The effects of two digital educational games on cognitive and non-cognitive math and reading outcomes

Stefanie Vanbecelaere[2,3*], Katrien Van den Berghe[3], Frederik Cornillie[3], Delphine Sasanguie[1], Bert Reynvoet[1], & Fien Depaepe[2,3]

[1] Brain & Cognition, KU Leuven Kulak, Kortrijk, Belgium

[2] Centre for Instructional Psychology and Technology, KU Leuven Kulak, Kortrijk, Belgium

[3] ITEC, KU Leuven, Belgium, also at imec Kapeldreef 75, B-3001 Leuven, Belgium

Author note

*Corresponding author. Faculty of Psychology and Educational Sciences, KU Leuven Kulak, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium. Tel: +32 56246235.

E-mail address: stefanie.vanbecelaere@kuleuven.be

URL: https://www.kuleuven-kulak.be/nl/onderzoek/itec

Abstract

Digital educational games play an increasingly important role in education. However, multiple questions about the effectiveness of educational games with respect to cognitive and non-cognitive effects remain unclear. The current study, a longitudinal, quasi-experiment with 336 first graders, examined the effects of two digital educational games, Number Sense Game (NSG) and Reading Game (RG). The NSG trained early numerical skills, the RG supported emergent reading. Children were pseudo-randomly assigned to either an experimental condition, comprising eight weeks of intensive game-based training, or a control condition in which they took part in regular education without game-based practice. A pretest-posttest design was used to examine the effects of the intervention on cognitive (digit comparison, number line estimation, letter knowledge, math and reading competence) and non-cognitive outcomes (math and reading anxiety). Delayed cognitive effects on math and reading competence were also investigated two months after the intervention. Furthermore, we examined variances of the impact of the training on cognitive outcomes as a consequence of differences in children's prior knowledge, prior affect and socio-economic status. For cognitive outcomes, results revealed that children who played a game performed better on number line estimation and reading competence, whereas no significant differences were observed for digit comparison, letter knowledge and math competence. Also, children who played a game showed better scores in the delayed reading posttest, but not in the delayed math posttest. For non-cognitive outcomes, game training did not affect math or reading anxiety. Regarding individual differences, children with less prior knowledge in the game play condition performed better on the number line estimation posttest compared to children in the control condition. Children with more prior knowledge in the game play condition still scored better on this test compared to the control condition, but the difference between the conditions was smaller.

Keywords: elementary education, improving classroom teaching, interactive learning environments, multimedia systems

1 Introduction

There is a growing body of literature that recognizes the importance of digital game-based learning (DGBL). DGBL refers to the usage of games to serve educational purposes (Zyda, 2005). It is generally assumed that digital educational games can positively contribute to children's academic knowledge (e.g., mathematics and reading), because "learning is most effective when it is active, experiential, situated, problem-based and provides immediate feedback" (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012, p. 661), features that are inherently connected to DGBL. However, several systematic reviews (e.g., Connolly et al., 2012; Kebritchi, Hirumi, & Bai, 2010) and meta-analyses (e.g., Clark, Tanner-Smith, & Killingsworth, 2016; Girard, Ecalle, & Magnan, 2013; Wouters, van Nimwegen, van Oostendorp, & van Der Spek, 2013) have revealed that the empirical evidence regarding the effectiveness of digital educational games is mixed.

In addition, methodological limitations of previous studies on DGBL further hamper an unambiguous interpretation of the findings in the domain. More particularly, prior research often investigated immediate, near transfer cognitive outcomes of one particular game, which contributes to only a part of our understanding about the effectiveness of DGBL (Girard et al., 2013). In the current study, we conducted a large-scale, longitudinal study in which first graders trained with a numerical and a reading game. The current paper consists of two main contributions in the domain of effectiveness of DGBL. First, the investigation focused on the effectiveness of DGBL in a very broad sense, including both cognitive (near transfer domain-specific cognitive factors, and far transfer general math and reading outcomes) and non-cognitive (math and reading anxiety) outcomes as well as immediate and delayed effects. In addition, the extent to which individual differences moderate the intervention effects was examined. Second, this study attempted to transcend the particularity of specific educational games by investigating the effects of two different educational games in two different domains (i.e., early math and early reading). Consequently, the effects of this study might have more potential to be generalized than studies with the focus on a single subject.

2 Theoretical background

Young children have been the primary target of intensive interventions to train early math and reading skills (Kadosh, Dowker, Heine, Kaufmann, & Kucian, 2013; Wanzek et al., 2018). This is crucial because early math and reading skills are important predictors of math and reading competence with lasting effects in performance during children's future school career (Melby-Lervåg, Lyster, & Hulme, 2012; Sasanguie, Göbel, Moll, Smets, & Reynvoet, 2013).

2.1 DGBL in early math

Early numerical skills which are predictive for later math competence are numerosity comparison, symbolic number comparison, number line estimation (NLE), and understanding the mapping between symbolic numbers and their corresponding numerosity (De Smedt, Noël, Gilmore, & Ansari, 2013; Sasanguie et al., 2013). Two meta-analyses by Schneider et al. (2017, 2018) revealed that especially comparison and number line tests are reliably associated with mathematical competence. Both tests can be conducted with either non-symbolic stimuli (e.g., dot arrays) or symbolic stimuli (e.g., digits or number words). In comparison tasks, participants have to indicate which of two presented numbers is the largest. In a NLE test, they have to place a presented number on an empty number line.

Several empirical studies used DGBL to improve early numerical skills. Kim, Jang, and Cho (2018) randomly assigned 56 first graders to an experimental condition (training numerical processing) or a control condition (following regular education). The experimental condition outperformed the control condition in non-symbolic comparison, but not in other early numerical skills (e.g., approximate arithmetic). Also, there was no transfer to math competence (e.g., exact calculation). Another study assigned 147 first graders to four conditions (Obersteiner, Reiss, & Ufer, 2013). Children of three experimental groups played a different version of the computer game Number Race (i.e., training only symbolic, only non-symbolic number processing, or both) and a control group (i.e., playing a language game). Results revealed improvements on the trained numerical skills, but not on other, untrained skills. In none of the conditions, an effect was observed in general math achievement. Van der Ven, Segers, Takashima, and Verhoeven (2017) pseudo-randomly assigned 103 first graders to either an experimental condition (training calculation fluency in a game environment) or to a control condition (continuing regular education). Children of the experimental condition significantly outperformed children of the control condition on dot subtraction efficiency, but not on numerical calculation. The intervention effect disappeared three months after the intervention. Also studies focusing on children with low prior knowledge (Räsänen, Salminen, Wilson, Aunio, & Dehaene, 2009) or children from low SES families (Wilson, Dehaene, Dubois, & Fayol, 2009) reported significant gains in favour of the DGBL condition compared to the control group on number comparison skills, but not in numerosity comparison or general math competence.

Next to optimising children's cognitive learning gains, it is hypothesized that educational games can also affect non-cognitive factors (Wouters et al., 2013). Math anxiety (MA) for example already appears in young children, and can negatively impact math competencies (Gunderson, Park, Maloney, Beilock, & Levine, 2018; Verkijika & De Wet, 2015). Some researchers have argued that children might feel less anxious in a DGBL context (Sun & Pyzdrowski, 2009). However, studies investigating this hypothesis are scarce (Dowker, Sarkar, & Looi, 2016; Kebritchi et al., 2010). Jansen et al. (2013) investigated whether 207 children of grades 3-6 experienced less MA when playing Math Garden, a digital game for practicing math, compared to a control condition in which children practiced math as usual. It was observed that MA significantly decreased in both conditions, but no differences were observed between the conditions. By contrast, results of a within-subjects longitudinal design with 36 students (age 10-16) revealed that students' MA could be reduced through DGBL when MA was assessed with physiological data during game play (Verkijika & De Wet, 2015).

Some studies also examined whether student factors moderated the effect of DGBL. In the study of Miller and Robertson (2011), 634 primary school children were assigned to an experimental condition (training calculation fluency with a game) or to a control condition. They observed that learning effects of DGBL vary depending on children's prior knowledge in favor of children with lower math competence. By contrast, Kebritchi et al. (2010) investigated the effect of a math game compared to a control game following regular math instruction in 193 high school students. They observed that prior knowledge, computer skills, and English language skills did not affect the game-playing children's achievement and motivation.

2.2 DGBL in early reading

Reading is a complex cognitive process for which phonological awareness and letter knowledge are domain-specific cognitive predictors for individual differences in children's later reading development (Melby-Lervåg et al., 2012; Sasanguie et al., 2013). 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).

DGBL has been used for enhancing emergent reading. For instance, Kyle, Kujala, Richardson, Lyytinen and Goswami (2013) assessed the effectiveness of two subgames of Graphogame by assigning 31 poor readers of 6-7 years old to either a game training rhyming skills, a game training letter-sound correspondences, or a control condition following regular education. They observed that both subgames improved immediate and delayed (after four months) reading, spelling and phonological skills in English compared to the control condition. In another study, 60 2nd and 3rd graders with special educational needs played the game ‘Letter Prince’ to train their early literacy skills (van de Ven, de Leeuw, van Weerdenburg, and Steenbeek-Planting, 2017). Results revealed immediate enhanced pseudoword reading fluency as well as improved text reading fluency in the DGBL condition compared to the control condition, but an effect on decoding of existing words was not observed. Delayed effects were observed on text reading fluency in favour of the DGBL condition. Van Gorp, Segers and Verhoeven (2017) trained 62 poor readers of the second grade with the ‘Racing game’, a game focusing on word decoding. Results revealed that the game condition outperformed the control condition on word decoding efficiency. These effects were maintained five weeks later. Kartal and Terziyan (2016) examined the effects of a game condition practicing phonological awareness compared to a control condition playing math games. In total 20 children of low-SES families participated in this study. The experimental condition outperformed the control condition on near transfer tests (e.g., phoneme segmentation, letter-name), but not on far transfer tests (e.g., rhyme, syllable blending).

Studies on the effects of DGBL on non-cognitive factors, such as reading anxiety (RA) are to date remarkably scarce (Punaro & Reeve, 2012), especially given the fact that RA seems to be an obstacle in early reading education. In this respect, Ramirez and colleagues (2019) observed in a longitudinal study with 607 1st and 2nd graders, that higher RA in the beginning of the school year was related to lower reading achievement later on.

Only a small number of studies investigated moderating effects of personal and environmental factors of DGBL in the domain of early reading. Concerning ethnicity, Segers and Verhoeven (2005) assigned 100 native Dutch and immigrant preschool children to a game training phonological awareness or to a control condition and evaluated them on various early literacy skills. They showed a significant three-way interaction between time, intervention and ethnicity indicating that the immigrant children (with lower rhyming skills at pretest) benefited more from the DGBL intervention compared to native speakers on the rhyming posttest. Concerning the impact of children's initial anxiety level, Yang, Lin, and Chen (2018) examined the effect of 43 fourth graders' language anxiety level on their learning performance in a game-based English learning system. They showed that high-anxiety learners performed worse than low-anxiety learners, but that DGBL was particularly beneficial to high-anxiety learners.

2.3 Gaps in research with respect to DGBL

Four unresolved issues in studies investigating the impact of DGBL can be detected. First, the effectiveness of DGBL is often investigated in a rather narrow sense, i.e. assessing whether the targeted skills during game play are improved; the so-called 'near transfer'. However, Clark et al. (2016) suggested that one should also assess to which extent the trained skills have been transferred to other, more general skills, such as math and reading competence in the current context (i.e., 'far transfer'). The few intervention studies evaluating far transfer effects however reported no effect on math competence (e.g., Kim et al., 2018; Obersteiner et al., 2013; Räsänen et al., 2009) and mixed results for reading (e.g., van de Ven et al., 2017).

Second, studies typically assess the effect of training immediately after the intervention, but most studies lack delayed posttests (Girard et al., 2013). Delayed assessment refers to children's skills, practiced in training, several weeks or months after leaving the training environment, mainly in order to control for short-term effects (All, Nuñez Castellar, & Van Looy, 2016). To date, some studies already reported sustained effects of game training, but more robust evidence is needed (Räsänen et al., 2009; Segers & Verhoeven, 2005).

Third, straightforward empirical evidence regarding the impact of DGBL on math and reading anxiety is inconclusive (Wouters et al., 2013). Young et al. (2012) stated that especially the very short training periods often used in studies might make it difficult to observe effects on children's non-cognitive factors with respect to learning (e.g., math anxiety, reading anxiety). Instead, longitudinal studies should be conducted.

Fourth, prior DGBL research mainly focused on the general effectiveness of educational games, not accounting for individual differences (All et al., 2016; Clark et al., 2016; Kadosh, et al., 2013; Yang & Quadir, 2018). Important factors which influence learning are prior knowledge, prior affect and SES (Chung, 2015; Miller & Robertson, 2011; Yang & Quadir, 2018). It is important to ensure that DGBL is as beneficial for good performing children or children with middle/high SES as for children with lower prior knowledge or low SES, because these children often perform low on executive functioning skills (Chung, Liu, McBride, Wong, & Lo, 2017).

Consequently, the present study addresses the above-mentioned gaps by focusing on four research questions:

1. What is the effect of the training on near and far transfer tests with regard to early math and reading?
2. What is the effect of the training after two months with regard to early math and reading?
3. What is the effect of the training on children's non-cognitive factors math and reading anxiety?
4. Does the impact of DGBL depend on children's individual differences?

4.1. Does children’s *prior knowledge* influence the immediate effect of DGBL on children’s performance?

4.2. Does children’s level of *math and reading anxiety* influence the immediate effect of DGBL on children’s performance?

4.3. Does *socio-economic status (SES)* influence the immediate effect of DGBL on children’s performance?

3 Methodology

3.1 Two digital educational games

In a recent research project (imec-ICON project LEAPS 2016-2018; Learning Analytics for AdaPtive Support), we investigated the effect of two digital educational games in math and reading: i.e., “the number sense game” (NSG) and “the reading game” (RG). The NSG was originally developed in view of a research project[1]. The RG was created by the publishing company Pelckmans (https://www.pelckmans.be/) and Sensotec (https://sensotec.be/). Both games provided the children with the necessary instructions in order that all children could autonomously interact and navigate through the game. In what follows, the content of the two games, the level structure and the instructional support integrated in the games are described.

3.1.1 NUMBER SENSE GAME (NSG)

The number sense game (NSG) (Linsen et al., 2015; Maertens, De Smedt, Sasanguie, Elen, Reynvoet, 2016) contains two types of exercises, namely a comparison game and a NLE game (see Figure 1). The levels of both the comparison game and the NLE game were presented in a fixed order and were characterized by increasing difficulty. In the comparison game, the levels were designed to vary in difficulty based on the numerosities (1-4, 1-9, 5-18), the display duration and the type of stimuli (non-symbolic, symbolic and mixed notation) used in the tasks. The difficulty of the levels in the number line game depended on the number of anchor points, the display duration, and the type of stimuli (non-symbolic, symbolic, and mixed notation). A detailed overview of the levels in the game can be found in Linsen et al. (2015, pp. 14-22).

[1] KU Leuven-GOA2012/010, coordinator Lieven Verschaffel

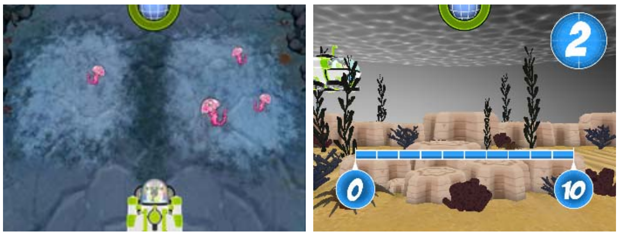


Figure 1. Screenshots of the comparison game (left) and the NLE game (right)

The NSG contained specific criteria which the participants needed to meet before they could go to the next level. If not they had to replay the level until the target score was obtained. For the comparison task, a trial was considered correct when the player selected the larger out of two numerosities. Children should obtain at least 80 percent accuracy before moving to the next level. For the NLE tests, the answer was considered correct when the indicated position of the number did not deviate more than 12.5 percent from its actual place on the number line. To avoid children getting stuck in a level because their performance was too low, the cut-off score to move to the next level was set to 50 percent.

In the NSG, auditory feedback was given every time a child gave an answer and a voice-over was used to encourage children when they waited too long before answering the exercise. Visual feedback was given by a blue bar on top of the screen. When children advanced through the level by playing, the blue bar changed accordingly. Furthermore, a narrative was included to make the game more attractive. At the start of the game, children get acquainted with two characters (Dudeman and Sidegirl) through a short video. Those two characters wanted to save the polluted world by using a vacuum cleaner to clean the world under water, on land and in the air. Children were invited to help them by doing exercises. Each time an exercise was answered correctly, the vacuum cleaner could make the area cleaner. At the beginning of each level, a voice-over explained the goal of the task, and this explanation was adapted to the characteristics of each level. Finally, personalization was realized by using cut-off scores that must have been obtained before moving to the next level.

3.1.2 READING GAME (RG)

The Reading Game (RG) aims to support emergent reading, and more specifically phonological awareness and letter knowledge (see Figure 2). Children were presented with tasks training these skills. For example, children trained phonological awareness, a.o. by hearing a word in parts, and were instructed to synthesize the words and to find the corresponding picture (left picture in Figure 2). Letter knowledge was trained by tasks involving particular letters and graphemes. One example of a task was when children hear a letter and have to choose the grapheme among four graphemes (right picture in Figure 2).

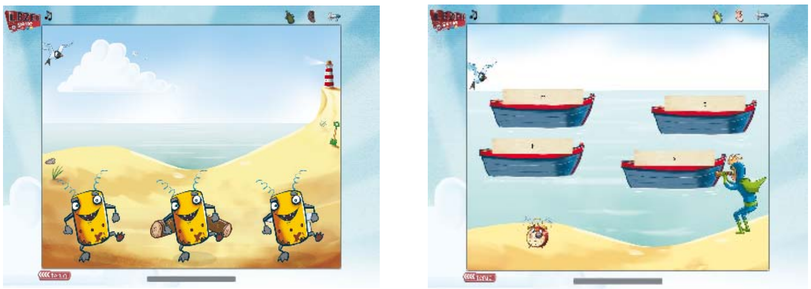


Figure 2. Screenshots of subgames in the RG

The RG consists of 65 levels in total, and within each level children could choose among three or four subgames which train the same content. All children started with level one and advanced independently level by level at their own pace. Children could earn a star each time they finished a subgame. A subgame was finished when children answered 80 percent of the trials correctly. When three stars were collected (i.e., when three subgames within a level were accomplished), the next level was unlocked and the student could continue to the next level (see Figure 3). If children did not pass a level, they were instructed to repeat the subgame, or to choose a comparable subgame within the same level.

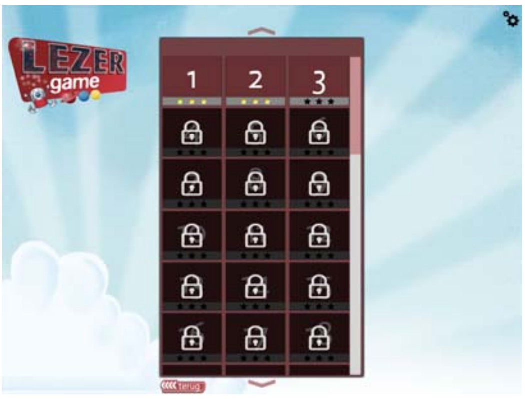


Figure 3. Screenshot of the learner's dashboard

Similar to the NSG, the RG provided instructional support in several ways. First, children received visual feedback through the horizontal bar displayed at the bottom of the screen. If a correct answer was given, the blue color in the bar moved accordingly. In case of a wrong answer the bar remained the same. Auditory feedback was provided by a different sound when a correct or wrong answer was given. Second, the game contained narrative elements. Three superheroes Run, Bit and Zip appeared as illustrations in the games (e.g., game "River": Run looming in the horizon watching boats). Third, interactivity was integrated so that children could experience control and could make decisions on the playing elements. Children could seek for help by choosing in the upper right corner for the 'machine', 'ear' or 'speaker'. If the machine was selected, they wanted some help with the exercise; when clicking the ear, the letter or word was repeated; selecting the speaker gave a repetition of the instructions. Fourth, modality was integrated by using audio channels for giving instructions, prompting words and letters etc. Fifth, personalization was addressed as children for each level could choose the subgame that they wanted to play. Also, cut-off scores were used that must have been obtained before children could move to the next level.

3.2 Participants and procedure

Participants were first graders 336 children from 10 primary schools in Flanders (157 boys, M_{age} = 6.37 years, *SD* = 0.42). Due to the ecologically valid nature of the study (i.e., where the games were implemented in existing class groups), classes (*N*=18) were randomly assigned as an intact group to the experimental condition or to the control condition (i.e., cluster randomized trial). The study took place at the start of Grade 1 at school during school hours. The experimental condition received both the digital game-based math and reading training and consisted of respectively 109 children[2] and 223 children. The control condition consisted of 113 children. Subjects were excluded from the study (*n*=28) if their parents did not accept the active informed consent or when they were not able to complete all pretests and/or posttests.

A pretest, posttest, delayed test design with an experimental and control condition was used to investigate the effects of two digital educational games on math and reading achievement. The outline of the study is shown in Figure 4. After the pretest measurements, the children of the experimental condition individually played six times NSG for 50 minutes, and eight[3] times RG for 50 minutes using a tablet, during a period of eight weeks. The content of the games was aligned with the Flemish curriculum in a sense that children start formal reading education at the start of the 1st grade. The control condition continued regular education, but with the explicit restriction of making use of educational games during that period. The effects of the intervention were measured at two time points, once one week after the intervention period (i.e., posttest) and again eight weeks after the first posttest (i.e.,

[2] The difference in sample size between the math training and the reading training is due to logging problems with the pilot version of the adaptive version of NSG, so that only half of the group who played the original version of the NSG remained (*N*=109). Half of the children of the experimental group (*N*=114) played a pilot version of NSG with integrated adaptivity, in this game children were presented with exercises that were related to each individual ability.
[3] The RG contained of more exercises which was the reason why we organized two more game interventions with the RG compared to the number of interventions with the NSG.

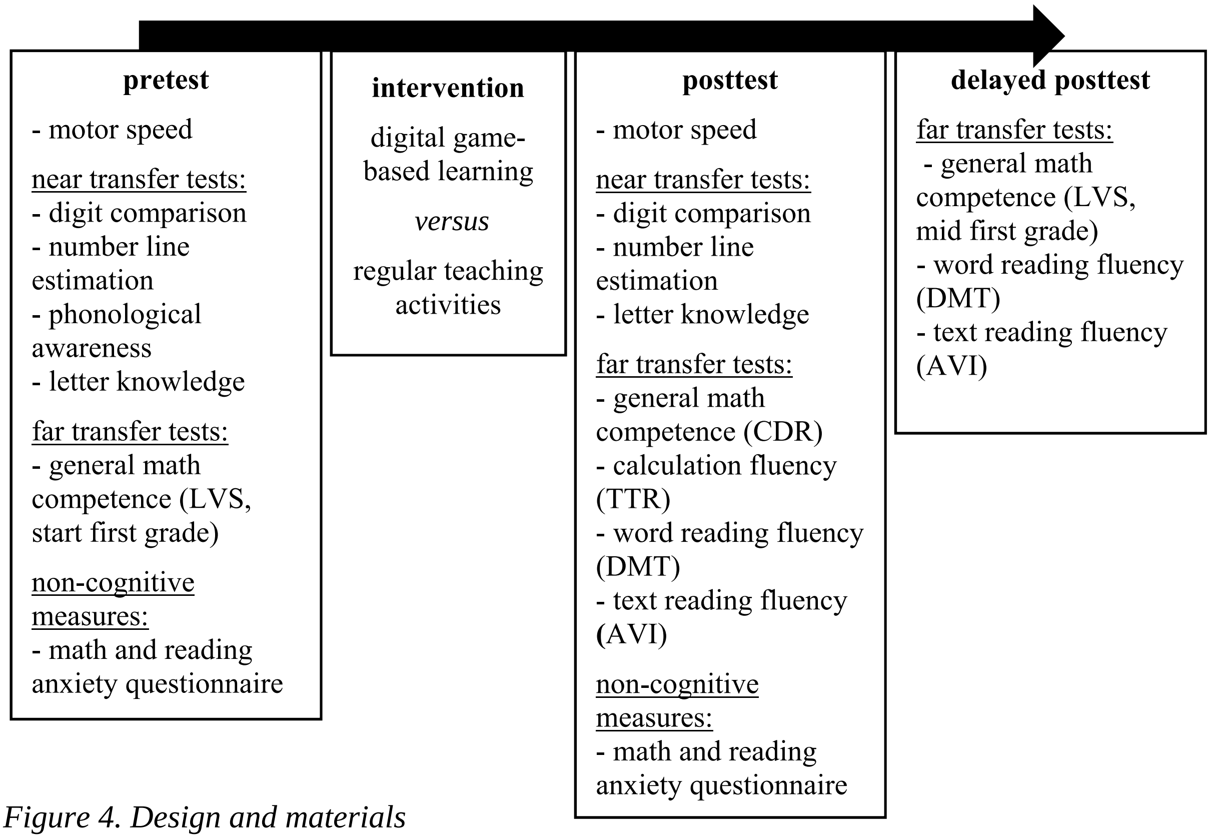
delayed posttest). We made use of an extensive test battery, therefore, some actions were taken to diminish the test load. First, the tests were spread over a period of five months. Second, most of the measurements did not take a lot of time (e.g., digit comparison only lasted for one minute). Third, prior to the study, we asked the teachers which tests they were obliged to conduct in their class irrespective of this particular study. Consequently, we made a decision based on the tests most of the teachers would have conducted in their class anyway (for example the LVS, the DMT and the AVI tests). Finally, we made sure that children had enough breaks between the tests. It can be concluded that children did not get tired during the administration of the tests and this was confirmed in the data as there was no systematic drop-out of children.

The children of the experimental and the control group spent the same amount of time on practicing early numerical skills and emergent reading during the regular math and reading lessons. As the content that was trained in the game conditions is part of the 1st grade curriculum, the children from the control condition practiced the same skills as the experimental condition. For early numerical skills, the educational methods used by the teachers also contained (symbolic and non-symbolic) number comparison, and to a lesser extent the number line content (cf. personal communication with teachers). For reading, the same letters as exercised in the RG were taught but the order in which these letters were taught varied according to the educational method. Importantly, in both conditions, children had no prior experience with the NSG and the RG as the first game was only used for scientific purposes so far, and the latter was a very recently developed game which the participating schools in this study did not purchase yet. Although it might be the case that children gained experience with similar (but probably more basic, freely available) apps training early math and reading at home, informal talks with the participating children learned us that they had only limited or no prior experience with math and reading games.

3.3 Materials

3.3.1 NEAR AND FAR TRANSFER TESTS

To examine near and far transfer effects, several validated instruments were used. Near transfer tests are tests that are similar to the tasks presented in the game (i.e., numerical comparison, NLE, letter knowledge). Far transfer tests measure the transfer of the trained tasks to higher-order learning outcomes such as general math and reading competence.



pretest
- motor speed

near transfer tests:
- digit comparison
- number line estimation
- phonological awareness
- letter knowledge

far transfer tests:
- general math competence (LVS, start first grade)

non-cognitive measures:
- math and reading anxiety questionnaire

intervention

digital game-based learning

versus

regular teaching activities

posttest
- motor speed

near transfer tests:
- digit comparison
- number line estimation
- letter knowledge

far transfer tests:
- general math competence (CDR)
- calculation fluency (TTR)
- word reading fluency (DMT)
- text reading fluency (AVI)

non-cognitive measures:
- math and reading anxiety questionnaire

delayed posttest

far transfer tests:
- general math competence (LVS, mid first grade)
- word reading fluency (DMT)
- text reading fluency (AVI)

Figure 4. Design and materials

3.3.1.1 Near transfer tests

Early numerical skills were tested with a digit comparison (Brankaer, Ghesquière, & De Smedt, 2017) and a NLE test (Siegler & Booth, 2004).

Digit comparison. 60 digit pairs were presented in four columns of 15 pairs each. Only single digits (1-9) were included and the distance between both digits was one on half of the items and three or four on the other half. All possible combinations with these distances were included. The number pairs were randomly presented. During the test, participants were asked to indicate the larger of the two digits. They were given 30 seconds to solve as many items as possible. Brankaer et al. (2017) reported satisfactory test-retest reliability (correlations between .72 and .86) and showed significant and stable correlations with math achievement, indicating satisfactory construct and criterion-related validity of the test. Because children's performance on the test could be influenced by general motor speed, a paper-and-pencil test to control for motor speed was taken (Brankaer et al., 2017). General motor speed and digit comparison skills were measured before and one week after the intervention (see Figure 4).

Number line estimation. Children were instructed to place a digit on an empty number line. The line was 25 cm long and labeled by 0 on the left side and by 10 on the right. The to-be-positioned number was presented in the middle of the page, 6 cm above the number line. All single digits (1-9) were presented to the children in a random order, resulting in a total of nine trials. In line with previous studies (e.g., Siegler & Booth, 2004), we computed the percent absolute error (PAE), as the index of number line performance using the formula of Siegler and Booth (2004):

$$\text{PAE} = \left|\frac{estimate - target\ number}{scale\ of\ estimates}\right| \text{ X } 100$$

NLE skills were measured before and one week after the intervention. In this study, Cronbach's alpha of the pretest and posttest data was respectively .70 and .84 indicating a good internal consistency of the test.

To evaluate children's emergent reading, phonological awareness and letter knowledge, different tests were used.

Phonological awareness. 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] (Aarnoutse, Beernink, & Verhagen, 2016). The test contained 24 items and the reliability of this test was considered good (Cronbach's alpha =.89; Verhagen, Aarnoutse, and Van Leeuwe, 2009). This test was only taken as a pretest due to a ceiling effect.

Letter knowledge. This test measured to which extent children could correspond letter sounds to graphemes (Aarnoutse et al., 2016). The children were presented with caterpillars containing 22 different graphemes in the different segments of each caterpillar. Children had to listen to the letter the instructor prompted and to color the corresponding grapheme in the caterpillar. The test contained 21 items and had a Cronbach's alpha of .90 (Verhagen et al., 2009). Letter knowledge skills were assessed before and one week after the intervention.

3.3.1.2 Far transfer tests

General math competence. The general math competence test was a curriculum-based standardized test for mathematics from the Flemish Student Monitoring System (*LeerlingVolgSysteem* - LVS, Van Rompaey & Vandenberghe, 2015). Usually, schools conduct this test three times a year at fixed moments to monitor children's general math achievement. In this study, the LVS tests of the first two test moments, i.e. "start 1st grade" and "middle 1st grade", were used. These tests consist of respectively 43 and 91 items (measuring number sense, arithmetic word problems, estimation, number decomposition, and number series). The first LVS test (start 1st grade) was used as a pretest for math competence, the test "middle 1st grade" served in this study as a delayed test of math competence 8 weeks after the training (see Figure 4). Internal consistency of the item scores was good as reported in previous studies (For both tests, resp..70 and .80; Van Rompaey & Vandenberghe, 2015). As a direct posttest, another general math test was used because there was no LVS available for that measurement time. The Cognitive Developmental aRithmetics test (*Cognitive Deelvaardigheden Rekenen* - CDR, Desoete, Roeyers, 2006), is a curriculum-based standardized test for children of the first grade assessing simple calculations in a number-problem format (e.g., 6+3=), a word-problem format (e.g., 1 less than 5 is…), or a context-problem format (e.g., the farmer has 6 chickens, he sells 2 chickens, how many chickens are left?). This pen-and-paper test consists of 25 items and Cronbach's alpha was .93 with children of the 1st grade in the study of Desoete and Roeyers (2006).

Calculation fluency. In addition to the general math tests, we also administered the Arithmetics Number Facts Test (*TempoToets Rekenen* – TTR, De Vos, 1992) as an immediate posttest (see Figure 4). The first graders only completed the first two subtests on addition and subtraction problems. Both subtests consisted of 40 problems of increasing difficulty. For each subtest, children were instructed to solve as many problems as possible in 1 minute. Cronbach's alpha was .86 for both the addition and subtraction subtest as reported in the study of De Vos (1992).

Word reading fluency. Word reading fluency was assessed by the Three-Minutes-Test (*Drie Minuten Toets* - DMT, Moelands, Kamphuis, & Rymenans, 2003), containing three reading cards with different word types of vowel (V) and consonant (C) combinations that must be read aloud during 1 min per card. The number of correctly read words in 1 minute was registered for each card. The words on each card differed in complexity. The reliability of the test is high (Cronbach's alpha for grade 1 between .86 and .90, Moelands et al., 2003). Word reading fluency was evaluated as immediate and delayed posttest.

Text reading fluency. In the Analysis of Individual Word Forms test (*Analyse Van Individualiseringsvormen* - AVI, Krom, Jongen, Verhelst, Kamphuis, & Kleintjes, 2010) children were instructed to read a short text containing short sentences as quickly and accurately as possible. The number of mistakes and the total reading time were measured. The reliability varied from .86 to .93 in the study of Krom et al. (2010). Children conducted this test as immediate and delayed posttest.

3.3.2 MATH AND READING ANXIETY

Math anxiety. We adapted the existing Child Math Anxiety Questionnaire – Revised (CMAQ-R, Ramirez, Chang, Maloney, Levine, & Beilock, 2016) to a new questionnaire for Flemish 1st graders. More particularly, the original questionnaire was translated to Dutch, the number of items were reduced (from 16 items to 8 items) and the level of the math problems presented in the items was adapted to the Flemish math curriculum of the 1st grade.

Reading anxiety. Parallel to the Flemish MA questionnaire, also a RA questionnaire was developed, consisting of 8 items in line with the Flemish reading curriculum. In the present study, the MA and RA questionnaire were completed before the training, and immediately after the training. The MA scale (8 items) and the RA scale (8 items) have a Cronbach's alpha of respectively .77 and .77 at pretest. The Cronbach's alpha's at posttest were respectively .75 and .75. More information about the MA and RA questionnaire can be found on the following web page: [https://numcoglableuven.be/static/materials/Procedure%20MA%20&%20RA%20questionnaire.pdf].

3.3.3 INDIVIDUAL DIFFERENCES

To examine the moderating effect of individual differences, we investigated possible effects of prior knowledge, anxiety levels and SES. The first two factors characterize the child at the individual level and SES gives more information about a child's home environment.

Prior knowledge. Prior knowledge was operationalized by using the pretest measurements of the near transfer tests (i.e., digit comparison, NLE, letter knowledge).

Math anxiety and reading anxiety. MA and RA were operationalized by using the pretest measures of respectively the MA questionnaire and the RA questionnaire.

Socio-economic status. This factor 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 a 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).

4 Data-analysis

The following sections address the data-analyses which were conducted to answer the research questions. First, to test the intervention's learning effect on *near* transfer tests, three Repeated Measures Analyses of Variance (ANOVA) were conducted with digit comparison, NLE, and letter knowledge as dependent variables and condition (experimental versus control) as between subjects factor and pre-post time effect as within subjects variable. For digit comparison, motor speed was additionally included as covariate. The effect on *far* transfer (i.e., general math competence, calculation fluency, simple word reading fluency, complex word reading fluency, text reading fluency) was examined with a one-way ANOVA with condition (experimental versus control) as factor.

To address the second research question, four one-way ANOVAs were conducted with the four delayed test results (i.e., general math competence, simple word reading fluency, complex word reading fluency and text reading fluency) as dependent variables and condition (experimental versus control) as between-subjects factor.

Third, to determine whether the use of educational games affected children's non-cognitive factors, a Repeated Measures ANOVA was conducted on children's MA and RA scores with test moment (2 levels: pretest versus posttest) as within-subjects variable and condition (experimental versus control) as between-subjects factor.

Finally, to examine whether children's (1) prior knowledge, (2) math and reading anxiety and (3) their family's SES moderated the effect of training, we conducted separate Analyses of Covariance (ANCOVA). For these analyses, we focused on the variables on which an effect was observed of condition (i.e., if the experimental condition outperformed the control condition) immediately after the intervention. This was the case for the following dependent variables: NLE, simple word reading fluency and text reading fluency. In total nine ANCOVAs were ran to analyze whether prior knowledge, MA and RA and SES moderated the effect of the training. For prior knowledge and anxiety the pretest scores were used. For SES the mother’s highest degree of education was used. Condition was entered as independent variable. Posttest scores of respectively NLE, word reading fluency (measured with DMT) and text reading fluency (measured with AVI) were used as dependent variables.

5 Results

Appendix A shows the means and standard deviations for each condition on the different tests (i.e., pretest, posttest and delayed test results of the near transfer tests, far transfer tests and non-cognitive measures). Correlations between the pretest variables (i.e., near transfer tests, far transfer tests, non-cognitive measures and degree of mother) are presented in Appendix B.

One-way ANOVAs on all pretest measures showed no differences between the experimental and control conditions: digit comparison ($F(1, 218) = 0.35$, $p = .56$), NLE ($F(1, 217) = 2.50$, $p = .12$), phonological awareness ($F(1, 330) = 0.06$, $p = .82$), letter knowledge ($F(1, 326) = 0.02$, $p = .88$), general math competence (LVS) ($F(1, 215) = 0.00$, $p = .97$), MA ($F(1, 213) = 3.02$, $p = .08$), RA ($F(1 ,319) = 0.68$, $p = .41$).

5.1 Immediate effects on *near* and *far* transfer tests

5.1.1 NEAR TRANSFER EFFECTS

With respect to NSG, we observed a main effect of test moment for digit comparison, $F(1, 214) = 54.21$, $p \leq 01$, $ŋ_p^2 = .20$. However, there was no test moment × condition interaction, indicating that children of the experimental and the control condition similarly improved on the digit comparison test. For the symbolic NLE test, a main effect of test moment, $F(1, 215) = 109.38$, $p \leq .01$, $ŋ_p^2 = .34$ and an interaction between test moment and condition was found, $F(1, 215) = 60.30$, $p \leq .01$, $ŋ_p^2 = .22$ showing that the children from the experimental condition outperformed those of the control condition after the training.

The effectiveness of the RG on letter knowledge revealed a main effect of test moment, $F(1, 322) = 919.07$, $p \leq .01$, $ŋ_p^2 = .74$, but no significant test moment × condition interaction effect.

5.1.2 FAR TRANSFER EFFECTS

For the NSG intervention, a significant difference between the conditions on the CDR was observed, but not in expected direction. The children of the control condition outperformed those of the experimental condition, $F(1, 217) = 5.72$, $p \leq .05$, $ŋ_p^2 \leq .05$ (see Appendix A). No differences were found for calculation fluency between the experimental and the control condition.

The far transfer tests after the RG training resulted in significant differences in favor of the experimental condition. There was a significant difference between the experimental and the control condition for simple word reading fluency immediate after the intervention, $F(1, 331) = 4.77$, $p \leq. 05$, $ŋ_p^2 \leq .01$. There were no differences between the experimental and the control condition for complex word reading fluency immediately after the intervention. Furthermore, children of the experimental condition performed better on text reading fluency than children of the control condition after the training ($F(1, 330) = 4.13$, $p \leq .05$, $ŋ_p^2 \leq .01$).

5.2 Delayed effects on *far* transfer tests

For the NSG, no differences on the general math test (LVS) between the experimental and the control condition were observed. For the RG, results showed that the experimental condition outperformed the control condition on simple word reading fluency after the training ($F(1, 329) = 9.13$, $p \leq .01$, $ŋ_p^2 \leq .05$), whereas no significant difference was observed for complex word reading fluency and text reading fluency.

5.3 Effects on math and reading anxiety

For MA, a main effect of test moment was present, $F(1, 206) = 14.06$, $p \leq .01$, $ŋ_p^2 = .06$, indicating that children were significantly less anxious after the intervention. There was no interaction between test moment and condition demonstrating that children’s anxiety was reduced to the same extent in both conditions. For RA, there was no main effect of test moment and no test moment x condition interaction effect, indicating that the DGBL intervention did not impact RA.

5.4 Effects of individual characteristics on the intervention effects

5.4.1 PRIOR KNOWLEDGE

Table 1 presents the results of three ANCOVAs. The main effects show that children’s prior knowledge strongly predicts their results for number line estimation, word reading fluency and text reading fluency, whereas condition did not yield a significant main effect. Furthermore, results revealed that only for the NLE test a significant interaction was observed between prior knowledge and condition, indicating that children with low NLE pretest scores in the experimental condition benefitted more from the intervention compared to the children with low NLE pretest scores in the control condition (Figure 5). Also, children with high NLE pretest scores in the experimental condition still performed better on posttest compared to high achievers in the control condition, but the difference between pre and posttest NLE scores was smaller than for the low achievers. To interpret Figure 5, it is important to know that *the lower the score* on the pre- and posttest number line estimation, *the better the student performed* on this test.

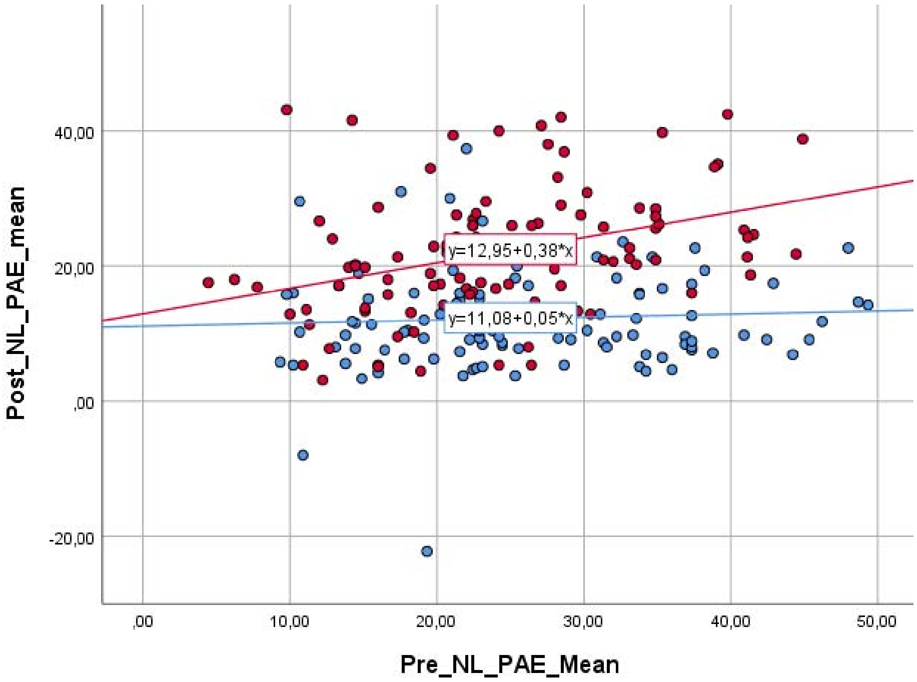


Figure 5. The interaction of prior knowledge and condition on the posttest scores of the NLE test. Blue: experimental condition, Red: control condition

With respect to simple word reading fluency and text reading fluency, no significant interactions were observed, suggesting that for reading fluency, children with high/low prior knowledge did not benefit differently from DGBL or regular education.

Table 1

Effects of prior knowledge on the intervention effects

	number line estimation			**simple word reading fluency**			**text reading fluency**		
	df	**F**	**η_p^2**	**df**	**F**	**η_p^2**	**df**	**F**	**η_p^2**
prior knowledge	1	12.26**	.05	1	96.08**	.23	1	97.43**	.23
condition	1	0.34	.00	1	2.81	.01	1	0.45	.00
condition x prior knowledge	1	7.54**	.03	1	0.78	.00	1	0.00	.00
Error	213			322			321		

**p ≤ 0.01, *p ≤ 0.05

5.4.2 MATH AND READING ANXIETY

Results show no significant main effects of condition and domain-specific anxiety on number line estimation, word reading fluency or text reading fluency. Table 2 presents the results of the ANCOVAs about the impact of MA and RA pretest scores on the number line estimation, simple word reading fluency and text reading fluency posttest scores. No moderating effects were found for MA or RA, this means that high MA/RA children and low MA/RA children did not benefit differently from regular education or game-based training.

Table 2

Effects of skill-related anxiety on the intervention effects

	number line estimation			**simple word reading fluency**			**text reading fluency**		
	df	**F**	**η_p^2**	**df**	**F**	**η_p^2**	**df**	**F**	**η_p^2**
anxiety[4]	1	0.05	.00	1	1.67	.01	1	0.02	.00
condition	1	1.04	.01	1	.23	.00	1	0.43	.00
condition x anxiety	1	0.11	.00	1	0.03	.00	1	0.11	.00
Error	209			315			313		

**p < 0.01, *p < 0.05

5.4.3 SOCIO-ECONOMIC STATUS (SES)

The results yield a significant main effect of SES on number line estimation, but no significant effect on word reading fluency and text reading fluency. Condition has a significant main effect on number line estimation, but not on word reading fluency and text reading fluency. No interactions were observed for SES, indicating that the effect of a game/no game on number line estimation, simple word reading fluency and text reading fluency was not moderated by SES (see Table 3).

Table 3

Effects of SES on the intervention effects

	number line estimation			**simple word reading fluency**			**text reading fluency**		
	df	**F**	**η_p^2**	**df**	**F**	**η_p^2**	**df**	**F**	**η_p^2**
SES	5	6.75*	.87	5	2.41	.71	5	2.36	.70

[4] In case the dependent variable concerned a math outcome, the MA pretest scores were included in the model. If the dependent variable concerned a reading outcome, the RA pretest scores were included in the model.

condition	1	48.23**	.68	1	4.25	.19	1	1.52	.11
condition x SES	5	0.50	.01	5	0.70	.01	5	1.06	.02
Error	185			283			283		

**p < 0.01, *p < 0.05

6 Discussion

This study provides unique insights in the effectiveness of DGBL as a wide range of cognitive, non-cognitive and differential effects of DGBL were investigated in two important domains of formal education within the same experiment. More particularly, whether (1) DGBL impacts near and far transfer skills, (2) the training effects last for a longer period, (3) DGBL impacts non-cognitive factors (e.g., anxiety), and (4) the training effects are the same for all children. The two games, NSG and RG, trained domain-specific skills (i.e., early numerical skills and emergent reading) at the start of first grade. This first year of formal schooling is critical for laying foundational math and reading skills. Children in the experimental condition played NSG and RG, while children in the control condition were involved in regular non-gamified education, practicing similar content.

The first research question investigated whether near and far transfer were realized through DGBL. Results on *near transfer test* revealed mixed findings. The children of the experimental condition outperformed those of the control condition on number line estimation, but no differences between the two conditions were observed concerning digit comparison and letter knowledge immediately after the intervention. A possible explanation is the intensive training of digit comparison and letter knowledge during the first months of the first grade in regular education. By contrast, NLE tests are used less frequently in that period (cf. personal communication with teachers). The effects of digital games on *far transfer tests* (i.e., general math and reading achievement) also revealed mixed findings. With regard to math, results showed no differences between the experimental and control condition on calculation fluency. However, for general math competence, children from the control condition outperformed children from the experimental condition. This is possibly due to the slightly different content which was tested in the LVS (measured as a pretest and delayed test) and the CDR (measured immediately after the training) instruments, with the CDR consisting of more word problems. Taking into account that we did not use the CDR measure as a pretest, the significance of this effect is questionable and therefore, in light of the other results, not strong enough to derive robust conclusions about the effect of DGBL on far transfer. By contrast, with regard to reading competence, children who played the RG improved significantly more in simple word reading fluency and text reading fluency, compared to children who did not play the RG. For complex word reading fluency, no differences between the conditions were observed. The mixed results with respect to far transfer are in line with previous DGBL research which also reported mixed findings on far transfer tests for math and reading immediately after the intervention (e.g., Kartal & Terziyan, 2016; Kim et al., 2018; Obersteiner et al., 2013; van Gorp et al., 2017).

Delayed tests were administered eight weeks after the first posttest to address the second research question, i.e. whether training effects remain for a longer period. The results revealed mixed effects. No differences were observed between both conditions on the delayed math test. In contrast, simple word reading fluency was still better in the experimental condition compared to the control condition. The positive effect of the RG on text reading fluency disappeared in the delayed test. Previous studies with respect to DGBL often lacked a delayed evaluation (e.g., Kartal & Terziyan, 2016; Obersteiner et al., 2013) or reported mixed findings. For example, Räsänen et al. (2009) reported a sustained effect of digit comparison, while van der Ven et al. (2017) failed to find any delayed advantage for the game play condition. With respect to DGBL studies on reading, delayed effects of the intervention were measured more frequently. Several studies reported that some effects were even maintained a few weeks after the intervention (e.g., Kyle et al., 2013; Segers & Verhoeven, 2005; van de Ven et al., 2017).

For both immediate and delayed assessment concerning far transfer, the differences between the results on math and reading might be explained in at least two ways. First, the content that was trained for reading (i.e., emergent reading) was more closely related to word reading fluency and text reading fluency (i.e., the far transfer reading tests), than the content that was trained for math (i.e., numerical comparison and NLE) was to calculation fluency and general math competence (i.e., the far transfer math tests). However, counterevidence for this first explanation is the fact that numerical comparison is identified as a predictor of math competence (Schneider et al., 2017) and that our general math competence tests also comprised comparison tests. Second, it might be possible that the integrated instructional support in each game is also partly responsible for the more positive results on reading compared to math (Vusić et al., 2018). In the RG children were able to follow their own proficiency through the game and were rewarded with a star after each finished subgame, two game features that were not present in the NSG. In addition, in the RG children were allowed to choose which subgame they wanted to play within a particular difficulty level, while the NSG lacked learner control. Finally, letters and words in the RG were often displayed in combination with oral support (i.e., modality principle; see also Mayer, 2014), while in the NSG numbers and numerosity were only printed. These three characteristics or instructional support in the RG are in line with the literature suggesting that game features such as feedback, interactivity and modality are beneficial for learner outcomes (Vusić et al., 2018). However, further work needs to be done to determine under which circumstances far transfer will appear.

The third research question examined the impact of games on non-cognitive factors. Similar to previous studies, we examined the impact of DGBL on math and reading anxiety operationalized via questionnaires (Gunderson et al., 2018; Jansen et al., 2013). We observed that children from all conditions showed significantly less anxiety towards math at the moment of the posttest compared to prior to the intervention, while reading anxiety scores remained the same. However, the experimental condition did not differ in (math or reading) anxiety as a result of playing a math or reading game compared to the control group. This finding could indicate a true null effect of DGBL on math and reading anxiety or – alternatively – could be a consequence of the way in which anxiety was measured. Indeed, it should be acknowledged that – especially in the case of young children – non-cognitive factors such as anxiety are difficult to assess (Wouters et al., 2013). The assessment of anxiety *after* children have played the game implies that children remember their affective state during game play – which is of course difficult to evaluate. Therefore, the use of *continuous* (state) measures such as physiological data *during* game play might be more appropriate to assess this research question in future studies (Verkijika et al., 2015).

In the fourth research question, it was investigated whether the effects of the DGBL intervention were moderated by prior knowledge, math and reading anxiety and/or SES. For number line estimation, besides a main effect of condition, also an interaction of prior knowledge and condition was found on the number line estimation posttest scores. This means that a bigger difference was observed between the experimental and control condition for children with low prior knowledge compared to the difference between both conditions for children with high prior knowledge. This might be due to the fact that children with more prior knowledge also without explicit instruction for number line estimation in class (control condition) might be able to solve these tests, while for children with lower prior knowledge an intensive training with these tests in the game condition is beneficial. Other than that, no interactions were found. This shows that children's individual differences (such as math and reading anxiety and SES) might play a rather small role when evaluating the cognitive effects of a tablet game intervention. These results are in line with the findings of Kebritchi et al. (2010) who did not find a moderation of effectivity by prior knowledge, computer skills and English language skills either.

A first limitation of this study is that the test battery we used did not lead to a one-to-one correspondence of each test with each time point (pre-post-delayed). For practical reasons of not over-assessing children only far transfer was evaluated in the delayed test. For future research, it would be interesting to also measure near transfer delayed as one can expect a larger impact of a game on near transfer compared to far transfer (e.g., Kyle et al., 2013). With regard to the far transfer tests, we made use of existing standardized tests to assess the cognitive and non-cognitive outcomes as much as possible. Moreover, it was chosen to conduct tests which would have been administered by the teacher anyway, irrespective of the study. The advantage of using these tests is that children are not over-assessed as well as these tests are reliable and they measure far transfer. Disadvantages are that they need to be administered on specific moments (according to what children mastered at that moment). However, we considered that the advantages of the current test battery (using standardized tests) were crucial to answer the research questions and do not outweigh the drawbacks. Second, as already indicated, it would be more reliable to adopt a multimodal approach to measure non-cognitive factors (e.g., semi-structured interviews, continuous physiological data during game play). Third, the current study takes a medium-comparison perspective in which a game condition is compared to a non-game condition. The mixed findings suggest that it is not the medium as such – i.e., a game – impacts learning outcomes, but rather the way in which the game is designed. By contrast, adopting a value-added approach in which the learning gains of children, enrolled in different versions of one educational game, are compared might be better suited for future studies to understand which specific game features foster learning. As suggested above, game features such as feedback, interactivity and multimodality might be especially beneficial in explaining learners' learning gains.

The current findings have some implications for teachers, parents and game designers. We would encourage teachers and parents to be open-minded towards DGBL as it is at least an equivalent or even better alternative to practice early math and reading. Meanwhile, they should also be critical towards the quality of the games (i.e., Which instructional support is provided?) and need to be aware that these games are no magic bullets for learning e.g. early math and reading skills, certainly given the mixed findings. Designers of educational games are recommended to discuss the development of the games closely with educational scientists and teachers to develop games which provide sufficient instructional support to promote learning and which are easy to implement in the classroom. Based on our study, it might be assumed that games including feedback, carefully following a child's own proficiency, providing exercises which are not only printed but also orally explained, and including rewards and choice might be more beneficial for children's enjoyment and learning compared to games which do not provide these game features. However, further research is needed to further unravel the game features that are crucial for learning. Furthermore, based on our experiences in the classroom, it might be suggested that classroom implementation could be enhanced if young children could login through finger or face recognition, instead of the typed logins used in the current study. This way, teachers could save a lot of time and all children would be able to start practicing without teacher's help.

In sum, despite some mixed findings, we can conclude that the present study provides evidence for the effectiveness of DGBL. Indeed, according to the recently developed comprehensive framework of All, Castellar and Van Looy (2015) a game is considered to be effective if it achieves similar or higher scores on cognitive and non-cognitive learning outcomes in comparison to 'business as usual' in cognitive and non-cognitive outcomes. Except for one test, our results of the training on near and far transfer tests, on immediate and delayed tests, showed equal or higher scores in the DGBL condition compared to the control condition immediately after the training and delayed.

7 References

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

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

Brankaer, C., Ghesquière, P., & De Smedt, B. (2017). Symbolic magnitude processing in elementary school children: A group administered paper-and-pencil measure (SYMP Test). *Behavior Research Methods*, *49*(4), 1361–1373.

Chung, K. K. (2015). Socioeconomic status and academic achievement. *International Encyclopedia of the Social & Behavioral Sciences*, *22*(2), 924-930.

Chung, K. K., Liu, H., McBride, C., Wong, A. M. Y., & Lo, J. C. (2017). How socioeconomic status, executive functioning and verbal interactions contribute to early academic achievement in Chinese children. *Educational Psychology, 37*(4), 402-420.

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.

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

De Smedt, B., Noël, M. P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education, 2*(2), 48-55.

De Vos, T. (1992). *Test voor het vaststellen van het rekenvaardigheidsniveau der elementaire bewerkingen (automatisering) voor het basis en voortgezet onderwijs: Handleiding* [Test to determine the mathematics ability level for elementary operations (automatization) in primary and secondary education: Manual]. Nijmegen, The Netherlands : Berkhout.

Desoete, A., & Roeyers, H. (2006). *Cognitieve Deelvaardigheden Rekenen (CDR). Rekentests voor 1ste, 2de en 3de graad.* Herenthals, Belgium: Vlaamse Vereniging voor Logopedisten (VVL).

Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology*, *7*(4), 1–16.

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.

Gunderson, E. A., Park, D., Maloney, E. A., Beilock, S. L., & Levine, S. C. (2018). Reciprocal relations among motivational frameworks, math anxiety, and math achievement in early elementary school. *Journal of Cognition and Development*, *19*(1), 21–46.

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

Kadosh, Roi, C., Dowker, A., Heine, A., Kaufmann, L., & Kucian, K. (2013). Interventions for improving numerical skills: present and future. *Trends in Neuroscience and Education*, *2*(2), 85–93.

Kartal, G., & Terziyan, T. (2016). Development and evaluation of game-like phonological awareness software for kindergarteners: JerenAli. *Journal of Educational Computing Research*, *53*(4), 519–539.

Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers and Education*, *55*(2), 427–443.

Kim, N., Jang, S., & Cho, S. (2018). Testing the efficacy of training basic numerical cognition and transfer effects to improvement in children's math ability. *Frontiers in Psychology*, *9*, 1775.

Krom, R., Jongen, I., Verhelst, N., Kamphuis, F., & Kleintjes, F. (2010). Wetenschappelijke verantwoording DMT en AVI [scientific justification of the three-minutes-test and AVI reading test]. Arnhem, The Netherlands: Cito.

Kyle, F., Kujala, J., Richardson, U., Lyytinen, H., & Goswami, U. (2013). Assessing the effectiveness of two theoretically motivated computer assisted reading interventions in the United Kingdom: GG Rime and GG Phoneme. *Reading Research Quarterly*, *48*(1), 61–76.

Linsen, S., Maertens, B., Husson, J., Van den Audenaeren, L., Wauters, J., Reynvoet, B., et al. (2015). Design of the game "Dudeman & Sidegirl: Operation clean world", a numerical magnitude processing training. In J. Torbeyns, E. Lehtinen, & J. Elen (Eds*.), Describing and studying domain-specific serious games* (pp. 9-26). Springer International Publishing.

Maertens, B., De Smedt, B., Sasanguie, D., Elen, J., & Reynvoet, B. (2016). Enhancing arithmetic in pre-schoolers with comparison or number line estimation training: Does it matter?. *Learning and Instruction*, *46*, 1-11.

Mayer, R. E. (2014). Multimedia instruction. In J. M. Spector, M. D. Merrill, J. Elen, M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (4th ed.) (pp. 385-399). New York, NY: Springer.

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.

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.

Obersteiner, A., Reiss, K., & Ufer, S. (2013). How training on exact or approximate mental representations of number can enhance first-grade students' basic number processing and arithmetic skills. *Learning and Instruction*, *23*(1), 125–135.

Punaro, L., & Reeve, R. (2012). Relationships between 9-year-olds' math and literacy worries and academic skills. *Child Development Research*, 1-12.

Ramirez, G., Chang, H., Maloney, E. A., Levine, S. C., & Beilock, S. L. (2016). On the relationship between math anxiety and math achievement in early elementary school: The role of problem solving strategies. *Journal of Experimental Child Psychology*, *141*, 83–100.

Ramirez, G., Fries, L., Gunderson, E., Schaeffer, M. W., Maloney, E. A., Beilock, S. L., & Levine, S. C. (2019). Reading Anxiety: An Early Affective Impediment to Children's Success in Reading. *Journal of Cognition and Development, 20*(1), 15–34.

Räsänen, P., Salminen, J., Wilson, A. J., Aunio, P., & Dehaene, S. (2009). Computer-assisted intervention for children with low numeracy skills. *Cognitive Development*, *24*(4), 450–472.

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.

Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., & De Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: a meta-analysis. *Developmental Science*, *20*(3).

Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: A meta-analysis. *Child Development*, *89*(5), 1467–1484.

Segers, E., & Verhoeven, L. (2005). Long-term effects of computer training of phonological awareness in kindergarten. *Journal of Computer Assisted Learning*, *21*(1), 17–27.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development*, *75*(2), 428–444.

Sun, Y., & Pyzdrowski, L. (2009). Using technology as a tool to reduce mathematics anxiety. *The Journal of Human Resource and Adult Learning*, *5*(2), 38–44.

van de Ven, M., de Leeuw, L., van Weerdenburg, M., & Steenbeek-Planting, E. G. (2017). Early reading intervention by means of a multicomponent reading game. *Journal of Computer Assisted Learning*, *33*(4), 320–333.

van der Ven, F., Segers, E., Takashima, A., & Verhoeven, L. (2017). Effects of a tablet game intervention on simple addition and subtraction fluency in first graders. *Computers in Human Behavior*, *72*, 200–207.

van Gorp, K., Segers, E., & Verhoeven, L. (2017). Enhancing decoding efficiency in poor readers via a word identification game. *Reading Research Quarterly*, *52*(1), 105–123.

Van Rompaey, A. & Vandenberghe, I. (2015). *Leerlingvolgsysteem LVS-VCLB Wiskunde 1-2*. Antwerpen: Garant.

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.

Verkijika, S. F., & De Wet, L. (2015). Using a brain-computer interface (BCI) in reducing math anxiety: Evidence from South Africa. *Computers and Education*, *81*, 113–122.

Vusić, D., Bernik, A., & Gečeк, R. (2018). Instructional design in game based learning and applications used in educational systems. *Technical Journal*, *1*(2), 11–17.

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 skills and word-level reading as children develop from beginning to skilled readers: a 5-year longitudinal study. *Developmental Psychology*, *33*(3), 468–479.

Wanzek, J., Stevens, E. A., Williams, K. J., Scammacca, N., Vaughn, S., & Sargent, K. (2018). Current evidence on the effects of intensive early reading interventions. *Journal of Learning Disabilities*, *51*(6), 612–624.

Wilson, A. J., Dehaene, S., Dubois, O., & Fayol, M. (2009). Effects of an adaptive game intervention on accessing number sense in low-socioeconomic-status kindergarten children. *Mind, Brain, and Eeducation*, *3*(4), 224–234.

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.

Yang, J. C., Lin, M. Y. D., & Chen, S. Y. (2018). Effects of anxiety levels on learning performance and gaming performance in digital game-based learning. *Journal of Computer Assisted Learning*, *34*, 324–334.

Yang, J. C., & Quadir, B. (2018). Effects of prior knowledge on learning performance and anxiety in an English learning online role-playing game. *Journal of Educational Technology & Society, 21*(3), 174-185.

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., … Tran, M. (2012). Our princess is in another castle : A review of trends in serious gaming for education. *Review of Educational Research*, *82*(1), 61–89.

Zyda, M. (2005). From visual simulation to virtual reality to games. *IEEE Computer Society, 38*(9), 25–32.

8 Appendices

A. Descriptives of the pretests, posttests and delayed tests

Table A1 shows the descriptives of all instruments measuring near and far transfer cognitive factors of two conditions at pretest, posttest and delayed test. Moreover the descriptive results on the MA and RA questionnaire are also reported.

Table A1

Descriptives for near transfer tests, the far transfer tests and the MA and RA questionnaire

instrument	**condition**	**pretest**			**posttest**			**delayed test**		
		N	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
near transfer tests										
digit comparison (# correct, max. = 60)	experimental	109	12.73	4.17	109	17.18	4,22	[a]	[a]	[a]
	control	111	12.41	4.08	108	16.85	4,35	[a]	[a]	[a]
motor speed (# correct, max. = 60)	experimental	109	11.78	5.13	109	16.85	4,51	[a]	[a]	[a]
	control	110	11.48	3.66	111	16.77	4,76	[a]	[a]	[a]
symbolic NLE (*PAE*)	experimental	109	26.09	9.63	109	12.27	7,72	[a]	[a]	[a]
	control	110	24.09	9.07	111	22.08	9,15	[a]	[a]	[a]
phonemic awareness (# correct, max. = 24)	experimental	220	17.54	3.78	[a]	[a]	[a]	[a]	[a]	[a]
	control	112	17.64	3.72	[a]	[a]	[a]	[a]	[a]	[a]
letter knowledge (# correct, max. = 21)	experimental	215	11.75	4.68	220	19.22	2.12	[a]	[a]	[a]
	control	113	11.67	4.20	111	18.48	2.39	[a]	[a]	[a]
far transfer tests										
general math competence (CDR) (# correct, max. = 25)	experimental	[a]	[a]	[a]	108	8.40	4.49	[a]	[a]	[a]
	control	[a]	[a]	[a]	111	9.97	5.21	[a]	[a]	[a]
calculation fluency (TTR) (# correct, max. = 80)	experimental	[a]	[a]	[a]	107	14.07	6.31	[a]	[a]	[a]
	control	[a]	[a]	[a]	109	15.03	7.29	[a]	[a]	[a]
general math competence (LVS) (# correct, pre: max. = 43, retention: max. = 91)	experimental	107	32.40	6.50	[a]	[a]	[a]	98	74.91	12.38
	control	110	32.36	6.86	[a]	[a]	[a]	104	73.74	13.21
simple word reading fluency (DMT_card1)	experimental	[a]	[a]	[a]	221	12.36	8.04	221	24.29	14.56

(# correct, max. = 150 words)										
	control	[a]	[a]	[a]	112	10.34	7.87	109	19.33	12.83
complex word reading fluency (DMT_card2) (# correct, max. = 150 words)	experimental	[a]	[a]	[a]	221	5.79	5.17	221	12.56	8.68
	control	[a]	[a]	[a]	112	6.04	6.19	109	11.47	8.78
text reading fluency (AVI) (time read)	experimental	[a]	[a]	[a]	220	0:04:00	0:01:10	219	0:02:59	0:01:20
	control	[a]	[a]	[a]	111	0:04:16	0:01:00	104	0:03:14	0:01:20
questionnaires										
math anxiety (total score, 8 items, 4-point likert scale)	experimental	107	26.98	5.18	107	28.34	4.08	[a]	[a]	[a]
	control	101	28.01	4.30	101	28.79	3.82	[a]	[a]	[a]
reading anxiety (total score, 8 items, 4-point likert scale)	experimental	212	26.63	5.19	212	26.64	4.89	[a]	[a]	[a]
	control	102	27.27	5.15	102	26.75	5.13	[a]	[a]	[a]

[a] The grey boxes indicate which instruments were not administered as a pretest, posttest or delayed test; *PAE* = percentage absolute error

B. Correlation matrix

The correlations between the pretest measures are presented in table B1.

Table B1

Correlations between the pretest measures

	1	2	3	4	5	6	7	8	9
1 motor speed test	-								
2 digit comparison test	.53**	-							
3 symbolic *NLE* test	-.08	-.13*	-						
4 student monitoring system for arithmetic performance	.29**	.49**	-.14**	-					
5 letter knowledge test	.25**	.41**	-.05	.42**	-				
6 phonological awareness	.19**	.37**	-.02	.44**	.47**	-			
7 math anxiety	.09	.11*	.01	.13*	.02	-.02	-		
8 reading anxiety	.02	.02	.00	-.02	-0.07	-.01	.67**	-	
9 degree mother	.16**	.27**	-.06	.30**	.09	.12*	.08	-.02	-

*p<.05, **p<.01