Effects of a Mathematics Game-Based Learning Environment on Primary School Students' Adaptive Number Knowledge

Boglárka Brezovszky[1], Jake McMullen[1], Koen Veermans[1], Minna M. Hannula-Sormunen[1,2], Gabriela Rodríguez-Aflecht [1], Nonmanut Pongsakdi[1] Eero Laakkonen[1] & Erno Lehtinen[1]

University of Turku

Author Note

[1]Centre for Research on Learning and Instruction and Department of Teacher Education, University of Turku, Finland

[2] Department of Teacher Education and Turku Institute for Advanced Studies, University of Turku, Finland

Published in Computers & Education 128 (2019) 63-74

https://doi.org/10.1016/j.compedu.2018.09.011

Corresponding Author:

Boglárka Brezovszky

Centre for Research on Learning and Instruction, Department of Teacher Education

Turun Yliopisto, 20014, Finland.

E-mail: bogbre@utu.fi

Funding

This work was supported by the Academy of Finland [grant number 274163; PI: Erno Lehtinen] and the Academy of Finland DECIN project [grant number 278579; PI: Minna Hannula-Sormunen].

1. Introduction

Flexible mathematical thinking and the adaptive use of arithmetic strategies have been highlighted by researchers of mathematics education (Baroody, 2003; Nunes, Dorneles, Lin, & Rathgeb-Schnierer, 2016) and flexibility is an important aim of national mathematics education standards and curricula in many countries (e.g., Finnish National Agency for Education, 2014; NCTM 2014). Flexibility is necessary to apply problem solving procedures in new contexts and across different representations, which is crucial for a transferable mathematical proficiency to everyday life (NCTM 2014; Nunes et al., 2016). However, pedagogical recommendations and models for enhancing flexibility and adaptivity are rare and mostly focused on teaching the use of a few strategies (Verschaffel, Luwel, Torbeyns, & Van Dooren, 2009), which are rarely applied in practice in spite of a strong curricular emphasis (Hickendorff, 2017). One explanation for this disconnect could be that, in order to flexibly apply arithmetic strategies in varying situations, it is not enough to teach these strategies, but students need to develop a well-connected mental representation of the natural number system which makes it possible for them to notice when and which strategies are applicable (Baroody, 2003; Brezovszky et al., 2015; Lehtinen, et al., 2015; Threlfall, 2002, 2009; Verschaffel et al., 2009).

Indeed, regular classroom instruction may struggle to provide students with the opportunity to experience the rich relations of numbers and operations (Baroody, 2003; Threlfall, 2009) and the type of practice needed for the development of flexible and adaptive mathematical skills (Lehtinen, Hannula-Sormunen, McMullen, & Gruber, 2017). Thus, it is important to study if game-based learning could be used in enhancing this highly emphasized, but pedagogically under-supported, aim of mathematics education. The open-ended and flexible nature of game-based learning environments can be ideal to support this intensive practice in an engaging way (Devlin, 2011; Ke, 2009; Squire, 2008), and the technological affordances allow the training to be scalable (Rutherford et al., 2014). These features of the game-based learning environment could provide teachers with an ideal complementary tool for enriching their methods in developing flexibility and adaptivity with arithmetic problem solving. The present study explores the potential of the Number Navigation Game (NNG) in strengthening fourth, fifth, and sixth grade students' flexible and adaptive arithmetic skills. Gameplay with the NNG is expected to develop a more well connected representation of the arithmetic relations between natural numbers in students, which should enable them to become more adaptive in their arithmetic problem solving.

1.1. Developing flexibility and adaptivity with arithmetic problem solving by strengthening students' adaptive number knowledge

The definition of flexibility varies greatly, but in general it is described as: (a) having a rich repertoire of arithmetic problem solving procedures, (b) being able to flexibly switch between these procedures, and (c) being adaptive with selecting strategies efficiently (Heinze, Star, & Verschaffel, 2009; Star & Rittle-Johnson, 2008; Verschaffel et al., 2009). It is problematic however that the meaning of flexible and adaptive can be very subjective, as it might depend on the characteristics of the problem, on personal preferences, or the context (Verschaffel et al., 2009), and efficient solutions can sometimes be described as elegant and sometimes as quick or fast (Heinze et al., 2009). With so much variability in what it means to be flexible and adaptive with arithmetic problem solving it is equally unclear what type of instructional design can support the development of this knowledge and skills.

It is suggested that instead of focusing on building a strategy repertoire, more emphasis should be given to supporting students in developing an understanding of the qualities of numbers and numerical relations (Threlfall, 2009). Studies show that both mathematically proficient primary school students (Blöte, Klein, & Beishuizen, 2000; Blöte, Van der Burg, & Klein, 2001) and expert mathematicians (Star & Newton, 2009) are likely to take into account the characteristics of numbers in a problem when deciding on solution strategies. This approach emphasizes the subjective nature of flexibility and shifts the focus towards the process of problem solving during which efficient strategies emerge and are applied adaptively as important features of a problem become visible (Baroody, 2003; Schneider, Rittle-Johnson, & Star, 2011; Threlfall, 2002, 2009; Verschaffel et al., 2009). From a more practical perspective, it is suggested that for students to recognize important features of a problem they need to develop a well-connected understanding of numerical relations for which they need opportunities to practice with various number combinations and numerical relations making their own discoveries (Baroody, 2003; Verschaffel et al., 2009).

The present study tested the efficiency of a game-based learning environment for strengthening these numerical relations defined as *adaptive number knowledge*. Adaptive number knowledge is a component of adaptivity with whole-number arithmetic and it refers to the well-connected, network-like knowledge of numerical characteristics and arithmetic relations between numbers (McMullen et al., 2016, 2017). Adaptive number knowledge enables the recognition and use of key numbers in adaptive arithmetic problem solving. This can include recognizing numbers with many factors or multiples, recognizing numbers in equations that are close to other numbers that are easy to work with, or the use of the base-ten structure and basic arithmetic principles like associativity or commutativity (McMullen et al., 2017). Adaptive number knowledge is related to arithmetic fluency and knowledge of arithmetic concepts and is a unique predictor of pre-algebra skills (McMullen et al., 2017). This suggests that having a well-connected representation of numerical relations is not only related to faster and more efficient problem solving, but also to a more complex understanding of arithmetic relations, which can be foundational to algebraic thinking (McMullen et al., 2017; Nunes et al., 2016).

Designing tools for developing adaptive number knowledge in the context of regular classroom teaching can be challenging. In order for students to develop their own flexible procedures they should be provided opportunities for extensive deliberate practice with various combinations of numbers and operations (Baroody, 2003; Lehtinen et al., 2017; Verschaffel et al., 2009). This requires extra resources from the teacher in developing and using individualized training methods. Indeed, the few existing intervention studies which are not limited to teaching a few strategies but aim at supporting students' discovery of flexible and adaptive problem solving procedures are case studies or have a highly controlled experimental design using small samples. For example, case studies by Heirdsfield and collegues (Heirdsfield & Cooper, 2004; Heirdsfield, 2011) suggest that using external representations like the number line or the number square can be useful in promoting students' engagement in discovering their own strategies when performing mental arithmetic. Additionally, there is evidence that reflecting on ones' solutions, comparing and contrasting different solution methods or re-solving the same problem using an alternative procedure can promote students' flexibility both in mental arithmetic and in equation solving (Blöte et al., 2000; Rittle-Johnson & Star, 2009; Star & Seifert, 2006).

The present study tested the effectiveness of the NNG game-based learning environment in strengthening primary school students' adaptive number knowledge and related arithmetic skills. Using a game-based format allows extensive opportunities for practice in an engaging open-ended context that does not pre-define 'optimal' solutions but prompts students to reflect on the efficiency of various alternatives. Through careful design, discovery learning and exploration are organic components of game-based learning (Ke, 2009; Squire, 2008) and including appropriate rules and challenges can promote reflection (Ke, 2008; Kiili, 2007). Additionally, due to technological affordances, the training is scalable and can take into account individual differences (Rutherford et al., 2014). This format made it possible to apply the training in naturalistic classroom settings on a large scale, which was not possible in previous interventions with similar aims. It is expected that the NNG would provide teachers with an ideal complementary tool for enriching their methods for developing flexibility and adaptivity with arithmetic problem solving and provide a novel training that can be flexibly used over three different grade levels

1.2. Developing adaptive number knowledge using game-based learning

Digital games can be powerful in presenting complex mathematical concepts as they can provide an alternative media where interaction with and exploration of the content is inherent (Devlin, 2011; Ke, 2009; Lowrie & Jorgensen, 2015). However, it is clear now that games as such do not represent a magic bullet and they do not automatically make the learning experience more motivating or effective (Wouters & Oostendorp, 2013; Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013; Young et al., 2012). As with most media, it is a question of a meaningful translation and integration of the learning content into the game context which makes game-based learning effective.

Thus, the added value of game-based learning depends on the design process where the learning content integrated and not just added on top of the game-based media. This allows students to interact with the content in a novel and alternative way that would otherwise be unavailable or difficult to achieve in the traditional classroom practice (Arnab et al., 2015; Devlin, 2011; Habgood & Ainsworth, 2011).

In spite of these affordances of game-based learning, many interventions aim to simply develop basic arithmetic skills or to strengthen already acquired skills (Cheung & Slavin, 2013; Li & Ma, 2010). Empirically tested game-based learning environments which target the development of complex mathematical skills and knowledge in primary school mathematics education are rare. Additionally, some exiting examples can be limited by the domain, for example training multiplicative reasoning (Bakker, van den Heuvel-Panhuizen, & Robitzsch, 2015) or the understanding of multiplicative relations (Habgood & Ainsworth, 2011) and only few game-based learning environments aim for more holistic mathematical learning outcomes such as flexibility and adaptivity with arithmetic problem solving. Such examples include the game-based intervention by Pope and Mangram (2015) which aimed at training students' flexibility with numbers and operations and an understanding of properties of numbers, or the large-scale intervention by van den Heuvel-Panhuizen, Kolovou, and Robitzsch (2013) which aimed to promote the development of reasoning about relations between quantities that are foundational to early algebra knowledge.

Although the learning aims of the NNG are quite close to the ones described by Pope and Mangram (2015) and van den Heuvel-Panhuizen and collegues (2013), that is the flexible use and recognition of numerical relations in arithmetic problem solving, the format of the NNG allows for a different experience with number combinations and numerical relations. Additionally, unlike previous game-based learning environments (e.g. Bakker et al., 2015; Habgood & Ainsworth, 2011) the NNG aims to develop the understanding of numerical relations, without limiting the scope of the game to one type of arithmetic operation (e.g. multiplicative or additive relations). Instead, the aim of the game is to develop the recognition and use of various numerical relations within the system of natural numbers as a whole. To our knowedge, the NNG represents the first attempt to develop elemantary school students' adaptive number knolwedge by strengthening their understanding of numerical relations.

Based on empirical results and theoretical suggestions for developing adaptive number knowledge, the game design should be able to (a) maintain students' long-term engagement to mentally execute and compare a large number of various equations in a meaningful way (Baroody, 2003; Threlfall, 2009; Verschaffel et al., 2009), (b) be flexible enough to promote students' curiosity in searching for alternative solutions (Rodríguez-Aflecht et al., 2018), (c) provide students with an external representation that supports the formation of rich networks of numbers and operations (Mulligan & Mitchelmore, 2013), and (d) promote reflection on the quality and efficiency of their solution process (Heirdsfield, 2011; Rittle-Johnson & Star, 2009; Star & Seifert, 2006). As the learning content is integrated within the core game mechanics (Devlin, 2011; Habgood & Ainsworth, 2011; Salen & Zimmerman, 2003), progress in the NNG means performing mental calculations and combining and comparing various alternative number-operation combinations. The hundred-square was selected as an external representation as it can be flexibly used to work within the system of natural numbers, decomposing and recombining numbers, and provides a clear visual representation of numerical relations and the base-ten system (Laski, Ermakova, & Vasilyeva, 2014).

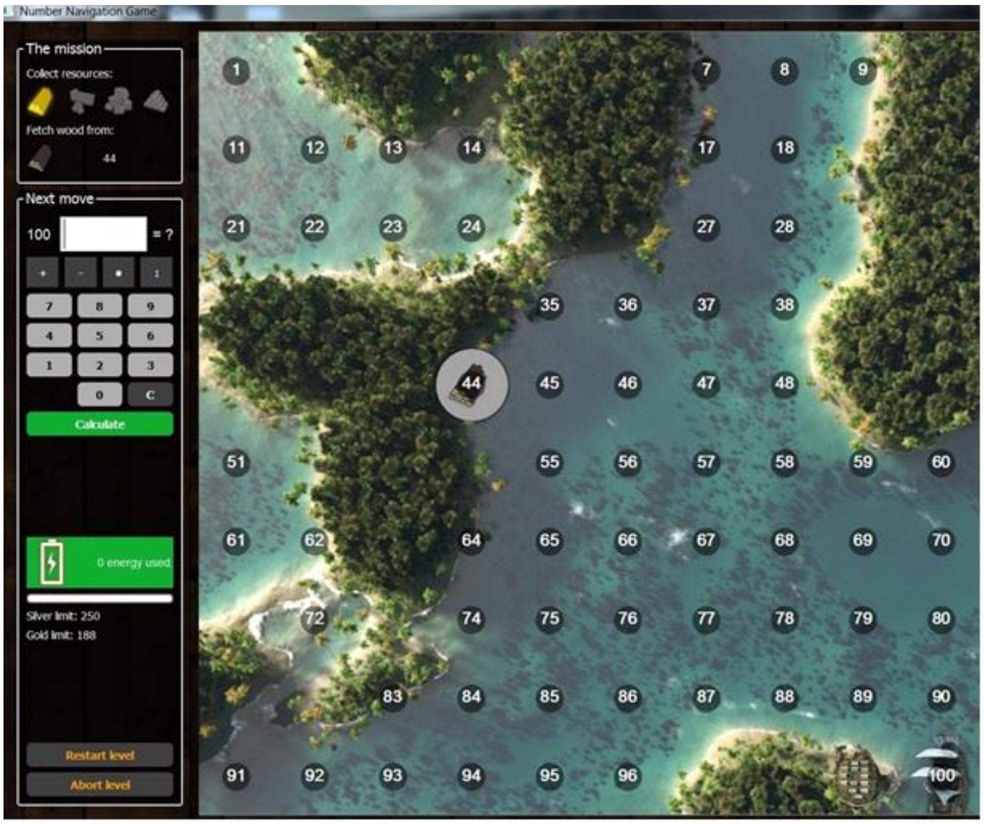

Figure 1. Example of a NNG map in the energy scoring mode (harbour at number 100, first target material at number 44).

Within the game, different game modes were developed to engage players in searching for alternative solutions and reflect on the efficiency of these different solutions. For example, on Figure 1 the game mode requires the player to minimize the sum of numbers used in their equations, so the player has to look for key numbers which are useful to achieve this goal. One useful solution can be to realize that both 90 and 88 are ideal to move forward using division with small numbers. The game is open-ended, therefore there are no right or wrong answers but players need to select their calculations considering the game rules and the position of islands (which define the available numbers in a map) or the placement of starting and target numbers. Taking into account these challenges, players need to mentally execute and compare several alternative routes when selecting their calculations. It is this type of mental practice with various combinations of numbers and operations which represents the core game action that is expected to develop students' adaptive number knowledge by strengthening the recognition and use of numerical characteristics and relations during arithmetic problem solving (Brezovszky, Lehtinen, McMullen, Rodriguez, & Veermans, 2013; Brezovszky et al., 2015; McMullen et al., 2016, 2017).

Results of a pilot testing of the NNG showed that players executed around 180 calculations in an hour of gameplay and used a large variety of number combinations (Brezovszky et al., 2013). Players were likely to compare and contrast their solutions prior to execution and reflect on their solution during gameplay. Additionally, results of a pilot intervention using a prototype of the NNG showed an increase in sixth grade students' adaptive number knowledge and math fluency after a seven-week long training (Brezovszky et al., 2015). The present study scales up this training and explores the effectiveness of the NNG in the development of adaptive number knowledge, math fluency, and pre-algebra knowledge of primary school students in grades four to six.

2. Current study

The aim of the present study was to explore the effects of a 10-week long training of regular mathematics teaching enriched by the NNG game-based learning environment on the development of primary school students' adaptive number knowledge, arithmetic fluency, and pre-algebra knowledge. The effectiveness of the game-based learning environment was examined in an ecologically valid setting where, in line with the Finnish educational context, teachers had the freedom to decide about the practicalities of the gameplay. Furthermore, the study explored the relationship between game performance and mathematical learning outcomes. Accordingly, the present study asked the following research questions

2.1. How does training with the NNG affect the development of primary school students' adaptive number knowledge, arithmetic fluency, and pre-algebra knowledge in different grade levels?

An important aim of primary school mathematics education is to enhance the creative and flexible use of mathematical knowledge by teaching alternative methods of problem solving (Nunes et al., 2016; NCTM 2014). However, recent research shows that solely teaching strategies is insufficient (Hickendorff, 2017). For students to develop more flexibility and adaptivity with arithmetic problem solving they need to have a chance to practice with various number-operation combinations (Baroody, 2003; Verschaffel et al., 2009). Research has shown advantageous numerical connections can be noticed by those mathematical experts who have a dense and strong network of numerical relations (Dowker, 1992). A well-connected representation of the natural number system enables noticing of strategically important numbers and leads to a more frequent use of flexible strategies (Heirdsfield & Cooper, 2004), which has been shown to be related to better general math abilities

(Star & Seifert, 2006). As well, adaptive number knowledge has been shown to be related to arithmetic skills and knowledge (McMullen et al., 2016), and later pre-algebra knowledge (McMullen et al., 2017).

In the NNG, students are encouraged to explore combinations of numbers and operations, look for key numbers, and discover number patterns. Thus, it is expected that as a result of this repeated practice students will develop their recognition and use of different numerical characteristics and arithmetic relations, as indexed by their adaptive number knowledge, and also develop their arithmetic fluency and pre-algebra knowledge. The NNG has a complex design with different game features aimed at triggering different types of mathematical thinking. Thus, it can be anticipated that students in the experimental classes, where regular teaching was enriched with the NNG play, outperform students in the control classes but that the effects will be varying in different grade levels.

2.2. Does students' performance on the NNG affect the development of mathematical learning outcomes?

Studies in the domain of education and game-based learning are often criticized for failing to align game goals and learning goals (Devlin, 2011; Young et al., 2012). In the NNG the educational content is integrated within the core game mechanics (Habgood & Ainsworth, 2011; Lehtinen et al., 2015). Thus, gaming is not delivered as a reward after students have engaged with the necessary mathematical content, but players make meaningful progress interacting with the educational content as they combine, compare, and strategically select the different calculations. Accordingly, it is expected that the more practice players have with the mathematically relevant content of the NNG the better their performance will be on the mathematical outcome measures.

3. Methods

3.1. Participants

Participants were 1,168 fourth to sixth grade primary school students (546 female) from four urban and suburban areas in the southwest and middle of Finland. The mean age of the fourth graders was 10.18, (*SD* = 0.42), M_{age} of fifth graders was 11.14, (*SD* = 0.38), and M_{age} of sixth graders was 12.20, (*SD* = 0.45). Table 1 provides a description of the distribution of participants by grade in the experimental and control groups.

Table 1
Number of Participants by Grade and Experimental Condition

Group	Grade			Total
	4	5	6	
Experimental	63	309	270	642
Control	72	297	157	526
Total	135	606	427	1,168

Participation was voluntary; informed consents from parents and assent of the students were obtained before data gathering. All students in the experimental classes played the NNG as part of their mathematical teaching and completed the pre- and post-tests as part of their regular school work. However, data was only gathered from students who had the consent to participate in the study. Ethical guidelines of the University of Turku were followed. All teachers in both the experimental and control classes were qualified primary school teachers with a masters' degree.

3.2. Procedure

The study was a large-scale cluster randomized trial. Randomization to experimental and control conditions was done on the classroom level because classrooms are usually considered as ecologically valid units of measurement in experimental designs in the field of education (Hedges & Rhoads, 2010; Winn, 2003). The two-level hierarchical design with covariates (pre-test) was used in the power analysis and design planning (Hedges & Rhoads, 2010). The parameters used in power analysis were effect size (d = .35), power (> .8) correlations within ($Rw^2 = 0.5$), and among ($Rs^2 = 0.8$) clusters and intraclass correlations (ICC = 0.1). The correlation estimates were the ones presented by Hedges and Rhoads (2010) and the intraclass correlation used in the calculation was based on the values presented in a comprehensive review of differences between schools in Nordic countries by Yang Hansen, Gustafsson and Rosén (2014). Based on the power analysis, the minimum number of classrooms needed were 12 for experimental and 12 for control condition. In order to make sub-group comparison possible and better control the variation in ways how teachers used the NNG in their classrooms, a substantially larger sample was used in the present study. There were in total 61 teachers that volunteered to participate and were randomly assigned into experimental (31 classes) or control groups (30 classes). Two control classes were not able to participate because they had to move to a temporary school building due to renovations.

The intervention started at the beginning of the spring semester when all participants completed the pre-test measures. After the pre-test, the experimental group participated in the intervention over a ten week period in which they used the NNG as part of their regular math classes. The game was distributed to the experimental group on individual pen-drives and was played on PCs. Gameplay was expected to be integrated into everyday math teaching, thus the intervention group did not receive more training compared to the controls. For ethical reasons and in order to keep volunteer teachers' motivation during the implementation, after the post-test, the conditions were switched, and the control group played the NNG for the rest of the semester (5 weeks). There was no pseudo-treatment in the control classes. Control classes were instructed to continue their regular textbook-based teaching. In all schools, in line with the national core curriculum (Finnish National Agency for Education 2014), the local curriculum emphasized the creative use of alternative arithmetic problem solving strategies.

Before data gathering, all teachers from both conditions were invited to participate in an information session regarding the main features of the NNG and the general outline and aim of the study. All teachers received written instructions including a link to a video guide on how to use the game and were offered e-mail support in case of questions or technical difficulties. Teachers only received general guidelines regarding the gaming sessions, since (a) the general design of the NNG is open and flexible and was aimed to be suitable for various grade levels and (b) the present study aimed to create an ecologically valid intervention in the Finnish context. Teachers were free to choose if they wanted their students to play the game in pairs or individually, and in case of pair play, it was the teacher's decision how to select the pairs. For all classes playing in pairs teachers sent back the list of pairs when the game log data was gathered during the post-test[1].

The guidelines for teachers in the experimental group were to aim for around ten hours playing time, with at least three playing sessions a week where a session is no shorter than 30 minutes. Implementation fidelity was checked using the game log data. Based on the log data the average time on task (effective gameplay) in the 29 experimental classes was 4 hours and 10 minutes, ranging between average 3 hours and average 5 hours 30 minutes. There were two classes with average time on task less than one hour. Because the attempt to have an ecologically valid design, these differences in time on task were interpreted as a natural variation when this type of method is introduced in the regular education. Therefore, all classes were included in the comparison of the experimental and control groups.

3.3. Measures

Paper-and-pencil measures of adaptive number knowledge, arithmetic fluency, and pre-algebra knowledge were administered by thirteen different trained testers (university master's students or researchers). Before data gathering, all testers took part in a training regarding test administration. The testers used an automatic slide show including standardized timer and sound signals. The three tests, took 45 minutes to complete and were administered during both time points in the following order: arithmetic fluency, adaptive number knowledge, and pre-algebra knowledge

3.3.1. Adaptive number knowledge

The Arithmetic Production Task was used in order to measure participants' adaptive number knowledge. The task is a timed, paper-and-pencil instrument which aims to capture students' ability to recognize and use different numerical characteristics and relations during their arithmetic problem solving (McMullen et al., 2017). Students are presented with four to five numbers and the four basic arithmetic operations. Using these numbers and operations, the aim is to produce as many arithmetic

[1] In the experimental group, repeated measures ANOVAs showed no significant interaction effect of time and mode of play (individually or in pairs) for any of the math learning outcomes. Therefore, mode of play was not included in the reporting of the results.

sentences that equal a target number as they can in 90 seconds. To increase the reliability of the task, three more items were added to the post-test. Table 2 shows the items for the two time points. At both times, the same example item was used to demonstrate the task. After explaining the instructions, the example item had to be solved in the same amount of time as a regular item, but students were free to ask clarification questions before and after solving it.

Table 2
Items of the Arithmetic Product Task for Pre- and Post-tests

Item	Pre-test		Post-test	
Example	1,2,3,4	=6	1,2,3,4	=6
1	2,4,8,12,32	=16	2,3,9,15,36	=18
2	1,2,3,5,30	=59	1,2,6,14,42	=8
3	2,4,6,16,24	=12	2,4,6,16,48	=24
4	3,5,30,120,180	=12	1,2,3,30,36	=14
5			2,6,8,32,54	=18
6			3,4,5,6	=7
7			2,3,6,10,18	=38

Students' solutions were transcribed and then automatically scored using Microsoft Excel Macros written for this purpose. The scoring criteria were developed based on the results of previously conducted studies using similar tasks (McMullen et al., 2016, 2017). In the present study, the scoring criteria examined the quantity and complexity of students' solutions on the Arithmetic Production Task. For quantity, the total number of mathematically correct solutions matching the instructions was counted (Correct). For complexity, the total number of multi-operational (Multi-op.) solutions was counted. A solution was multi-operational if two or more different arithmetic operations were used in order to reach the target number (i.e., $2 * 4 + 8 = 16$ multi-operational, but $12 + 2 + 2 = 16$ not). Final scores were made up of the total average of correct solutions and total average of multi-operational solutions. Cronbach's α reliability value for correct solutions was .70 for pre-test and .86 for post-test and for multi-operational solutions .63 for pre-test and .80 for post-test

3.3.2. Arithmetic fluency

Basic arithmetic fluency was measured by the Woodcock-Johnson Math Fluency sub-test (WJ III® Test of Achievement), which consists of two pages with a total of 160 items. Students have to complete as many arithmetic problems (simple addition, subtraction and multiplication) as possible during three minutes (for more details see Schrank, McGrew, & Woodcock, 2001).

3.3.3. Pre-algebra knowledge

The task measuring pre-algebra knowledge consisted of short-answer and multiple choice questions on equation solving (i.e., 12 + ___ = 11 + 15). The task consisted of six multiple-choice items during pre-test and six multiple-choice plus six fill-in items during post-test. At both time points students had eight minutes to solve all items. Each correct item was worth one point, so the maximum pre-test score could be six, while the maximum post-test score could be twelve. Cronbach's α reliability value for the pre-algebra knowledge test was .73 for pre-test and .88 for post-test. Due to the different number of items during the two time points, the standardized sum scores of the pre-algebra knowledge task were used for analyses.

3.3.4. Game performance

The number of maps completed was selected as an indicator of game performance in the present study. As the educational content and game mechanics are integrated in the NNG interaction with the game means also interaction with the relevant mathematical learning content. Additionally, progress in the game (unlocking new levels) is only possible if students complete a set of maps within a certain performance range. Thus, it can be hypothesized that the number of maps completed can be used as a reliable estimate of students' practice with different number combinations and numerical relations.

A map is the basic unit of progress in the NNG, there are 64 maps in total, with 4 target materials needing to be picked up and returned within each map. Each map has a start and an end and provides a substantial amount and variation of mathematically relevant game activity in-between depending also of the active scoring modes within a map. The layout of each map is different, and every target material is placed on different numbers. Together with the two scoring modes (moves and energy) this context provides a wide range of unique arithmetic combinations throughout gameplay and ample opportunities for players to explore number patterns and establish arithmetic connections. As players progress in the NNG maps get progressively harder and there are more and more maps in the energy mode.

During the intervention, each student or pair of students received the NNG on an individual pen drive. All game action was saved and stored in time-stamped text files on these pen drives. The pen drives were collected at the time of the post-test. After the log data was copied, the pen drives were returned to the students as a token of reward for participation. The game log data was summarized and analysed using Microsoft Excel Macros written for this purpose.

3.4. Analyses

As the study was conducted in a naturalistic classroom setting and the unit of randomization was classroom, a variance component analyses was conducted to explore classroom effects and assess the need for multilevel analyses. While it is generally accepted that small ICC values indicate no need for multilevel analyses, there is no agreement about exact guidelines regarding specific cut-off values. Recommendations for considering multilevel methods can range from 0.10 (e.g., Lee, 2000) to 0.25 (e.g., Bowen & Guo, 2011), but the decision is largely dependent on the specific context and study design. For the present study, the ICC values at the classroom level using post-test scores were 0.11 for arithmetic fluency, 0.11 for the number of correct solutions, 0.09 for multi-operational solutions and 0.10 for pre-algebra knowledge.

In order to account for the possible effects of the hierarchical study design, intervention effects were tested using the linear mixed model (LMM) procedure of SPSS version 24 (Heck, Thomas, & Tabata, 2010; West, 2009). For each dependent variable (correct solutions, multi-operational solutions, math fluency, or pre-algebra knowledge), two models were run and compared using the likelihood-ratio test (Field, 2009; West, 2009). The first of the two models included only fixed effects to explore the interaction of treatment and time. The second model was a random intercept model including the same fixed effects and additionally random classroom effects. Thus four separate fixed effects and four separate random effects models were run using the dependent variable measurement occasion and treatment as factors. For each separate dependent variable, level1 variables included observations from the repeated measure nested within the students (level2), who in turn were nested within the classes (level3). The time variable that refers to the measurement occasion (pre- to post-test) was included at level1. The treatment variable that indicates if students received training was added on level2 as a fixed covariate and 'time * treatment' was added as a cross-level interaction term. In the random models, in addition to this fixed model structure, on level3 random intercepts (classroom membership) were estimated. Covariance structures were: 'unstructured' for level1 and 'variance components' for level3.

Based on the comparison of the fixed and random effects models, results from the model with the best fit are reported for all dependent variables. Within the whole sample, compared to the model including fixed effects only, the random intercept model including the classroom effects had the best fit ($p < .001$) for all measured math learning outcomes. Similarly, in grades five and six compared to the fixed effects only model, the random intercept model had the best fit for most of the math learning outcomes ($p < .001$). The two exceptions were in grade six where for math fluency the random intercept model had the best fit using the .05 cutoff value ($p = .04$) and for the multi-operational solutions the random effects model did not have a significantly better fit than the fixed effects model ($p = .11$). In grade four, adding random classroom effects to the model did not significantly affect the model fit.

To answer the second research question, hierarchical linear regression was used to explore the relationship of game performance and mathematical learning outcomes. Post-test scores of correct solution, multi-operational solution, arithmetic fluency and pre-algebra knowledge were used separately as dependent variables. For each analysis grade level was entered as a predictor in the first step, followed by the matching pre-test score of the mathematical outcome used as a dependent variable in the second step, and finally by game performance entered in the third step.

As the second research question examined the relationship of game performance (amount of gameplay) and mathematical learning outcomes, using game log data, only students with sufficient amount of gameplay were included (cf. Bakker et al., 2015; van den Heuvel-Panhuizen et al., 2013). Thus from the experimental group, only students with at least one map completed were used in the regression analyses. Based on this criterion 25 cases were excluded from the analysis, out of a total of 642 participants in the experimental group. On average, students in this group completed 27.02 maps (SD = 12.65) with a range of 1-88 maps completed. The total number of maps completed does not mean unique maps completed; students could repeat the same in order to achieve a better score.

4. Results and discussion

The linear mixed model analyses showed a significant interaction effect between experimental condition and time-point for the number of correct solutions $F(1,1053) = 11.52$, $p < .001$, the number of multi-operation solutions $F(1,1073) = 5.93$, $p = .02$, and arithmetic fluency $F(1,1028) = 5.57$, $p = .02$, but not for pre-algebra knowledge, $F(1,1065) = 1.90$, $p = .17$. Overall, training with the NNG seemed to be more effective than traditional instructional methods in developing students' adaptive number knowledge and math fluency, but not their pre-algebra knowledge. However, students' prior mathematical knowledge in grades four, five and six is very different and this across grade-level analysis may conceal the more substantial impact of the NNG training within the grade levels. Thus, the main examination of the intervention effects was made using separate analysis for each of the three different grade levels.

4.1. Intervention effects by grade level

Tables 3, 4 and 5 show the raw average scores and the results of the linear mixed model analyses of the experimental and control groups during pre- and post-test for grades four, five, and six, respectively.

Table 3

Linear Mixed Model (Fixed Effects): Interaction Effect of Group and Time for Grade Four (n = 133)

Variables		Pre-test *M (SD)*	Post-test *M (SD)*	*F(df)*	*p*
Correct solutions	Exp.	2.08 (.78)	2.69 (.94)	8.11(1,117)	.01
	Cont.	2.29 (.98)	2.48 (1.23)		
Multi-op. solutions	Exp.	.51 (.43)	.81 (.59)	1.17(1,121)	.28
	Cont.	.63 (.46)	.82 (.59)		
Arithmetic fluency	Exp.	55.30 (16.61)	66.15 (14.77)	13.37(1,112)	.00
	Cont.	65.75 (18.27)	70.58 (18.43)		
Pre-algebra knowledge	Exp.	.54 (.32)	.28 (.22)	.38(1,120)	.54
	Cont.	.66 (.30)	.44 (.30)		

Note. Exp. = Experimental group; Cont. = Control group; Multi-op. = Multi-operational

Table 4

Linear Mixed Model (Fixed and Random Effects): Interaction Effect of Group and Time for Grade Five (n = 599)

Variables		Pre-test *M (SD)*	Post-test *M (SD)*	*F(df)*	*p*
Correct solutions	Exp.	2.60 (1.04)	2.94 (1.20)	12.04(1,544)	.00
	Cont.	2.83 (1.22)	2.91 (1.25)		
Multi-op. solutions	Exp.	.77 (.61)	1.15 (.82)	6.75(1,555)	.01
	Cont.	.86 (.72)	1.11 (.83)		
Arithmetic fluency	Exp.	69.24 (18.38)	77.43 (19.76)	.07(1,530)	.80
	Cont.	68.14 (17.00)	75.97 (18.57)		
Pre-algebra knowledge	Exp.	.72 (.30)	.53 (.30)	.11(1,552)	.74
	Cont.	.71 (.30)	.53 (.31)		

Note. Exp. = Experimental group; Cont. = Control group; Multi-op. = Multi-operational

Table 5

Linear Mixed Model (Fixed and Random Effects): Interaction Effect of Group and Time for Grade Six (n = 423)

Variables		Pre-test *M (SD)*	Post-test *M (SD)*	*F(df)*	*p*
Correct solutions	Exp.	3.03 (1.22)	3.33 (1.36)	.00(1,385)	.97
	Cont.	2.90 (1.11)	3.11 (1.24)		
Multi-op. solutions	Exp.	.94 (.67)	1.37 (.82)	.06(1,394)	.82
	Cont.	.92 (.65)	1.36 (.80)		
Arithmetic fluency	Exp.	75.66 (16.77)	84.94 (19.11)	3.00(1,381)	.08
	Cont.	74.91 (16.43)	82.43 (20.23)		
Pre-algebra knowledge	Exp.	.77 (.29)	.67 (.29)	6.68(1,389)	.01
	Cont.	.81 (.27)	.62 (.32)		

Note. Exp. = Experimental group; Cont. = Control group; Multi-op. = Multi-operational.

As results show, practice with the NNG had varying effects on different arithmetic skills in different grade levels. In grade four, results showed a significant interaction of time and group for correct solutions and arithmetic fluency, but no interaction effects were found for multi-operational solutions or pre-algebra knowledge. In grade five, results show significant interaction effects of time and group for correct solutions and multi-operational solutions, but no significant interaction effects for math fluency and pre-algebra knowledge. In grade six, results show a significant interaction effect for pre-algebra knowledge.

These results suggest that the NNG was able to support the development of different aspects of arithmetic and mathematical development at different ages. In grade four, the game supported the development of basic calculation fluency and the more basic aspect of adaptive number knowledge (finding correct solutions). In grade five, where calculation fluency is already more established, there was a positive effect of gameplay on both correct solutions and also in the more complex aspect of adaptive number knowledge (finding multi-operational solutions). Effects for pre-algebra knowledge were only found among sixth graders who otherwise did not benefit much from the game, the settings of which might have been not challenging enough for them. Alternatively, it could be that pre-algebra knowledge is a more complex skill that develops on top of other abilities, and practice with the NNG is only beneficial for pre-algebra skills if students already have a relatively high level of prior knowledge in arithmetic. Finding pre-algebra knowledge intervention effects only in grade six is similar to the results of van den Heuvel-Panhuizen and colleagues (2013) who found the strongest intervention effects in grade six – when compared to fourth and fifth graders – as a result of a game-based training which aimed to train quantitative reasoning that is foundational to early algebra knowledge.

In order to progress in the NNG, players need to mentally execute and compare many calculations, look for strategically important numbers (i.e., numbers with many divisors, numbers close to the target number), and continuously refine their strategies according to the different game modes and challenges. This requires noticing and using numerical characteristics and relations in order to arrive at efficient problem solving strategies in mental calculations (Brezovszky et al., 2015; McMullen et al., 2017; Threlfall, 2009). This practice is in line with the idea that in order to develop more flexibility and adaptivity with arithmetic problem solving, students need to have a chance to practice with various number-operation combinations (Baroody, 2003; Verschaffel et al., 2009). It seems that the playful practice in the NNG indeed enhances students' awareness of numerical relations, so they can notice more of these relations (i.e., correct solutions) and by having a richer repertoire of number relations come up with more complex solutions (i.e., multi-operational solutions). With more attention on calibrating gameplay for particular grades, it may be possible to elicit even more pronounced effects. For example, offering younger students opportunities to strengthen their understanding of the natural number base-ten system, while stimulating older students to explore their understanding of arithmetic concepts such as the inverse nature of multiplication and division more deeply.

The lack of methodologically sound interventions and large-scale randomized control studies is a common finding of many reviews and meta-analyses in game-based learning (Hainey, Connolly, Boyle, Wilson, & Razak, 2016; Wouters et al., 2013; Young et al., 2012). This is problematic because it can distort the picture regarding the effectiveness of game-based learning as smaller and more controlled quasi-experimental designs might inflate intervention effects, while effect sizes are generally very small for large-scale randomized control trials (Cheung & Slavin, 2013; Wouters et al., 2013). In spite of the possible small effects and loss of control in implementation, conducting large-scale randomized interventions is important as they provide an opportunity to gain more ecological validity and realistic estimate regarding the efficiency of game-based learning environments in the classroom setting (Winn, 2003; Cook, & Payne, 2001).

4.2. Effects of game performance

The second research question of the present study explored the relationship of game performance in the NNG and the development of students' mathematical learning outcomes. Hierarchical linear regression analyses were conducted in order to investigate the impact of game performance on the improvement of the experimental group's mathematical learning outcomes. As Table 7 shows, after taking into account grade level and pre-test scores, the number of maps completed still explains part of the variance for all outcome measures. Since the linear regression cannot take into account the nested data structure, to confirm the significance level of beta values linear mixed model analyses including pre-test scores, grade level and maps completed as fixed effects and random classroom effects was performed. Results showed a significant main effect for the number of maps completed $p < .001$.

Table 7
Hierarchical Linear Regression Analyses: Impact of Game Performance on the Mathematical Learning Outcomes in the Experimental Group (n = 617)

		Post-test							
		Correct solutions		Multi-op. solutions		Math Fluency		Pre-algebra	
		β	*ΔR²*	*β*	*ΔR²*	*β*	*ΔR²*	*β*	*ΔR²*
1	Grade-level	-.06	.03***	.02	.04***	.02	.09***	.19***	.12***
2	Pre-test	.68***	.45**	.53***	.32***	.78***	.57***	.41***	.19***
3	Game performance	.09**	.01**	.20**	.03***	.09**	.01**	.21***	.04***
	Total R^2		.49		.39		.66		.36

Note 1. * *p* < .05, ** *p* < .01, *** *p* < .001. *Note 2.* Pre-test = Corresponding pre-test variable to the post-test variable (correct solutions, multi-operational solutions, math fluency or pre-algebra). *Note 3.* Multi-op. = Multi-operational

Results show that students who had more practice with the NNG also benefited more from the gameplay. This suggests that the type of action necessary in order to make meaningful progress in the NNG could be transferred outside the game environment as well. These results are in line with a previous small-scale intervention study using the NNG (Brezovszky et al., 2015).

One issue with many game-based learning environments is that players engage with content or features that are irrelevant from the perspective of the learning aims (Clark et al., 2011; Wouters & Oostendorp, 2013). Even if students use the environment as intended, connecting the in-game learning material with educational content outside of the game-based learning environment is often problematic (Clark et al., 2011; Lajoie, 2005). In the NNG, at least eight paths from a starting point to a target were taken within a completed map. But, considering that not all numbers are available and that players need to adapt their strategies according to the active game mode (moves or energy), by the time a map was completed a player most likely undertook a large number of mental calculations (Brezovszky et al., 2013). In light of the results, it seems that extended practice with the NNG was able to develop students' recognition and use of numerical characteristics and relations.

4.3. Limitations

A few major limitations of the study are connected to the design decision which was to conduct an ecologically valid large-scale intervention study. First, one of the major issues with this design is the loss of control over the implementation details, which in the present study were only accounted for by analyzing the game log data. It was an important aim of the present study to design an intervention that is as close to everyday classroom practices as possible, especially given the large autonomy Finnish teachers have. Thus, guidelines regarding the implementation of the NNG did not specify strict details and teachers were free to decide, for example, the mode of play. This freedom could result in a substantial amount of variation and can make the interpretation of the results difficult. However, as students' average time on task did not show a large variation across the classrooms, it can be assumed that most students benefited from the training to a similar extent.

Connected to methodological decisions, it is important to mention that randomization on a classroom level might have affected results. However, a classroom is a natural unit of analysis in educational research and randomization by individual students is rarely done as it is almost impossible for this type of study designs. Multilevel statistical analyses were conducted that takes into account the nested nature of the data. Additionally, ICC scores were low and the amount of participating classes was much higher than the necessary number suggested by the power analyses which strengthens the generalizability of results despite the cluster randomized design.

Using regular instruction and no alternative intervention method (game-based or traditional) in the control group affects the interpretation of the results in the present study. While the time period for math instruction was the same for both groups (i.e. NNG play replaced a part of normal mathematics lessons), the novelty of playing the NNG may have had an effect on student performance on the post-tests. However, in a separate study of situational interest using the NNG, results showed substantial variation in situational interest between students and across sessions which was mainly explained by prior personal mathematics interest (Rodríguez-Aflecht et al., 2018). This suggests that the novelty effect cannot explain group differences in the mathematical learning outcomes of the current study. With regards to the type and quality of the regular teaching practice it is important to add here that developing flexibility and adaptivity with arithmetic is an aim of the Finnish National Core Curriculum and teachers are encouraged to use various representations, games, discovery learning, and discussion, as well as plenty of group work in their everyday math classroom practice (Kupari, 2008). For more comprehensive conclusions, future studies should compare gameplay with the NNG with other game-based learning environments and more elaborated training methods of flexibility with arithmetic problem solving without digital games.

A number of issues with the instruments used in the present study should also be addressed in future studies. First, the relatively low pre-test reliability of the Arithmetic Production Task raises concerns. These low values could be explained by the low number of items used during pre-test and/or the novelty of the task type. Increasing the number of items in future studies, as was done on the post-test in the present study, is recommended. With regards to the measure of game performance, it has to be acknowledged that the number of maps completed is a crude indicator and there might be more subtle differential patterns underlying this measure (i.e., performance within a map, replay trials of the same map, more qualitative aspects of the problem solving strategies used etc.). As the NNG logs different aspects of the game data, future studies could explore this question in more detail.

5. Conclusions

NNG is, to our knowledge, the first game-based learning environment directly focused on enhancing adaptive arithmetic knowledge and skills, which have been difficult to support in traditional classrooms teaching. This suggests that the NNG could be a flexible tool to develop complex mathematical skills and knowledge in a naturalistic classroom setting. Adaptive number knowledge has been suggested to underlie proficiency with whole-number arithmetic problem solving strategies (McMullen et al., 2017), a core feature of arithmetic development (Nunes et al., 2016; Verschaffel et al., 2009). By promoting students' adaptive number knowledge, the NNG is a valuable pedagogical tool for supporting students' development of flexible and adaptive arithmetic problem solving. This type of support may have long-lasting value in mathematical development, for example with learning algebra. Future studies that focus on integrating gameplay and more typical classroom activities (Clark, Tanner-Smith, & Killingsworth, 2016; Wouters & Oostendorp, 2013) may improve both the embedding of the NNG in the curriculum and its positive outcomes for students.

References

Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., de Freitas, S., Louchart, S., … De Gloria, A. (2015). Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology*, *46*(2), 391–411. https://doi.org/10.1111/bjet.12113

Bakker, M., van den Heuvel-Panhuizen, M., & Robitzsch, A. (2015). Effects of playing mathematics computer games on primary school students' multiplicative reasoning ability. *Contemporary Educational Psychology*, *40*, 55–71. https://doi.org/10.1016/j.cedpsych.2014.09.001

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In *The Development of Arithmetic Concepts and Skills: Constructive Adaptive Expertise* (pp. 1–33). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410607218

Blöte, A. W., Klein, A. S., & Beishuizen, M. (2000). Mental computation and conceptual understanding. *Learning and Instruction*, *10*(3), 221–247. https://doi.org/10.1016/S0959-4752(99)00028-6

Blöte, A. W., Van der Burg, E., & Klein, A. S. (2001). Students' flexibility in solving two-digit addition and subtraction problems: Instruction effects. *Journal of Educational Psychology*, *93*(3), 627–638. https://doi.org/10.1037/0022-0663.93.3.627

Bowen, N. K., & Guo, S. (2011). *Structural equation modeling.* Oxford University Presshttps://doi.org/10.1093/acprof:oso/9780195367621.001.0001.

Brezovszky, B., Lehtinen, E., McMullen, J., Rodriguez, G., & Veermans, K. (2013). Training flexible and adaptive arithmetic problem solving skills through

Brezovszky, B., Lehtinen, E., McMullen, J., Rodriguez, G., & Veermans, K. (2013). Training flexible and adaptive arithmetic problem solving skills through exploration with numbers: The development of NumberNavigation game. In *7th European Conference on Games Based Learning, ECGBL 2013* (Vol. 2, pp. 626–634).

Brezovszky, B., Rodríguez-Aflecht, G., McMullen, J., Veermans, K., Pongsakdi, N., Hannula-Sormunen, M. M., & Lehtinen, E. (2015). Developing adaptive number knowledge with the number navigation game-based learning environment. In J. Torbeyns, E. Lehtinen, & J. Elen (Eds.), *Describing and Studying Domain-Specific Serious Games* (pp. 155–170). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-20276-1_10

Cook, T. D., & Payne, M. R. (2001). Objecting to the objections to using random assignment in educational research. In F. Mosteller, & R. F. Boruch (Eds.), *Evidence matters* (pp. 150-177). Washington, DC: Brookings Institution Press.

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, *9*, 88–113. https://doi.org/10.1016/j.edurev.2013.01.001

Clark, D. B., Nelson, B. C., Chang, H. Y., Martinez-Garza, M., Slack, K., & D'Angelo, C. M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers and Education*, *57*(3), 2178–2195. https://doi.org/10.1016/j.compedu.2011.05.007

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. https://doi.org/10.3102/0034654315582065

Devlin, K. (2011). *Mathematics education for a new era: Video games as a medium for learning*. A K Peters.

Dowker, A. (1992). Estimation Strategies of Computational. *Journal for Research in Mathematics Education*, *23*(1), 45–55.

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.

Finnish National Agency for Education (2014). Core curriculum for basic Education 2014. http://www.oph.fi/english/curricula_and_qualifications/basic_education/curricula_2014

Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating Children to Learn Effectively: Exploring the Value of Intrinsic Integration in Educational Games. *Journal of the Learning Sciences*, *20*(2), 169–206. https://doi.org/10.1080/10508406.2010.508029

Hainey, T., Connolly, T. M., Boyle, E. A., Wilson, A., & Razak, A. (2016). A systematic literature review of games-based learning empirical evidence in primary education. *Computers & Education*, *102*, 202–223. https://doi.org/10.1016/j.compedu.2016.09.001

Heck, R. H., Thomas, S. L. T., & Tabata, L. N. (2010). *Multilevel and Longitudinal Modeling with IBM SPSS.* Routledge. New York.

Hedges, L. V., & Rhoads, C. (2010). Statistical power analysis. In *International Encyclopedia of Education* (pp. 436–443). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.01356-7

Heinze, A., Star, J. R., & Verschaffel, L. (2009). Flexible and adaptive use of strategies and representations in mathematics education. *ZDM - International Journal on Mathematics Education*, *41*, 535–540. https://doi.org/10.1007/s11858-009-0214-4

Heirdsfield, A. M., & Cooper, T. J. (2004). Factors affecting the process of proficient mental addition and subtraction: Case studies of flexible and inflexible computers. *Journal of Mathematical Behavior*, *23*(4), 443–463. https://doi.org/10.1016/j.jmathb.2004.09.005

Heirdsfield, A. M (2011). Teaching mental computation strategies in early mathematics. *Young Children, 66*(2), pp. 96-102.

Hickendorff, M. (2017). Dutch sixth graders' use of shortcut strategies in solving multidigit arithmetic problems. *European Journal of Psychology of Education*, 1–18. https://doi.org/10.1007/s10212-017-0357-6

Ke, F. (2008). A case study of computer gaming for math: Engaged learning from gameplay? *Computers and Education*, *51*(4), 1609–1620. https://doi.org/10.1016/j.compedu.2008.03.003

Ke, F. (2009). A Qualitative Meta-Analysis of Computer Games as Learning Tools. *Handbook of Research on Effective Electronic Gaming in Education*, 1–32. https://doi.org/10.4018/978-1-59904-808-6.ch001

Kiili, K. (2007). Foundation for problem-based gaming. *British Journal of Educational Technology*, *38*(3), 394–404. https://doi.org/10.1111/j.1467-8535.2007.00704.x

Kupari, P. (2008). Mathematics education in Finnish comprehensive school: characteristics contributing to student success. In *Proceedings of the XI International Congress in Mathematics Education*.

Lajoie, S. P. (2005). Extending the scaffolding metaphor. *Instructional Science*, *33*(5–6), 541–557. https://doi.org/10.1007/s11251-005-1279-2

Laski, E. V., Ermakova, A., & Vasilyeva, M. (2014). Early use of decomposition for addition and its relation to base-10 knowledge. *Journal of Applied Developmental Psychology*, *35*(5), 444–454. https://doi.org/10.1016/j.appdev.2014.07.002

Lee, V. E. (2000). Using Hierarchical Linear Modeling to Study Social Contexts: The Case of School Effects, *Educational Psychologist, 35(2),* 125-141, doi: 10.1207/s15326985ep3502_6

Lehtinen, E., Brezovszky, B., Rodríguez-Aflecht, G., Lehtinen, H., Hannula-Sormunen, M. M., McMullen, J., … Jaakkola, T. (2015). Number Navigation Game (NNG): Design Principles and Game Description. In J. Torbeyns, E. Lehtinen, & J. Elen (Eds.), *Describing and Studying Domain-Specific Serious Games* (pp. 45–61). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-20276-1_4

Lehtinen, E., Hannula-Sormunen, M., McMullen, J., & Gruber, H. (2017). Cultivating mathematical skills: from drill-and-practice to deliberate practice. *ZDM - Mathematics Education*, *49*(4), 625–636. https://doi.org/10.1007/s11858-017-0856-6

Li, Q., & Ma, X. (2010). A Meta-analysis of the Effects of Computer Technology on School Students' Mathematics Learning. *Educational Psychology Review*, *22*(3), 215–243. https://doi.org/10.1007/s10648-010-9125-8

Lowrie, T., & Jorgensen, R. (2015). Digital Games and Learning: What's New Is Already Old? (pp. 1–9). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9517-3_1

McMullen, J., Brezovszky, B., Rodríguez-Aflecht, G., Pongsakdi, N., Hannula-Sormunen, M. M., & Lehtinen, E. (2016). Adaptive number knowledge: Exploring the foundations of adaptivity with whole-number arithmetic. *Learning and Individual Differences*, *47*, 172–181. https://doi.org/10.1016/j.lindif.2016.02.007

McMullen, J., Brezovszky, B., Hannula-Sormunen, M. M., Veermans, K., Rodríguez-Aflecht, G., Pongsakdi, N., & Lehtinen, E. (2017). Adaptive number knowledge and its relation to arithmetic and pre-algebra knowledge. *Learning and Instruction*, *49*, 178–187. https://doi.org/10.1016/j.learninstruc.2017.02.001

Mulligan, J. T., & Mitchelmore, M. C. (2013). Early awareness of mathematical pattern and structure. In L. D. English & J. T. Mulligan (Eds.), Reconceptualising early mathematics learning (pp. 29-45). New York: Springer. https://doi.org/10.1007/978-94-007-6440-8_3

National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. Reston.

Nunes, T., Dorneles, B. V., Lin, P.-J., & Rathgeb-Schnierer, E. (2016). *Teaching and Learning About Whole Numbers in Primary School*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-45113-8

Pope, H., & Mangram, C. (2015). Wuzzit Trouble: The Influence of a Digital Math Game on Student Number Sense. *International Journal of Serious Games*, *2*(4). https://doi.org/10.17083/ijsg.v2i4.88

Rittle-Johnson, B., & Star, J. R. (2009). Compared With What? The Effects of Different Comparisons on Conceptual Knowledge and Procedural Flexibility for Equation Solving. *Journal of Educational Psychology*, *101*(3), 529–544. https://doi.org/10.1037/a0014224

Rodríguez-Aflecht, G., Jaakkola, T., Pongsakdi, N., Hannula-Sormunen, M., Brezovszky, B., & Lehtinen, E. (2018). The development of situational interest during a digital mathematics game. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12239

Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Kibrick, M., Graham, J., … Martinez, M. E. (2014). A Randomized Trial of an Elementary School Mathematics Software Intervention: Spatial-Temporal Math. *Journal of Research on Educational Effectiveness*, *7*(4), 358–383. https://doi.org/10.1080/19345747.2013.856978

Salen, K., & Zimmerman, E. (2003). *Rules of Play: Game Design Fundamentals*. The MIT Press.

Schneider, M., Rittle-Johnson, B., & Star, J. R. (2011). Relations among conceptual knowledge, procedural knowledge, and procedural flexibility in two samples differing in prior knowledge. *Developmental Psychology*, *47*(6), 1525–1538. https://doi.org/10.1037/a0024997

Schrank, F. A., McGrew, K. S., & Woodcock, R. W. (2001). *Technical Abstract (Woodcock-Johnson III Assessment Service Bulletin No. 2)*. Itasca, IL: Riverside Publishing. Retrieved from http://www.riverpub.com/clinical/pdf/WJIII_ASB2.pdf

Star, J. R., & Seifert, C. (2006). The development of flexibility in equation solving. *Contemporary Educational Psychology*, *31*(3), 280–300. https://doi.org/10.1016/j.cedpsych.2005.08.001

Star, J. R., & Rittle-Johnson, B. (2008). Flexibility in problem solving: The case of equation solving. *Learning and Instruction*, *18*(6), 565–579. https://doi.org/10.1016/j.learninstruc.2007.09.018

Star, J. R., & Newton, K. J. (2009). The nature and development of experts' strategy flexibility for solving equations. *ZDM - International Journal on Mathematics Education*, *41*(5), 557–567. https://doi.org/10.1007/s11858-009-0185-5

Squire, K. (2008). Open-Ended Video Games: A Model for Developing Learning for the Interactive Age. In Salen K., *The ecology of games: Connecting youth, games, and learning* (p. 167-198). Cambridge MA: The MIT Press.

Threlfall, J. (2002). Flexible Mental Calculation. *Educational Studies in Mathematics*, *50*(1), 29–47. https://doi.org/10.1023/A:1020572803437

Threlfall, J. (2009). Strategies and flexibility in mental calculation. *ZDM*, *41*(5), 541–555. https://doi.org/10.1007/s11858-009-0195-3

van den Heuvel-Panhuizen, M., Kolovou, A., & Robitzsch, A. (2013). Primary school students' strategies in early algebra problem solving supported by an online game. *Educational Studies in Mathematics*, *84*(3), 281–307. https://doi.org/10.1007/s10649-013-9483-5

Verschaffel, L., Luwel, K., Torbeyns, J., & Van Dooren, W. (2009). Conceptualizing, investigating, and enhancing adaptive expertise in elementary mathematics education. *European Journal of Psychology of Education*, *24*(3), 335–359. https://doi.org/10.1007/BF03174765

West, B. T. (2009). Analysing longitudinal data with the linear mixed models procedure in SPSS. *Eval Health Prof*, *32*(3), 207–228. https://doi.org/10.1177/0163278709338554

Winn, W. (2003). Research Methods and Types of Evidence for Research in Educational Technology. *Educational Psychology Review*, *15(4),* 367-373. https://doi.org/10.1023/A:1026131416764

Wouters, P., & Oostendorp, H. Van. (2013). Computers & Education A meta-analytic review of the role of instructional support in game-based learning. *Computers & Education*, *60*(1), 412–425. https://doi.org/10.1016/j.compedu.2012.07.018

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. https://doi.org/10.1037/a0031311

Yang Hansen, K., Gustafsson, J.-E. , & Rosén, M. (2014). School performance difference and policy variations in Finland, Norway and Sweden. In K. Yang Hansen, J.-E. Gustafsson, M. Rosén, S. Sulkunen, K. Nissinen, P. Kupari, R.F. Ólafsson, J.K. Björnsson, L.S. Grønmo, L. Rønberg, J. Mejding, I.C. Borge, & A. Hole. *Northern Lights on TIMMS and PIRLS 2011*. TemaNord 2014:528, 25-48.

Young, M. F., Slota, S., Cutter, B., R., Jalette, G., Mullin, G., … Yukhymenko, 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. https://doi.org/10.3102/0034654312436980