, $\eta_p^2=.03$, indicating that children scored better on the posttest compared to the pretest. There was no significant interaction of moment by condition, which means that children did not benefit more or less from one or the other condition. Children also scored better on the short letter-sound knowledge test after the intervention, $F(1,188)=106.39$, $p=.00$, $\eta_p^2=.36$. The interaction effect with condition was not significant, indicating that the improvement in letter-sound knowledge was similar in all conditions. Table 1 shows the descriptive results of the pre- and posttests.

Letter knowledge was also measured after the intervention with a more extensive version of the test. No differences were noticed between the conditions. Similarly, for reading fluency, no differences were observed between the conditions.

To conclude, significant learning gains were found immediately following the completion of training in all conditions. The group receiving adaptive training did not perform better than the children of the non-adaptive and active control condition.

[Insert Table 1: *Impact of training on cognitive measures*]

Furthermore, we tested for possible moderating effect of prior knowledge, home language and SES. No significant interaction effects of condition and prior knowledge ($F(2,185)=0.54$, $p=.58$, $\eta_p^2=.01$), condition and home language ($F(6,177)=2.03$, $p=.06$, $\eta_p^2=.07$) and condition and SES ($F(10,167)=1.19$, $p=.30$, $\eta_p^2=.07$) were observed indicating that none of these factors moderated the cognitive outcomes of the intervention .

4.2 Effects on non-cognitive factors

Table 2 shows the descriptives and the results of the ANOVA for the questionnaire data of interest in reading and self-concept. No effects of condition were observed for all these items.

[Insert Table 2: *ANOVA of interest (3 items) and self-concept (1 item)*]

5 Discussion

Although training with digital educational games is promising, previous studies mainly assessed the effects of these trainings by contrasting a game versus a no game intervention. However, several scholars claim that it is more appropriate to measure the effects of a specific game feature rather than the medium if one aims to evaluate the effectiveness of games. One particular game feature that might be promising is adaptivity of the content to learners' content knowledge. The aim of the present study was to conduct an empirical study investigating the effects of adaptivity on a game training emergent reading.

The first research question concerned the difference between adaptive gameplay, non-adaptive gameplay and pen-and-paper exercises on cognitive factors. Previous studies revealed mixed evidence as in most of the studies children benefitted more from adaptive compared to non-adaptive training on near transfer tasks, while for far transfer skills no differences between these conditions were observed (Holmes et al., 2009; Hooshyar et al., 2018). Furthermore, results have revealed no differences between the non-adaptive and the control condition, but they did observe benefits in favour of the adaptive condition compared to the control condition (Ketamo, 2003; Sampayo-Vargas et al., 2013). The results of our empirical study showed that children significantly improved from pretest to posttest on phonological awareness and letter knowledge (near transfer) in all conditions. Also on the extended letter knowledge task and reading fluency test (far transfer) which were only measured in the posttest, no differences were found between the three conditions. Further analyses showed also that children's individual differences (i.e. prior knowledge, home language, SES) did not moderate the impact of the training.

The lack of significant differences between (non-)adaptive gameplay and pen-and-paper exercises is in line with previous studies that the results of the children did not differ after the non-adaptive training when compared to the active control condition (Sampayo-Vargas et al., 2013). The reason is that children of the active control condition practiced the same content as in the non-adaptive condition. This means that the non-adaptive game can be used by teachers as a variation to other learning activities for example and as such provides new and powerful affordances, but with similar learning gains (Clark et al., 2016). Though a benefit of the adaptive condition compared to the non-adaptive and active control condition was expected, this was not confirmed in this study. A possible explanation is that the training period was not intensive enough to observe differences between the conditions. However, other studies consisted of trainings which were less intensive than in this study (with training duration less than one hour) but nevertheless reported positive findings in favour of the adaptive condition (e.g., Ketamo, 2003; Sampayo-Vargas et al., 2013). Another explanation is related to the validity of the instruments. In this study validated standardized instruments were used as recommended by All and colleagues (2016) in order to not bias the outcomes of the study, but it might be the case that these instruments are not as aligned to the intervention as for example self-developed instruments or in-game performance tests which were applied multiple times. For example, in the study of Hooshyar and colleagues (2018) children's phonological awareness and letter knowledge were assessed based on children's in-game performance to establish children's learning gains due to the (adaptive) training.

The second aim of this study was to investigate the effects non-adaptive and adaptive games on non-cognitive factors (i.e., interest in reading, self-concept). Only a few studies investigated the effects of adaptive learning on non-cognitive factors. The results revealed no differences between the adaptive and non-adaptive condition (Sampayo-Vargas et al., 2013). In line with this, our empirical study also observed no differences between the conditions. The reason for not finding an effect might be due to the implementation of the adaptivity in the RG. Similar to the non-adaptive condition in which children might be frustrated that items were too easy or too difficult, the adaptive condition might have caused some frustration for children as the requirement was to solve 65% of the items correctly within one level before they can proceed to a next level. Consequently, for some children, after playing the same level several times, frustrations popped up and their motivation to continue the game might have decreased. Although this study implemented a dynamic approach of adaptive learning, a more fine-grained learner model might be more appropriate. More specifically, only adjusting the number of exercises children are provided with to practice a certain content is not sufficient. Observations from the log data confirmed that children played previously unlocked levels although they were instructed to always advance to the next unlocked level. It might be the case that weak performing children were not motivated to learn new content by only making more exercises training the same content. Rather, the adaptive system should also decrease the difficulty level of the exercises which are based on reliable learner estimates obtained from previous data collection. By doing so, children's engagement during play might be strengthened through which persistent training could be realized.

To conclude, the results of the empirical study revealed no convincing evidence for a positive effect of adaptivity. This is possibly due to the specific design of our own empirical study and of previous studies. In the next paragraphs, three elements which might have led to the null effect are explained, and accordingly, recommendations for future research are given. First, the adaptivity algorithm might not have been fine-grained enough which might have hampered the expected gains on cognitive and non-cognitive factors. Our empirical study has shown that when the adaptivity algorithm is not able to also decrease the level of difficulty, this might affect children's motivation in the sense that they feel unable to proceed. Previous studies suggested that children's interest in a certain topic and their self-concept is associated with academic performance, even in children in the first year of school (Dowker, Cheriton, Horton, & Mark, 2019). Although the direction of causality is not established yet, it is explicitly stated that "it is advised to focus in early years on reducing distress at failure and on preserving confidence in ability to perform" (Dowker, et al., 2019, p. 225). Future studies should implement games which dynamically adapt according to children's performance, by both decreasing and increasing the level of difficulty during gameplay, which might affect children's motivation to practice, and accordingly also their cognitive outcomes.

Second, measuring non-cognitive factors is complex, and especially difficult in young children. In this regard, a limitation of the current study is that 'interest' and 'self-concept' were only measured once, after the intervention. Due to the fact that children were partly random assigned to a condition, and we did not find any differences between the conditions with regard to cognitive pretest data, we assumed that children did not differ at the moment of the pretest in terms of non-cognitive factors. It would however be of interest for future studies to control for this measure by testing children's non-cognitive factors in advance in order to confirm this assumption. Regarding the instrument to measure non-cognitive factors, self-reported data was used. A characteristic of self-reports is however that it is assumes that children can express how they feel already at a very young age. In addition, it can be questioned whether children at this age are able to think retrospectively about one of the multiple activities they experienced in the classroom. As the questionnaire data were collected after the intervention, the differences in children's affect during gameplay were not captured. In order to meet these shortcomings, a multimodal approach is recommended for the assessment of non-cognitive factors in which questionnaires and qualitative data (such as interviews) are complemented with unobtrusive data (e.g., increased heartbeat might suggest negative emotions) (e.g., Hunt, Bhardwa, & Sheffield, 2017).

Third, we have found few studies which included efficiency measures revealing promising outcomes. In the empirical study, the efficiency outcome was not considered due to the fact that the training time was fixed (30 minutes per week over a period of five weeks). It is suggested that the effectiveness might be established not only by the comparison across the conditions of children's (non-)cognitive outcomes, but also by taking into account children's outcome of efficiency. Especially for adaptive learning, this efficiency outcomes might be an important goal assuming that each student gets tasks adjusted to the own level of ability, and so, speed up the learning process (All et al., 2015; van Oostendorp et al., 2014). Therefore, future studies might register the time children are involved in training with the game. As such, children's learning efficiency can be derived and it might be interesting to examine if adaptive game learning leads to a reduction in time spent to learn a certain content matter.

6 References

Aarnoutse, C., Beernink, J., & Verhagen, W. (2016). *Toetspakket beginnende geletterdheid* [Early literacy]. Amersfoort, The Netherlands: CPS.

Ali, A., & Sah, M. (2017). Adaptive game-based e-learning using semantic web technologies. In *2017 International Conference on Open Source Systems & Technologies (ICOSST)* (pp. 15-23). IEEE.

All, A., Nuñez Castellar, E. P., & Van Looy, J. (2015). Towards a conceptual framework for assessing the effectiveness of digital game-based learning. *Computers and Education*, *88,* 29–37.

All, A., Castellar, E. P. N., & Van Looy, J. (2016). Assessing the effectiveness of digital game-based learning: Best practices. *Computers & Education, 92*, 90-103.

Andrade, G., Ramalho, G., Gomes, a. S., & Corruble, V. (2006). Dynamic game balancing: An evaluation of user satisfaction. *AIIDE*, *6*, 3–8.

Aunola, K., Leskinen, E., Onatsu-Arvilommi, T., & Nurmi, J. E. (2002). Three methods for studying developmental change: A case of reading skills and self-concept. *British Journal of Educational Psychology*, *72*(3), 343-364.

Backlund, P., & Hendrix, M., (2013). Educational games - are they worth the effort? A literature survey of the effectiveness of serious games. In *Games and virtual worlds for serious applications (VS-GAMES)* (pp. 1–8).

Byun, J., & Joung, E. (2018). Digital game-based learning for K–12 mathematics education: A meta-analysis. *School Science and Mathematics, 118*(3-4), 113-126.

Choi, Y. S. (2015). Effectiveness of game based learning to minimize boolean functions. *Multimedia Tools and Applications, 74*(17), 7131-7146.

Chu, F. W., Van Marle, K., & Geary, D. C. (2016). Predicting children's reading and mathematics achievement from early quantitative knowledge and domain-general cognitive abilities. *Frontiers in Psychology*, *7*(5), 1–14.

Chung, K. K. H. (2015). Socioeconomic status and academic achievement. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., Vol. 22, pp. 924–930). Oxford: Elsevier.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: a systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79–122.

Dowker, A. (2017). Intervention for primary school children with difficulties in mathematics. *Advances in Child Development and Behavior*, 53, 255–287.

Dowker, A., Cheriton, O., Horton, R., & Mark, W. (2019). Relationships between attitudes and performance in young children's mathematics. *Educational Studies in Mathematics*, *100*(3), 211-230.

Feldon, D. F. (2010). Why magic bullets don't work. Change: *The Magazine of Higher Learning*, *42*(2), 15–21.

Girard, C., Ecalle, J., & Magnan, A. (2013). Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning, 29*(3), 207-219.

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental Science*, *12*(4), 9–15.

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*(2), 293–310.

Hunt, T. E., Bhardwa, J., & Sheffield, D. (2017). Mental arithmetic performance, physiological reactivity and mathematics anxiety amongst U.K. primary school children. *Learning and Individual Differences, 57*, 129–132.

Jagušt, T., Botički, I., & So, H. J. (2018). Examining competitive, collaborative and adaptive gamification in young learners' math learning. *Computers and Education, 125*, 444–457.

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

Jansen, B. R., Louwerse, J., Straatemeier, M., Van der Ven, S. H., Klinkenberg, S., & Van der Maas, H. L. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences, 24*, 190-197.

Ketamo, H. (2003). An adaptive geometry game for handheld devices. *Educational Technology and Society, 6*(1), 83–95.

Klinkenberg, S., Straatemeier, M., & Van Der Maas, H. L. J. (2011). Computer adaptive practice of math ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education, 57*(2), 1813–1824.

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

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.

Miljanovic, M. A., & Bradbury, J. S. (2018, November). Making serious programming games adaptive. In *Joint International Conference on Serious Games* (pp. 253-259). Springer, Cham.

Miller, D. J., & Robertson, D. P. (2011). Educational benefits of using game consoles in a primary classroom: A randomised controlled trial. *British Journal of Educational Technology, 42*(5), 850-864.

Moelands, F., Kamphuis, F., & Rymenans, R. (2003). Drie-Minuten-Toets voor Vlaanderen (DMT-V): Wetenschappelijke verantwoording [Three-Minutes-Test for Flanders (DMT-V): Scientific justification]. Arnhem, Netherlands: Citogroep.

Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child development*, 800-814.

Nurmi, J. E., & Aunola, K. (2005). Task-motivation during the first school years: A person-oriented approach to longitudinal data. *Learning and Instruction*, *15*(2), 103–122.

Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human Behavior*, *24*(5), 2415–2433.

Papoušek , J., & Pelánek , R. (2015). Impact of adaptive educational system behaviour on student motivation. In *Artificial Intelligence in Education* (Vol. 9112, pp. 348–357).

Park, J. Y., Joo, S. H., Cornillie, F., van der Maas, H. L., & Van den Noortgate, W. (2019). An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments. *Behavior research methods*, *51*(2), 895-909.

Pelánek, R., Papoušek, J., Řihák, J., Stanislav, V., & Nižnan, J. (2017). Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction, 27*(1), 89–118.

Plass, J.L., Homer, B.D., Pawar, S., Brenner, C., & MacNamara, A.P. (2018). The effect of adaptive difficulty adjustment on the effectiveness of a game to develop executive function skills for learners of different ages. *Cognitive Development, 49,* 56–67.

Sampayo-Vargas, S., Cope, C. J., He, Z., & Byrne, G. J. (2013). The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers and Education, 69,* 452–462.

Sasanguie, D., Göbel, S. M., Moll, K., Smets, K., & Reynvoet, B. (2013). Approximate number sense, symbolic number processing, or number-space mappings: What underlies mathematics achievement? *Journal of Experimental Child Psychology*, *114*(3), 418–431.

Shute, V., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59-78). Springer, Cham.

Shute, V., & Towle, B. (2010). Adaptive E-Learning. *Educational Psychologist, 38*(2), 105–114.

Szűcs, D., & Myers, T. (2017). A critical analysis of design, facts, bias and inference in the approximate number system training literature: A systematic review. *Trends in Neuroscience and Education, 6,* 187–203.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., … Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted, 27*(2–3), 119–145.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

van Oostendorp, H., van der Spek, E. D., & Linssen, J. (2014). Adapting the complexity level of a serious game to the proficiency of players. *EAI Endorsed Transactions on Game-Based Learning, 1*(2), 8-15.

Vandercruysse, S., Vandewaetere, M., & Clarebout, G. (2012). Game-based learning: A review on the effectiveness of educational games. In *Handbook of research on serious games as educational, business and research tools* (pp. 628-647). IGI Global.

Vandewaetere, M., & Clarebout, G. (2014). Advanced technologies for personalized learning, instruction, and performance. In *Handbook of research on educational communications and technology* (pp. 425-437). Springer, New York, NY.

van Roy, R., & Zaman, B. (2017). Why gamification fails in education and how to make it successful: introducing nine gamification heuristics based on self-determination theory. In *Serious Games and edutainment applications* (pp. 485-509). Springer, Cham.

Verhagen, W. G. M., Aarnoutse, C. A. J., & Van Leeuwe, J. F. J. (2009). The predictive power of phonemic awareness and naming speed for early Dutch word recognition. *Educational Research and Evaluation*, *15*(1), 93–116.

Viljaranta, J., Kiuru, N., Lerkkanen, M. K., Silinskas, G., Poikkeus, A. M., & Nurmi, J. E. (2017). Patterns of word reading skill, interest and self-concept of ability. *Educational Psychology*, *37*(6), 712–732.

Viljaranta, J., Tolvanen, A., Aunola, K., & Nurmi, J. E. (2014). The developmental dynamics between interest, self-concept of ability, and academic performance. *Scandinavian Journal of Educational Research, 58*(6), 734–756.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., … Garon, T. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: a 5-year longitudinal study. *Developmental Psychology, 33*(3), 468–479.

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers and Education, 58*(4), 1183–1193.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van Der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*(2), 249–265.

Wouters, P., & Van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. *Computers and Education, 60*(1), 412–425.

Data availability statement

Data available on request due to privacy/ethical restrictions.