

Effects of playing mathematics computer games on primary school students' multiplicative reasoning ability



Marjoke Bakker^{a,b,*}, Marja van den Heuvel-Panhuizen^a, Alexander Robitzsch^c

^a Freudenthal Institute for Science and Mathematics Education, Faculty of Science & Faculty of Social and Behavioural Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands

^b Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

^c Federal Institute for Education Research, Innovation and Development of the Austrian School System, Alpenstraße 121, 5020 Salzburg, Austria

ARTICLE INFO

Article history:

Available online 7 October 2014

Keywords:

Educational computer games
Mathematics education
Multiplicative reasoning
Primary school

ABSTRACT

This study used a large-scale cluster randomized longitudinal experiment ($N = 719$; 35 schools) to investigate the effects of online mathematics mini-games on primary school students' multiplicative reasoning ability. The experiment included four conditions: playing at school, integrated in a lesson (E_{school}), playing at home without attention at school (E_{home}), playing at home with debriefing at school ($E_{\text{home-school}}$) and, in the control group, playing at school mini-games on other mathematics topics (C). The mini-games were played in Grade 2 and Grade 3 (32 mini-games in total). Using tests at the end of each grade, effects on three aspects of multiplicative reasoning ability were measured: *knowledge* of multiplicative number facts, *skills* in multiplicative operations, and *insight* in multiplicative number relations and properties of multiplicative operations. Through path analyses it was found that the mini-games were most effective in the $E_{\text{home-school}}$ condition, where both students' skills and their insight were positively affected as compared to the control group (significant d s ranging from 0.22 to 0.29). In the E_{school} condition, an effect was only found for insight in Grade 2 ($d = 0.35$), while in the E_{home} condition no significant effects were found. Students' gameplay behavior (time and effort put in the games) was in some cases, but not always, related to their learning outcomes.

© 2014 The Authors. Published by Elsevier Inc. All rights reserved.

1. Introduction

1.1. Educational computer games

Computer games have often been suggested as promising educational tools (e.g., Egenfeldt-Nielsen, 2005; Malone, 1981; Prensky, 2001; Tobias, Fletcher, Dai, & Wind, 2011). The most commonly mentioned benefit of computer games for education is their motivational aspect (e.g., Garris, Ahlers, & Driskell, 2002; Malone, 1981; Malone & Lepper, 1987; Prensky, 2001). In addition, games are assumed to be beneficial for learning because they can provide immediate feedback. Players often instantly see the consequences of their actions in the game (e.g., Prensky, 2001). Moreover, games allow players to try, make mistakes, and then try again without losing face (e.g., Gee, 2005). Because of this risk-free environment and the immediate feedback provided by

the computer, players are stimulated to explore and experiment, as was pointed out by Kirriemuir (2002). In other words, games can offer students opportunities for experiential learning (e.g., Egenfeldt-Nielsen, 2005; Garris et al., 2002), enabling them to discover new rules and strategies.

Because of these presumed advantages, computer games are more and more becoming part of primary school education (e.g., Williamson, 2009). In accordance with the expected educational benefits of computer games, a meta-analysis by Wouters, Van Nimwegen, Van Oostendorp, and Van der Spek (2013) reported an overall positive effect of educational computer games in comparison to conventional instruction. However, when only randomized studies were taken into account, they did not find a significant effect. Furthermore, other review studies revealed that there is still insufficient experimental evidence for the effectiveness of educational computer games in the school practice (Tobias et al., 2011; Vogel et al., 2006; Young et al., 2012), and that large-scale in-class longitudinal studies are needed (Tobias et al., 2011; Young et al., 2012). Authors of review articles argued that studies on the effects of games and other educational software quite often suffer from methodological shortcomings, such as not using a control group (e.g., Vogel et al., 2006), not applying random assignment to condi-

* Corresponding author at: Freudenthal Institute, Utrecht University, Princetonplein 5, 3584 CC Utrecht, the Netherlands. Tel.: +31 30 253 1179; fax: +31 30 253 7494.

E-mail addresses: m.bakker@uu.nl, m.bakker@let.ru.nl (M. Bakker), m.vandenheuvel-panhuizen@uu.nl (M. van den Heuvel-Panhuizen).

tions (e.g., Slavin & Lake, 2008), using a small sample (e.g., Bai, Pan, Hirumi, & Kebritchi, 2012), and not accounting for the nested data structure (e.g., Honey & Hilton, 2011; Slavin & Lake, 2008).

Also in primary mathematics education, computer games and other educational software are often used (e.g., Mullis, Martin, Foy, & Arora, 2012). Yet, also for the domain of mathematics, evidence for the effects of educational computer games is still insufficient, as is apparent from Bai et al.'s (2012) literature overview. Meta-analyses by Li and Ma (2010) and Slavin and Lake (2008) did show that in general the use of ICT in mathematics education positively affects learning outcomes, but in these analyses games were not taken as a separate category.

To gain evidence about the effectiveness of deploying computer games in mathematics education, we conducted a large-scale randomized experiment, with a longitudinal design. The focus was on mini-games in the domain of multiplicative reasoning (multiplication and division) in the early grades of primary school, where formal instruction of multiplicative reasoning commonly commences (e.g., Department for Education UK, 2011; NCTM, 2006; Van den Heuvel-Panhuizen, 2008).

1.2. Using computer games in mathematics education

1.2.1. Mini-games

A frequently used type of computer game in mathematics education is the so-called mini-game (e.g., Jonker, Wijers, & Van Galen, 2009; Panagiotakopoulos, 2011). Mini-games are short, focused games that are easy to learn (e.g., Frazer, Argles, & Wills, 2007; Jonker et al., 2009). They are often easily accessible (commonly free of charge), and usually have a flexible time duration; one game often takes only a few minutes and can be repeated endlessly (e.g., Jonker et al., 2009). Earlier studies have shown that mini-games have potential for mathematics education. In an evaluation study by Panagiotakopoulos, Sarris, and Koleza (2013), for example, positive learning outcomes were found in fifth-grade students who worked with a number mini-game. Furthermore, Miller and Robertson (2011) showed the effectiveness of handheld mathematics mini-games in improving 10- and 11-year-olds' mental computation skills.

1.2.2. Multiplicative number fact knowledge, skills, and insight

In learning multiplicative reasoning, it is important to develop ready knowledge of number facts (the multiplication tables), and skills in calculating multiplication and division operations. In addition, students need to develop insight in, or understanding of, multiplicative number relations (e.g., Anghileri, 2006; Nunes, Bryant, Barros, & Sylva, 2012). They should, for example, have insight into the factors of numbers and the properties of multiplication (see, e.g., Chang, Sung, Chen, & Huang, 2008), like the commutative property (e.g., $3 \times 7 = 7 \times 3$) and the distributive property (e.g., $6 \times 7 = 5 \times 7 + 1 \times 7$). These three aspects of multiplicative reasoning ability – number fact knowledge, operation skills, and insight – parallel the three types of knowledge often distinguished in mathematics education: declarative knowledge, procedural knowledge, and conceptual knowledge (see, e.g., Miller & Hudson, 2007).

Many of the computer games and other educational software currently used in primary school mathematics education focus on the first two aspects: number fact knowledge and operation skills (e.g., Mullis et al., 2012). However, computer games can also be employed for developing mathematical insight (see, e.g., Van Borkulo, Van den Heuvel-Panhuizen, Bakker, & Loomans, 2012). Jonker et al. (2009), for example, described a mini-game for enhancing primary school students' understanding of divisibility, and two studies reported by Klawe (1998) showed the effectiveness of computer games in fostering fifth-graders' understanding of several mathematical concepts. In fact, Ke (2009), in her review article,

noted that games seem more useful to promote higher-order thinking than factual knowledge acquisition. The instructional power of games that are focused on insight development is often related to the educational theory of experiential learning (see, e.g., Kebritchi, Hirumi, & Bai, 2010). In such games, students can learn new concepts and rules by experimenting with different mathematical strategies and discovering which strategies are convenient. To make this learning process happen, reflection is crucial, as is stated, for example, by Egenfeldt-Nielsen (2005) and Garris et al. (2002). Students can utilize reflection to generalize what they have learned, which leads to transfer. In this way, what is learned can also be applied outside the game (see, e.g., Tobias et al., 2011). However, many researchers argue that this reflection does not occur spontaneously in students (e.g., Leemkuil & De Jong, 2004). It is proposed that class discussion after playing a game is needed to encourage reflection (e.g., Egenfeldt-Nielsen, 2005; Garris et al., 2002; Klawe, 1998). In such a discussion – also called debriefing (e.g., Garris et al., 2002) – the learning points from the game are emphasized and different possible strategies are compared (e.g., Klawe, 1998). Indeed, Wouters et al. (2013), in their meta-analysis, found that interventions with computer games are more effective when the games are supplemented with other instructional methods, such as debriefing sessions, than when they are presented as a stand-alone activity. Also support before and during the game is assumed to foster learning (e.g., Leemkuil & De Jong, 2004).

1.3. Playing games at school vs. at home

Mini-games can be played at school (a formal setting) as well as at home (an informal setting; see, e.g., Honey & Hilton, 2011). Because of the involvement of the teacher, playing in a formal setting at school has the advantage that all instructional aspects of the games can be exploited by discussing them in a lesson. Moreover, the teacher has control over whether the games are played. However, playing in an informal setting at home, which also occurs a lot (e.g., Ault, Adams, Rowland, & Tiemann, 2010; Jonker et al., 2009), has advantages as well. Jonker et al. (2009), for example, reported that the Dutch mathematics games website Rekenweb is visited mainly during after-school hours, which, for the students involved, implies an extension of the time that is spent on mathematics. According to researchers like Honey and Hilton (2011) and Tobias et al. (2011), an important characteristic of educational computer games is that their motivational effect can cause students to be involved in a learning activity for a longer time period than is regularly the case. In a study by Sandberg, Maris, and De Geus (2011), for example, primary school students who were offered a mobile game were found to voluntarily spend extra time on language learning, which led to increased learning. Besides the advantage of extra learning time, playing at home may imply that students have more control over the learning activity. This so-called learner control is often mentioned as an important motivating factor of educational computer games (e.g., Malone & Lepper, 1987), and can lead to improved learning. In a study by Cordova and Lepper (1996), for example, learner control in the form of choice of avatars and character names in a mathematics game resulted in enhanced learning outcomes. Freedom of choice concerning which game is played, and when and for how much time it is played, can also be considered an aspect of learner control (e.g., Wouters et al., 2013). When educational games are played in students' free time, this freedom of choice is larger than when they are played at school, which may lead to higher motivation in students, and consequently to higher learning outcomes.

A possible approach that combines the advantages of playing at school and those of playing at home, is to play the games at home with a debriefing at school. In this way, students are stimulated to reflect upon their experiences in the games (see Section 1.2.2).

This manner of utilizing computer games in education was, for example, found to be effective in an experiment focused on informal algebraic reasoning in primary school (Kolovou, Van den Heuvel-Panhuizen, & Köller, 2013). Such an approach is in line with Honey and Hilton's (2011) suggestion of bridging formal and informal learning contexts using educational games.

1.4. Gameplay behavior

When using games in education, the amount of time and effort students spend on the games may be an important predictor of their learning outcomes. Indeed, Jansen et al. (2013) found that students who had practiced more problems in a game environment on the automatization of number facts, exhibited higher gains in their number fact knowledge. However, the relation might be less clear when using games meant to contribute to gaining mathematical insight. Kolovou et al. (2013), for example, did not find a relation between students' online game involvement – measured as a composite variable consisting of logged-in time and online game actions – and their gain in understanding co-varying quantities. They explained this finding by suggesting that through the class debriefing sessions, students who had not played the game at home could have learned from the experiences of the students who had played. As another possible explanation, Kolovou and colleagues hypothesized that students might require only a limited amount of experience to discover the concepts to be learned in the game, and any further game playing would not result in more learning.

1.5. Our study

In the current study we investigated the effects of multiplicative mini-games on students' multiplicative reasoning ability in Grade 2 and Grade 3. In terms of the genres of game research distinguished by Mayer (2011), our research falls within the cognitive consequences genre, investigating what students learn from the mini-games. We examined the effectiveness of three different ways of deploying the mini-games: playing at school, playing at home, and playing at home with debriefing at school. The mini-games used in the study focused both on automatizing multiplicative number facts and multiplicative operations (through practicing), and on developing insight in multiplicative number relations (through exploring and experimenting).

The aim of our study was to investigate the effects of a mini-games intervention when implemented as part of the regular educational practice. As such, we studied the added value of the mini-games when employed as part of the regular multiplicative reasoning curriculum.

We had previously performed a preliminary analysis on the effects of the mini-games in Grade 2 (Bakker, Van den Heuvel-Panhuizen, Van Borkulo, & Robitzsch, 2012, 2013), using a combined measure of multiplicative ability that included multiplicative operation skills and insight. While the analysis revealed no significant effects, a marginally significant effect was found for the condition in which the games were played at home and debriefed at school ($p = .07$, $d = 0.23$). The current study covered a two-year intervention in Grade 2 and Grade 3 and investigated the effects of the games on three aspects of students' multiplicative reasoning ability: ready *knowledge* of multiplicative number facts, multiplicative operation *skills*, and *insight* in multiplicative number relations. Additionally, we examined the role of gameplay behavior.

The following research questions were investigated:

1. Does an intervention with multiplicative mini-games – either played at school, played at home, or played at home and debriefed at school – affect students' learning outcomes in multiplicative reasoning?

2. Does an intervention with multiplicative mini-games affect students' learning outcomes in all three aspects of multiplicative reasoning: knowledge, skills, and insight?
3. In what setting – playing at school, playing at home, or playing at home with debriefing at school – are the multiplicative mini-games most effective?
4. Are students' learning outcomes related to their gameplay behavior?

Our hypothesis for Research question 1 was that, in each of the three game-playing settings, the intervention with multiplicative mini-games would positively affect the learning of multiplicative reasoning, in comparison to the regular mathematics curriculum without these mini-games. This hypothesis is based on the motivating environment and immediate feedback provided by educational games, and on the possibilities for experiential learning offered by the games. Regarding Research question 2, we hypothesized that the mini-games would be effective in enhancing all three aspects of students' multiplicative reasoning ability. With respect to Research question 3, we hypothesized the mini-games to be most effective when played at home and debriefed at school. In this setting the advantage of playing at home (extra time-on-task, more learner control) is combined with the advantage of playing at school (debriefing). Additionally, for Research question 4, our hypothesis was that students' gameplay behavior would be positively related to their learning outcomes with respect to number fact knowledge and skills. The relation with their insight learning outcomes may be less clear, but if there is a relation we expect it to be positive.

2. Method

2.1. Research design

To answer our research questions, we used a cluster randomized longitudinal experiment containing three experimental conditions (E_{school} , E_{home} , and $E_{\text{home-school}}$) and a control condition (C):

E_{school} : Playing multiplicative mini-games *at school*, integrated in a lesson.

E_{home} : Playing multiplicative mini-games *at home*, with no attention at school.

$E_{\text{home-school}}$: Playing multiplicative mini-games *at home*, with *debriefing at school*.

C: Pseudo-intervention: playing at school mini-games on other mathematics domains, including spatial orientation, addition, and subtraction.

In all conditions, the teachers were asked to keep the total in-class lesson time that was spent on each mathematics domain the same as would have been the case had the school not been participating in the study. In this way, we could compare the regular curriculum for multiplicative reasoning (in the control group) with a multiplicative reasoning curriculum including an intervention with mini-games (in the experimental groups). The pseudo-intervention in the control group prevented the effect of the mini-games from being obscured by the positive effect that participating in an experiment may have by itself (Hawthorne effect, see Parsons, 1974; Rosas et al., 2003).

Fig. 1 shows the time schedule of the study. In both Grade 2 and Grade 3 there were two game periods, in which the mini-games were played according to one of the aforementioned conditions. To monitor students' learning of multiplicative reasoning, multiplicative ability tests were administered at three measurement

	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
Grade 1 (2009/2010)										Measurement point 1: Skills Test 1
Grade 2 (2010/2011)	Game period 1					Game period 2				Measurement point 2: Knowledge Test 2 Skills Test 2 Insight Test 2
Grade 3 (2011/2012)	Game period 3					Game period 4				Measurement point 3: Knowledge Test 3 Skills Test 3 Insight Test 3

Fig. 1. Time schedule of the study. Skills Test = test of multiplicative operation skills; Knowledge Test = test of multiplicative fact knowledge; Insight Test = test of insight in multiplicative number relations.

Table 1
Numbers of schools, classes, and students in the study.

Condition	Recruited sample		Sample that completed the study		Analysis sample	
	Schools (classes)	Students	Schools (classes)	Students	Schools (classes)	Students
C	21 (25)	519	17 (19)	356	16 (18)	327
E_{school}	15 (18)	381	8 (9)	168	6 (7)	112
E_{home}	15 (19)	394	13 (16)	284	9 (11)	202
$E_{\text{home-school}}$	15 (19)	367	9 (11)	185	4 (5)	78
Total	66 (81)	1661	47 (55)	993	35 (41)	719

Note. As some classes merged or split up in the course of the research project, the numbers of participating classes varied somewhat between grades. In the "Recruited sample" column the number of classes in Grade 1 (start of the study) is reported; in the other columns the number of classes in Grade 3 (end of the study) is reported.

points: at the end of Grade 1, at the end of Grade 2, and at the end of Grade 3.

2.2. Participants

When recruiting schools for our study, we aimed for a sample of schools that was representative of the primary schools in the Netherlands with respect to urbanization level, average level of parental education, and school size. When contacting schools by phone (response rate ca. 15%), e-mail (response rate ca. 2%), and an advertisement on a mathematics games website, we found 66 schools to be willing to participate. To evenly distribute the recruited schools over the research conditions, we used a method of blocking. Schools were matched in sets of four or five on the basis of similarity in school characteristics (urbanization level, average parental education, and school size), and random assignment was used to assign from each set of schools one school to each of the experimental conditions, and one or two schools to the control condition. Table 1 shows how the 66 schools, with 81 participating classes and 1661 students, were distributed over the four conditions.

For various reasons, such as teacher changes, organizational problems, and problems with computers, some schools dropped out in the course of the research project. There were five schools that administered the first test in Grade 1 but did not continue the project in Grade 2. Furthermore, seven schools dropped out during the Grade 2 intervention, six schools dropped out after Grade 2, and one school dropped out during the Grade 3 intervention. Thus, 47 schools stayed in the project till the end.

To measure the effects of the interventions in the different conditions as accurately as possible, we included in our analyses only those classes in which in both grades more than half of the games

were treated (see Section 2.3.3). Furthermore, we excluded one school in the E_{school} condition, because for this school the students' individual gameplay behavior could not be measured since the students played the games in dyads. Finally, we ended up with 35 schools, with 41 participating classes, which were used for the analyses. Of these classes, students who moved to another class or school during the experiment or did not complete any of the multiplicative ability tests were excluded, resulting in a sample of 719 students (see Table 1).

The initially recruited sample was found to be representative of the population of Dutch primary schools as well as of the population of Dutch primary school students with respect to the school characteristics urbanization level, average parental education, and school size, and the student characteristics gender and parental education. The analysis sample, however, differed from the population with respect to parental education. The students in this sample had parents with a higher level of education than had the students in the population (respectively 90.4% and 86.6% of the students had parents who completed at least secondary education, $\chi^2(1) = 8.84, p < .01$). Also with respect to the schools' average level of parental education, the analysis sample was not representative of the population ($t(34) = 3.88, p < .001$).¹

When checking for selective dropout, we found that the initially recruited students who were not included in the analysis sample had a significantly lower level of parental education than had the students who were included (respectively 81.9% and 90.4% of the students had parents with at least secondary education, $\chi^2(1) = 4.47, p < .05$). Moreover, not-included students more often

¹ Population data were taken from CBS (2012; student data) and CFI (2011; school data).

had a foreign home language (17.7%, as compared to 7.6% for included students, $\chi^2(1) = 4.38, p < .05$), and had, on average, a lower Grade 1 score on general mathematics ability ($M = 39.8, SD = 16.5$, as measured by the Cito mathematics test end Grade 1, Janssen, Verhelst, Engelen, & Scheltens, 2010) than had included students ($M = 45.2, SD = 14.7, t = 3.83, p < .001$). Regarding gender and age there was no difference between included and not-included students ($p > .05$). At the school level, we found that the recruited schools that were not included in the analysis sample had significantly lower average parental education than the schools that were included (averages of, respectively, 86.7% and 94.8% students with parents with at least secondary education, $t(64) = -2.70, p = .009$). Also, non-included schools had a higher level of urbanization ($M = 3.06, SD = 1.26$; on a five-point scale from non-urban (1) to highly urban (5)) than included schools ($M = 2.43, SD = 1.20, t(64) = 2.10, p = .040$). With respect to school size and mean general mathematics ability the included schools did not differ from the non-included schools ($p > .05$).

2.3. Intervention program

The intervention program included four game periods, each lasting 10 weeks (see Fig. 1). In each game period eight different mini-games were offered; every week a new game, except for the fifth and tenth week, which were meant for repeating earlier presented games.

2.3.1. The mini-games

The mini-games that were used in the experimental conditions were mostly adapted versions of multiplicative mini-games selected from the Dutch mathematics games website Rekenweb (www.rekenweb.nl, English version: www.thinklets.nl). The adaptations concerned the inclusion of a scoring mechanism and some changes in the games' difficulty level to make them fit the students' stage in the learning trajectory. Moreover, we modified some games to create more learning opportunities, for example by emphasizing connections between different multiplication problems, and relations between representation and formal notation. The games we used in the control group were existing mini-games from Rekenweb about spatial orientation, addition, and subtraction. For both the experimental groups and the control group, the games were made available online at a games website created using the Digital Mathematics Environment (DME).²

The games in the experimental conditions focused on automating multiplicative number facts and multiplicative operation skills (through practicing), and on developing insight in multiplicative number relations and properties of multiplicative operations (through exploration and experimentation). The properties embedded in the games were the principles of commutativity, distributivity, and associativity. Furthermore, the games promoted the use of derived fact strategies such as one more and one less, and doubling and halving. In addition, some games were meant to provide insight into the multiplication-related characteristics of numbers, such as factors of numbers and the divisibility of numbers. In most of the games, the mathematics content was intrinsically integrated into the main activity of the game (see Habgood & Ainsworth, 2011). In agreement with other researchers' (e.g., Leemkuil & De Jong, 2004) suggestion that support provided before playing a game may stimulate learning, we added to the games instruction videos, which lasted about 3 min each. In these videos someone plays the game while thinking aloud and thus introduces in a natural way how the game is played and which strategies can be used.

A list of the mini-games that were used in the four game periods of the experimental groups intervention program is included in Appendix A. As an example, two of the mini-games are shown in Fig. 2. In the game "Making groups" (Fig. 2a), the student had to make rectangular groups of smileys and then determine the number of smileys in the group. In this game, the student practiced solving multiplication problems (either as memorized multiplication facts or, for example, by repeated addition). Furthermore, the game could contribute to gaining insight into the relations between multiplication problems; for example, 4 rows of 5 is the same as 5 rows of 4 (commutative property), and if 5 rows of 4 is 20, then 6 rows is 4 more, resulting in 24 (derived fact strategy of one more, or distributive property). In the game "Frog" (Fig. 2b) the student was asked to come up with his or her own multiplication problem, after which the frog asked for the answer to a related multiplication problem. Also in this game, the student practiced solving multiplication problems and could potentially gain insight into the relations between multiplication problems.

2.3.2. Instructions for the teachers

Before each game period, the teachers were given a manual in which for each game it was described how it had to be treated in class. Briefly, the manuals for the different research conditions gave the following instructions:

E_{school}: The teacher introduces the new game in a whole-class lesson (20 min), using a worksheet. Afterwards, the students watch the instruction video and play the game. After all students have played the game for approximately 10 min, the game is debriefed in a class discussion (15 min), using a digital blackboard or a class computer. The manual indicates which topics should be treated in this discussion. The goal is for the class to discuss which strategies are faster or more useful in the game. After this discussion, the students play the game for another 10 min, during which they can try the strategies that have been discussed.

E_{home}: The teacher announces that there is a new game on the games website and that the students can play this game at home. They can also play the earlier presented games. Apart from this announcement, no attention is paid to the game. The teacher does not check whether the students have played the game.

E_{home-school}: At the beginning of the week the teacher announces that there is a new game on the games website and that the students can play this game at home. They can also play the earlier presented games. Furthermore, the teacher announces that the new game will be discussed in class at the end of the week. The class discussion (ca. 15 min), for which the instructions in the teacher manual are the same as in the *E_{school}* condition, focuses on what the students have discovered in the game and which strategies they find useful. Like in the *E_{home}* condition, the teacher does not check whether the students have played the game.

C: The teacher introduces a game from the control group program in a whole-class lesson (10 min), using the digital blackboard or a computer. After this, the students play the game in one or two sessions of 10 min.

In each grade, before the start of the intervention we organized a meeting to inform the teachers of the experimental groups about the intervention program. The teachers were told that there were different research conditions and that it was important to adhere to the instruction of their own condition, to make sure the different conditions could properly be compared. The control group teachers were informed through an extensive information letter sent by (e-)mail. These teachers were not told that other research conditions

² The DME has been developed by our colleague Peter Boon at the Freudenthal Institute of Utrecht University. See <http://www.fi.uu.nl/wisweb/en/>.

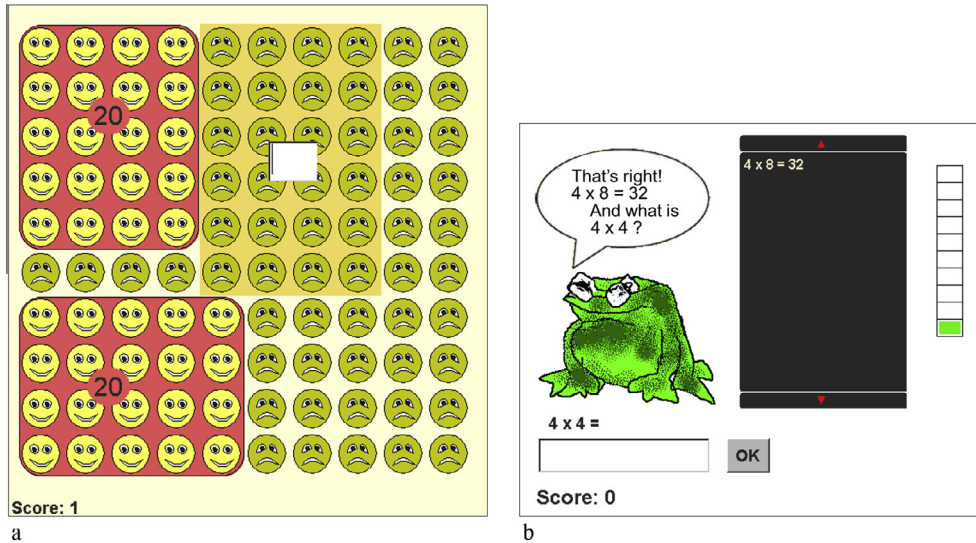


Fig. 2. Example games from the experimental groups intervention program. (a) “Making groups”. (b) “Frog”.

were included in the study, nor that the study was about multiplicative reasoning.

In all conditions a letter for the students’ parents was handed out. In the E_{home} and $E_{\text{home-school}}$ conditions, this letter explained the role of the parents in the playing at home. Parents were told not to urge their child to play the games; they should just give their child the opportunity to do so, for example by helping their child to get online. Also, it was indicated that the child needed to watch the instruction video before playing a game for the first time.

2.3.3. Intervention fidelity

To monitor the intervention fidelity we asked the teachers to keep a logbook, in which they could note each week whether they had performed the intervention as described in the teacher manual. The logbook data indicated that in several classes not all games were dealt with, due to lack of time or because the teacher had forgotten it. A similar picture arose from the automatically logged gameplay data (see Section 2.3.4). To be sure that the students had had sufficient experience with the games, we used an intervention fidelity criterion of more than half of the games having been treated (that is, played at school in E_{school} and C, debriefed at school in $E_{\text{home-school}}$, and announced in E_{home}) for deciding whether classes would be included in our analysis (see Section 2.2). We explicitly also used this intervention fidelity criterion in the control condition, in order to keep groups comparable. The decision whether a class met the fidelity criterion was primarily based on the teacher logbooks, as these provided information on teacher actions performed (e.g., debriefing sessions, announcements of new games) in addition to whether games were played. However, because of the possibility of unreliability of the logbook data (teachers may have exhibited socially desirable behavior in filling in the logbook, or may have filled in the logbook at a later time and not remembered exactly what they did), these data were verified using the logged gameplay data.³

In the analysis sample obtained using the mentioned intervention fidelity criterion, in Grade 2 on average 14.5 of the 16 games were treated. In Grade 3 this average was 14.0 (see Table 2).

³ In the case of missing logbook data (this concerned about five schools per game period), the number of games treated was estimated on the basis of the logged gameplay data combined with information obtained through communications with the school.

Table 2
Number of games treated in each condition.

Condition	Grade 2			Grade 3		
	N (classes)	M	SD	N (classes)	M	SD
C	16	14.8	2.0	17	14.2	2.4
E_{school}	6	15.0	0.9	6	15.2	1.6
E_{home}	9	14.2	2.5	10	13.5	2.0
$E_{\text{home-school}}$	5	13.8	1.8	5	13.2	1.9
Total	36	14.5	1.9	38	14.0	2.1

Note. The higher number of classes in Grade 3 is because two of the Grade 2 classes were split into two classes when transferred to Grade 3.

2.3.4. Students’ gameplay behavior

In the experimental conditions, the DME was used to log data on each student’s gameplay behavior. To indicate the extent to which the games were played, Table 3 reports per condition the time students spent on the games and the number of different games they played, in Grade 2 (game periods 1 and 2) and in Grade 3 (game periods 3 and 4). In both grades the games were played most frequently in the E_{school} condition and least frequently in the E_{home} condition, with all between-condition differences being significant ($p \leq .001$).⁴ These differences between conditions correspond to the set-up of the conditions: In E_{school} there was the most teacher guidance, in E_{home} the least. Furthermore, in all conditions the games were played more in Grade 2 than in Grade 3 ($p < .001$).⁵

The collected log data were used to compute four measures of gameplay behavior for each student: Time, Effort, Success, and NumberOfGames. The first three were computed per game as logarithmic transformations of, respectively, the time spent on the game in seconds, the number of attempted problems in the game, and the number of correct attempts. A logarithmic transformation ($f(x) = \log(x + 1)$) was employed to make the variables conform to a normal distribution and to diminish the impact of outliers. The

⁴ Because of the non-normal distribution of the playing time data, the non-parametric Mann–Whitney U test was used. Grade 2: $E_{\text{home}}-E_{\text{school}}$: $z = -12.36$, $r = -.70$; $E_{\text{home-school}}-E_{\text{school}}$: $z = -9.79$, $r = -.71$; $E_{\text{home-school}}-E_{\text{home}}$: $z = 3.26$, $r = .20$. Grade 3: $E_{\text{home}}-E_{\text{school}}$: $z = -15.08$, $r = -.85$; $E_{\text{home-school}}-E_{\text{school}}$: $z = -10.35$, $r = -.75$; $E_{\text{home-school}}-E_{\text{home}}$: $z = 4.68$, $r = .28$.

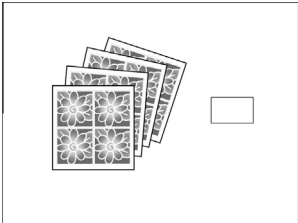
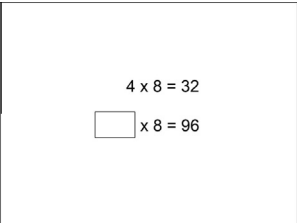
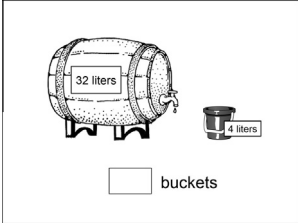
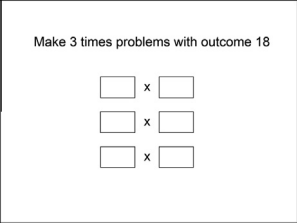
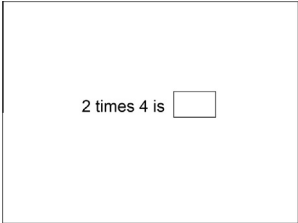
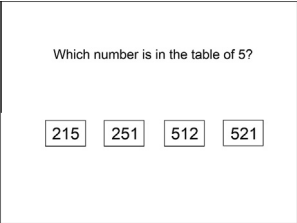
⁵ Because of the non-normal distribution of the playing time data, the non-parametric Wilcoxon signed-rank test was used. E_{school} : $z = 7.90$, $r = .75$. E_{home} : $z = 10.03$, $r = .71$. $E_{\text{home-school}}$: $z = 5.62$, $r = .64$.

Table 3
Time students spent on games (in minutes) and number of different games they played in the three experimental conditions.

Condition	Total time spent on games					Number of different games played				
	<i>M</i>	<i>SD</i>	<i>Mdn</i>	Min	Max	<i>M</i>	<i>SD</i>	<i>Mdn</i>	Min	Max
<i>Grade 2</i>										
<i>E_{school}</i>	366	84	351	187	642	15.4	1.4	16	7	16
<i>E_{home}</i>	120	228	43	0	1813	4.6	4.5	4	0	16
<i>E_{home-school}</i>	139	130	120	0	569	8.1	5.2	8	0	16
<i>Grade 3</i>										
<i>E_{school}</i>	299	97	275	98	493	14.3	1.5	15	10	16
<i>E_{home}</i>	12	35	0	0	307	1.2	2.1	0	0	10
<i>E_{home-school}</i>	60	133	0	0	860	3.2	3.9	1.5	0	16

Note. *E_{school}*: *n* = 112; *E_{home}*: *n* = 202; *E_{home-school}*: *n* = 78. Min = minimum; Max = maximum.

Table 4
Overview of the three multiplicative ability tests.

Knowledge Test	Skills Test	Insight Test
<i>What is measured?</i> Ready knowledge of multiplication number facts (declarative knowledge)	<i>What is measured?</i> Operation skills in multiplication and division (procedural knowledge)	<i>What is measured?</i> Insight in multiplicative number relations, and in properties of multiplicative operations (conceptual knowledge)
<i>Test description</i> Time-limited paper-and-pencil test with bare number multiplication problems	<i>Test description</i> Part of online test. Multiplication and division problems presented with or without a context (no time limit)	<i>Test description</i> Part of online test. Non-straightforward problems requiring explicit insight in multiplicative number relations and properties of operations
<i>Scoring</i> Number correct	<i>Scoring</i> Scale scores	<i>Scoring</i> Scale scores
<i>Sample items</i> $9 \times 2 = \dots$	<i>Sample items</i> 	<i>Sample items</i> 
	“Four sheets with four stickers. How many stickers altogether?”	“Four times eight is 32. How many times eight is 96?”
$6 \times 7 = \dots$		
	“Four liters of water go in one bucket. The barrel contains 32 liters of water. How many buckets can be filled?”	“Make three times problems with outcome 18. You are not allowed to make times problems with the number one.”
$4 \times 8 = \dots$		
	“Two times four is ...”	“Which number is in the table of five?”

transformed values were then z-standardized and, subsequently, weighted sums of the Time, Effort and Success variables were computed for each student over the intervention in Grade 2 and Grade 3 separately. The weights were based on the mean amount of time students spent on each game (averaged over students), such that games that were, on average, played more often were weighted more heavily. The fourth measure, NumberOfGames, was computed for

Grade 2 and Grade 3 as the number of different games the student played in these grades, ranging from 0 to 16. As our four measures of gameplay behavior were highly correlated for both Grade 2 and Grade 3 (correlations ranging from .76 to .96, $p < .001$), we computed summary measures of gameplay by taking the average of the four measures (after z-standardizing them). This resulted in the gameplay measures Gplay2 for Grade 2 and Gplay3 for Grade 3.

2.3.5. Multiplicative reasoning activities outside the intervention

2.3.5.1. In-class time spent on multiplicative reasoning. As we mentioned before, in all conditions, the teachers were asked to keep the total in-class lesson time that was spent on each mathematics domain the same as would have been the case if the school had not participated in the study. Thus, in the experimental conditions, the in-class parts of the mini-games intervention were scheduled as part of the time that was normally spent on the topic of multiplicative reasoning, whereas in the control group, the intervention was scheduled as part of the time normally spent on the topics of addition, subtraction, and spatial orientation. This means that the total in-class time spent on multiplicative reasoning was not influenced by the condition the school was assigned to.

To get an idea of the in-class time that was spent on multiplicative reasoning in the different conditions, we asked the teachers to fill in an online questionnaire at the end of each game period. In this questionnaire, teachers were requested to estimate the average time per week that was spent in class on different mathematics topics, including the domain of multiplicative reasoning. Averaged over the four game-periods, we found roughly similar estimates for all conditions for the in-class time spent on multiplicative reasoning (E_{school} : $M = 106$ min, $SD = 29$ min; E_{home} : $M = 119$, $SD = 31$; $E_{\text{home-school}}$: $M = 108$, $SD = 20$; C : $M = 103$, $SD = 24$; $F(3, 31) = 0.725$, $p > .10$).

2.3.5.2. Use of other educational software. Because we wanted to investigate the effects of embedding the mini-games in the real educational practice, no restrictions were placed on the contents of the multiplicative reasoning curriculum outside the mini-games intervention program. This means that teachers and students were not forbidden to work with other educational software as well, as this would also happen in normal school practice.⁶ Thus, our study investigated the effectiveness of our mini-games intervention beyond the effects of possible other educational software used. To get an indication of the total amount of educational software for the multiplicative reasoning domain that was used in the different conditions, the abovementioned teacher questionnaire also contained a question on how much in-class time, on average per week, was spent on educational software/games in different mathematics domains, including multiplicative reasoning. Based on the setup of our study, we would expect the average amount of in-class time per week spent on multiplicative reasoning software to be highest in the E_{school} condition, in which the intervention consisted of playing multiplicative mini-games at school. The teacher estimates confirmed this: in the E_{school} condition, the estimated amount of time was significantly higher than in each of the other conditions (E_{school} : $M = 21.0$ min, $SD = 4.0$ min; E_{home} : $M = 9.7$, $SD = 4.0$; $E_{\text{home-school}}$: $M = 8.8$, $SD = 6.0$; C : $M = 10.2$, $SD = 3.7$; t values ranging from 3.95 to 6.05, $p < .01$). However, also in the C condition some time was spent on educational software on multiplicative reasoning. This should be kept in mind when interpreting our results: we compare a curriculum including the mini-games intervention with a curriculum in which this intervention is not included, but which does include some working with other educational software on multiplicative reasoning.

⁶ We note that the Rekenweb games on which most of the games in the experimental groups intervention program were based, are on a freely available website. This means that students in the control group could have played some of the original Rekenweb games. To check for this, in the teacher questionnaire we asked teachers which mathematics educational software or games were used in class outside the research project. For each game-period, only zero to three of the control group teachers mentioned Rekenweb as an answer to this question. Based on these data, we can assume that playing Rekenweb games by control group students occurred infrequently. Moreover, if some students in the control group played some of the original Rekenweb games, they did this in a different way than the students in the experimental groups (i.e., using the original, non-adapted games, without the accompanying lessons/discussions, and without instruction videos).

2.4. Measurement instruments

In the current study, three dependent measures were used to assess the students' learning of multiplicative reasoning (see Table 4 for an overview): the Knowledge Test, measuring students' knowledge of multiplication number facts (declarative knowledge); the Skills Test, measuring students' multiplicative operation skills (procedural knowledge); and the Insight Test, measuring students' insight in, or understanding of, multiplicative number relations (conceptual knowledge). These tests were administered both at the end of Grade 2 and at the end of Grade 3, while the Skills Test was also administered as a pretest at the end of Grade 1 (see Fig. 1).

2.4.1. Knowledge Test

To measure students' ready knowledge of multiplicative number facts, we used the multiplication subtest of the TempoTest Automatiseren (De Vos, 2010), which we refer to as the Knowledge Test. To conceal from the teachers and students in the control group the study's focus on multiplicative reasoning, we also administered the addition and subtraction subtests of this test, but these were not used in our analyses. The multiplication subtest is a time-limited paper-and-pencil test, consisting of a sheet of 50 bare number problems, with the operation being represented by the \times symbol. Students get 2 min time to solve as many of the problems as possible; the score is the number of correct answers. The test has a split-half reliability of .96 (De Vos, 2010). Because in the Netherlands the \times symbol is commonly not introduced yet in Grade 1, the automaticity test was only administered in Grade 2 (Knowledge Test 2) and Grade 3 (Knowledge Test 3).

2.4.2. Skills and Insight Tests

The Skills and Insight Tests were administered together as an online test. In order to match the development of the students, at each measurement point a different online test was administered (Test 1, Test 2, and Test 3), of which the first test consisted only of a Skills Test. To be able to put the test scores at the different measurement points on a common scale, the tests were linked through anchor items. Each test was piloted at two schools that did not participate in the study.

2.4.2.1. Composition of the tests. The Skills Tests contained straightforward multiplicative problems, including both bare number problems and problems presented in a context. The Insight Tests consisted of problems in which students had to use their knowledge of multiplication and division at a higher comprehension level. These non-straightforward problems required explicit insight in multiplicative relations between numbers (e.g., factors of numbers) and the properties of multiplicative operations (e.g., the commutative and distributive property). For example, problems were included which were actually beyond the mathematics content taught to the students so far, which means that the students could only solve them by making use of their understanding of multiplicative number relations. Example items from the Skills Tests and the Insight Tests are presented in Table 4.

Besides the multiplicative problems of the Skills and Insight Tests, the online tests also contained some "distractor" items on spatial orientation, addition, and subtraction. These items, which were not used in our analyses, were meant to conceal from the students and teachers in the control group that the focus of the study was on multiplicative reasoning. Table 5 shows the numbers of items of different types in each test. Also the numbers of anchor items are given.

To control for order effects, for each measurement point four different versions of the online test were constructed. For this purpose the items of each test were organized into clusters. Test

Table 5
Numbers of (anchor) items in the online tests at the three measurement points (Test 1, Test 2, and Test 3), and reliability estimates.

Item type	Number of (anchor) items			WLE-reliability		
	Test 1	Test 2 ^a	Test 3 ^b	Test 1 (n = 689)	Test 2 (n = 665)	Test 3 (n = 694)
Skills items	28	29 (16)	31 (8, 12)	.84	.69	.71
Insight items	–	21 (0)	25 (0, 12)	–	.76	.77
Distractor items	12	16	16			
Total	40	66	72			

^a Between parentheses is the number of items in Test 2 that were also in Test 1.

^b Between parentheses are, respectively, the number of items in Test 3 that were also in Test 1 and Test 2, and the number of items in Test 3 that were also in Test 2 but not in Test 1.

Table 6
Initial differences in student characteristics between conditions.

Condition	n	Gender % female	Age % not delayed	Parental education % secondary	Home language % Dutch	GMath score M (SD)	Skills Test 1 score M (SD)
C	327	50.2	93.9	93.0	94.5	46.8 (14.7)	0.09 (1.36)
E_{school}	112	41.1	88.4	86.6	85.7	45.0 (16.6)	0.09 (1.33)
E_{home}	202	50.5	88.1	87.1	93.6	42.9 (15.0)	–0.12 (1.14)
$E_{\text{home-school}}$	78	35.9	80.8	93.6	89.7	43.5 (12.8)	–0.17 (1.47)
Total	719	47.3	90.0	90.4	92.4	45.1 (15.1)	0.00 (1.31)
Wald $\chi^2(3)$		10.15*	6.54†	3.08	1.45	1.03	1.44
η^2		.011	.019	.011	.014	.018	.007

Note. Wald tests (comparable to one-way ANOVAs) were performed in Mplus, using cluster-robust standard errors (see Section 2.7). The η^2 effect sizes were calculated based on regular ANOVA results.

† $p < .10$.

* $p < .05$.

1 contained four clusters of 10 items each, which were presented in different orders in the different test versions. For Test 2, to be able to assess a larger variety of items, including insight items, we used six clusters of 11 items each. Each version of Test 2 contained four of these clusters. With this design, we could later compute the total score over all 29 items of Skills Test 2 and all 21 items of Insight Test 2, using a Rasch model (see below). The same approach was used for Test 3, in which we used six clusters of 12 items each, with four clusters per test version. The different test versions were randomly assigned to the students.

2.4.2.2. Test procedure. The online tests were administered through the previously mentioned Digital Mathematics Environment (DME). Online test administration facilitated our large-scale data collection and ensured a relatively formal, standardized test setting. Each test item was individually displayed on the screen, and the accompanying question was read aloud by the computer. The tests were administered at school, facilitated by the class teacher. The duration of each test was, on average, approximately 20–30 min.

2.4.2.3. Correction of input errors. Since the text boxes in which the students had to type their answers accepted all kinds of input, not all responses were in the form of a number. Input errors for which it was clear which number was meant, such as “40” or “4o” instead of “40”, or “vier” (Dutch for “four”) instead of “4”, were corrected. For Test 1, this resulted in a change to a correct answer for 0.60% of the item answers; for Test 2 and Test 3, this was the case for 0.08% and 0.05% of the item answers, respectively.

2.4.2.4. Scaling of test scores. Because different tests were administered at the different measurement points, and because at measurement points 2 and 3 the different versions of the tests contained different subsets of the total set of Skills and Insight items, item response modeling was needed to put the Skills Test and Insight Test scores of the different measurement points and

different test versions on a common scale. For the Skills Tests, the items of Skills Test 1, Skills Test 2, and Skills Test 3 were first separately scaled by a Rasch model, employing the Conquest software (Wu, Adams, Wilson, & Haldane, 2007). Using this procedure, the students' raw test scores were converted into scale scores (weighted likelihood estimates, or WLE) for each test. Subsequently, to put all three Skills Test scores on a common scale, we employed mean–mean linking (Kolen & Brennan, 2004), with the assumption that (for equal student ability) the item difficulties of the anchor items were equal on average in the different tests. The same procedure was employed for the two Insight Tests.

2.4.2.5. Reliability. Table 5 presents the WLE reliability estimates (Wu et al., 2007) of the scale scores of the Skills Tests and Insight Tests, which can be interpreted in the same way as Cronbach's alpha. The tests can be considered adequately reliable, although the reliability of Skills Test 2 is just below the .70 boundary. Remaining unreliability was accounted for in our analyses (see Section 2.7).

2.5. Treatment of missing data

As is inevitable in a large-scale longitudinal study carried out in real school practice, not all data were available for all students. The percentage of missing scores ranged from 2.0% to 8.1% per test. To make estimates for the missing test scores, we employed multiple data imputation (see Graham, 2009). We specified an imputation model involving student background data, test scores, and, for the students in the experimental conditions, the gameplay data. Because the gameplay data can be expected to have a different relation with the learning outcomes in the different conditions, the imputation procedure was performed for each condition separately. To account for the clustered data structure (students nested within schools), we also included school mean test scores as predictors in the imputation model. The data imputation was run using the “mice” software (Van Buuren & Groothuis-Oudshoorn, 2011), and resulted in 50

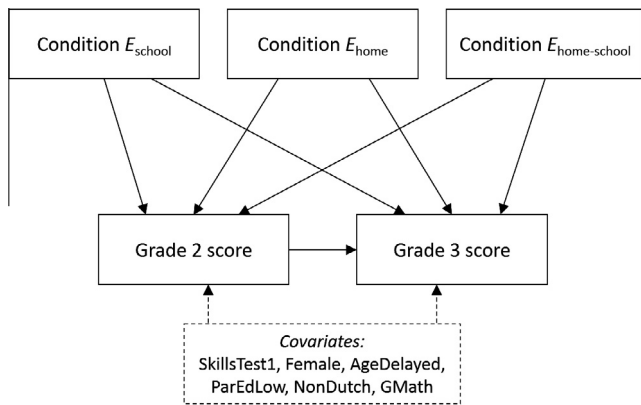


Fig. 3. Path model used for comparing the three experimental conditions to the control group (reference category).

imputed datasets. Statistical analyses were performed on these 50 datasets and results were combined using Rubin's rule (see Graham, 2009).

2.6. Initial differences between conditions in student characteristics

Although we employed blocking and random assignment to distribute the participating schools over the four research conditions, differences between groups may have arisen with respect to their student composition. Therefore, after data-imputation, we examined whether there were differences between the conditions with respect to students' gender, age (students with a grade-appropriate age vs. older, delayed students), parental education (higher vs. lower education⁷), and home language (monolingual Dutch vs. other), and their Grade 1 scores on general mathematics ability (GMath; measured by the Cito mathematics test end Grade 1, Janssen et al., 2010) and multiplicative reasoning ability (Skills Test 1). As is shown in Table 6, we found a significant difference between conditions for gender and a marginally significant difference ($p = .088$) for age. In addition to gender and age, effect sizes were non-trivial ($\eta^2 \geq .01$) for parental education, home language, and GMath score. To be conservative, we decided in all analyses to control for gender (dummy variable Female), age (dummy variable AgeDelayed), parental education (dummy variable ParEdLow), home language (dummy variable NonDutch), and GMath score.

2.7. Data analysis

2.7.1. Path analysis

We analyzed our data using path analysis in Mplus (Muthén & Muthén, 1998–2010). Path analysis was employed to be able to simultaneously study the effects of the intervention in Grade 2 and the intervention in Grade 3, as well as their combined effect.⁸

The path model we used in answering research questions 1–3 is displayed in Fig. 3. This model can be interpreted as testing two ANCOVAs simultaneously, one with the Grade 2 score as the dependent variable, and one with the Grade 3 score as the dependent variable. As is shown in Fig. 3, as predictors we used three dummy variables for the three experimental conditions (the

control condition was modeled as the reference category). As covariates we used the pretest score (Skills Test 1) and the covariates related to initial differences between conditions (Female, AgeDelayed, ParEdLow, NonDutch, and GMath). The model was separately specified for the three aspects of multiplicative reasoning ability – knowledge, skills, and insight. In addition, we specified a joint model in which the standardized paths for the three types of multiplicative ability tests were constrained equal, to test the effect of the mini-games interventions averaged over the three aspects of multiplicative reasoning ability. This joint model can be seen as testing the effect of the games on students' overall multiplicative reasoning ability. For answering Research question 4, the model in Fig. 3 was extended, as we will explain later.

In the model, an arrow from a condition variable to Grade 2 score represents the direct effect of the particular condition on the Grade 2 score, that is, the effect of the Grade 2 intervention in that condition on the score at the end of Grade 2. Similarly, an arrow from a condition variable to Grade 3 score represents the direct effect of the Grade 3 intervention in this condition on the Grade 3 score. In addition to these direct effects, we also examined the total effect of the interventions in Grade 2 and Grade 3 on Grade 3 score (computed as the sum of the direct and indirect effect of condition on Grade 3 score). This total effect can be seen as the effect of the combined Grade 2–3 intervention.^{9,10}

To control for unreliability in the Skills and Insight Test scores, these scores were modeled as latent variables, with their residual variance fixed at $[1 - \text{reliability of test scores}] * \text{variance of test scores}$ (see Hayduk, 1987). Dependent of whether our hypotheses were directional or not, we use one-tailed or two-tailed significance tests. When multiple equalities were tested at once (e.g., in a Wald test), two-tailed tests were used.

2.7.2. Dealing with clustered data

To account for the clustered data structure (students nested within classes/schools), we employed cluster-robust standard errors (see Angrist & Pischke, 2009) in our analyses, using the TYPE = COMPLEX option in Mplus. As the level of clustering we used the school, because random assignment to conditions was done at the school level and because participating classes within schools were sometimes merged or split up in the course of the research project. We do not expect large differences compared to an approach using class as the level of clustering, as in 80% of the schools only one class was participating.

2.7.3. Path coefficients and effect size measures

For all analyses, we report standardized or partially standardized path coefficients. For dummy (binary) predictors (the condition variables), we employed partially standardized coefficients β_{ps} , which represent the amount of change in standard deviation units in the dependent variable associated with a change in the dummy predictor from 0 to 1 (STDY in Mplus). β_{ps} is thus practically equivalent to a d effect size of the difference between the 0 and 1 category (interpretation guidelines: 0.20 (small effect), 0.50 (medium effect), 0.80 (large effect), see Cohen, 1988). For continuous predictors (the gameplay variables), the standardized coefficient β represents the amount of change in standard deviation units in the dependent variable associated with a one standard

⁷ Higher means that at least one parent has completed secondary education; lower means that none of the parents has completed secondary education.

⁸ As we had dependent variables at two time points (end Grade 2 and end Grade 3), a regular ANCOVA approach was not appropriate. Repeated measures ANOVA would have been a possibility, but we decided not to use such an analysis, as we had three measurement points for the Skills Test, but only two measurement points for the Knowledge Test and the Insight Test. Using repeated measures ANOVA, then, would have led to different, incomparable analyses for the different tests of multiplicative reasoning ability.

⁹ Total effects were tested using the Delta method as implemented in Mplus (Muthén & Muthén, 1998–2010). Although we are aware that this method can be rather conservative, other methods like Bootstrap cannot easily be adapted for clustered samples.

¹⁰ In accordance with the usual terminology for path analysis, the parameters of these models are called effects. This does not necessarily imply that these parameters can be interpreted as causal effects. Whether effects are causal depends on the research design: in our case, the effects of the condition variables can be seen as causal effects (conditions were experimentally manipulated).

deviation change in the predictor variable (STDYX in Mplus). This coefficient is practically equivalent to an *r* effect size, for which the values .10, .30, and .50 can be interpreted as a small, medium, and large effect, respectively (Cohen, 1988). For completeness, we also provide regular *d* or *r* values, in cases where these values differ (slightly) from the β_{ps} and β values, respectively.¹¹

2.7.4. Dealing with selective dropout

To account for the selective dropout (see Section 2.2), we also ran the path analyses on the effects of the games (research questions 1 and 2) using the method of inverse probability weighting, in which students that are underrepresented in the analysis sample are weighted more heavily (see Seaman & White, 2013). The results using this weighting did not deviate much from the results without weighting (see Footnotes 14 and 15), indicating that the selective dropout did not have a substantial influence on our results. For simplicity, then, we only report the non-weighted results.

3. Results and discussion

Before describing our path analysis results, to give the reader a first view of the outcomes we first performed regular ANCOVAs to compare the four conditions on the three aspects of multiplicative ability in Grade 2 as well as Grade 3, and pairwise comparisons comparing each of the experimental conditions to the control condition. As covariates we used pretest scores and the abovementioned covariates related to initial differences between conditions (see Section 2.6). ANCOVAs were modeled as multiple regression models in Mplus; the pairwise comparisons were assessed by testing the significance of the regression coefficients of the condition dummy variables (*z* statistics in Mplus), which reflect the contrasts between the experimental conditions and the control condition (reference category). ANCOVA *F* values comparing the four conditions were not significant ($p > .10$), but some of the pairwise comparisons were.¹² Table 7 displays the means and standard deviations of the scores on the multiplicative ability tests in the four conditions, together with the pairwise comparison results.

The pairwise comparisons in Table 7 are given just as a background for the reader. We base the answers to our research questions on the results of the abovementioned path models, because these models allowed us to test ANCOVAs for Grade 2 and Grade 3 simultaneously, together with the combined effect of the Grade 2–3 intervention. The reader will notice that results for the direct effects of the Grade 2 and the Grade 3 intervention are similar to the pairwise comparisons results.

3.1. Effects of the interventions on overall multiplicative reasoning ability

To investigate whether the interventions with multiplicative mini-games positively affected students' overall multiplicative

¹¹ The *d* values were calculated by dividing the raw coefficients by the pooled standard deviation of the dependent variable; the *r* values were calculated by multiplying the raw coefficient by the standard deviation of the predictor, and then dividing by the pooled standard deviation of the dependent variable. For Skills Test scores and Insight Test scores, in accordance with the abovementioned reliability correction (Hayduk, 1987), we used a standard deviation adjusted for unreliability: $SD = \sqrt{[variance * reliability]}$. Note: in cases where Mplus did not provide (partially) standardized coefficients (for total effects and for paired comparisons between coefficients), we computed them using the same formulas as we used for computing *d* and *r* values. In these cases, thus, β_{ps} and β values are by definition equal to *d* and *r* values, respectively. Also for analyses using the joint model, β_{ps} and β values are by definition equal to *d* and *r* values.

¹² As our research questions focused on the effect of each experimental condition as compared to the control group (orthogonal planned comparisons), rather than the general difference between the four conditions, it was justified to perform pairwise comparisons in the absence of significant *F* values (see Keppel & Wickens, 2004).

Table 7 Descriptives of all multiplicative ability test scores, and pairwise comparisons between the experimental groups and the control group.

Condition	n	Grade 1		Grade 2		Grade 3												
		Skills Test 1		Knowledge Test 2		Skills Test 2		Insight Test 2		Knowledge Test 3		Skills Test 3		Insight Test 3				
		M (SD)	d	M (SD)	d	M (SD)	d	M (SD)	d	M (SD)	d	M (SD)	d	M (SD)	d			
C	327	0.09 (1.36)		20.50 (9.66)		2.40 (1.46)		31.52 (12.08)		3.50 (1.32)		31.52 (12.08)		3.50 (1.32)		1.52 (1.38)		
E _{school}	112	0.09 (1.33)	0.05	20.15 (8.56)	0.05	2.41 (1.31)	0.43	0.08	0.31 (1.44)	1.74*	0.31	28.60 (11.93)	-1.20	-0.20	0.47	0.05	1.60 (1.33)	0.11
E _{home}	202	-0.12 (1.14)	-0.66	18.22 (7.55)	-0.66	2.12 (1.38)	-0.19	-0.03	-0.05 (1.51)	1.36*	0.16	29.78 (12.02)	0.45	0.05	0.33	0.03	1.24 (1.50)	-0.80
E _{home-school}	78	-0.17 (1.47)	0.30	20.45 (7.75)	0.30	2.38 (1.30)	1.06	0.15	0.12 (1.37)	1.68*	0.26	32.28 (10.48)	0.86	0.11	1.54†	0.16	1.53 (1.28)	1.20
Total	719	0.00 (1.31)		19.80 (8.80)		2.32 (1.40)		0.03 (1.48)		0.03 (1.48)		30.66 (11.94)		3.45 (1.36)		1.45 (1.40)		

Note. The scores on the Skills Tests and Insight Tests are scale scores; the scores on the Knowledge Tests are number correct scores (see Section 2.4). The *z* and *d* values are based on the regression coefficients contrasting the experimental conditions to the control group (C), controlling for pretest scores (for Grade 3 also Grade 2 scores were controlled for) and the covariates Female, AgeDelayed, ParEdLow, NonDutch, and GMath.
 † $p < .10$. One-tailed.
 * $p < .05$. One-tailed.

Table 8
Effects of conditions E_{school} , E_{home} , and $E_{\text{home-school}}$ on overall multiplicative reasoning ability and on the three aspects of multiplicative reasoning ability (as compared to the control group).

Effect	Overall multiplicative reasoning ability ^a		Aspect of multiplicative reasoning ability ^b								
			Knowledge			Skills			Insight		
	β_{ps} (=d)	SE	β_{ps}	SE	d	β_{ps}	SE	d	β_{ps}	SE	d
<i>Condition E_{school}</i>											
Effect of Grade 2 intervention	0.18	0.19	0.01	0.24	0.01	0.10	0.24	0.09	0.39 [†]	0.22	0.35
Effect of Grade 3 intervention	-0.10	0.11	-0.20	0.17	-0.20	0.02	0.15	0.02	-0.07	0.14	-0.06
Total effect of Grade 2–3 intervention	0.01	0.18	-0.20	0.23	-0.20	0.09	0.16	0.09	0.13	0.17	0.13
<i>Condition E_{home}</i>											
Effect of Grade 2 intervention	0.00	0.15	-0.16	0.23	-0.16	-0.04	0.20	-0.03	0.21 [†]	0.15	0.19
Effect of Grade 3 intervention	0.01	0.08	0.05	0.11	0.05	0.06	0.15	0.05	-0.14	0.13	-0.12
Total effect of Grade 2–3 intervention	0.01	0.10	-0.05	0.16	-0.05	0.03	0.13	0.03	-0.02	0.11	-0.02
<i>Condition $E_{\text{home-school}}$</i>											
Effect of Grade 2 intervention	0.19	0.17	0.08	0.26	0.08	0.21	0.20	0.18	0.32 [*]	0.19	0.29
Effect of Grade 3 intervention	0.11	0.09	0.11	0.13	0.11	0.13	0.16	0.12	0.07	0.13	0.06
Total effect of Grade 2–3 intervention	0.22 ^{**}	0.09	0.16	0.13	0.16	0.26 [*]	0.15	0.26	0.22 [*]	0.11	0.22

Note. $N = 719$. SkillsTest1, Female, AgeDelayed, ParEdLow, NonDutch, and GMath were included as covariates (see Fig. 3). β_{ps} = partially standardized coefficient.

^a These are the effects averaged over the three aspects of multiplicative reasoning ability (standardized paths of the path models of the three aspects constrained to be equal).

^b The model was separately specified for each of the three aspects of multiplicative reasoning ability.

[†] $p < .10$. One-tailed.

^{*} $p < .05$. One-tailed.

^{**} $p < .01$. One-tailed.

reasoning ability (Research question 1), we employed the joint path model as mentioned in Section 2.7. The direct and total effects in this joint model are presented in Table 8 (first columns).¹³ We found a significant total effect of the $E_{\text{home-school}}$ condition ($d = 0.22$), as compared to the control group (the reference category). Thus, the combined Grade 2–3 intervention where mini-games were played at home and debriefed at school was effective in enhancing students' overall multiplicative reasoning ability. This finding contributes to the still relatively sparse knowledge on the educational effectiveness of (mathematics) computer games (e.g., Bai et al., 2012; Wouters et al., 2013). Contrary to our expectations, for the E_{school} and the E_{home} condition we did not find significant effects on overall multiplicative reasoning ability ($p > .10$). Apparently, the combination of extra time-on-task and debriefing, both included in $E_{\text{home-school}}$, was needed for the mini-games to actually have added value as compared to the regular program for multiplicative reasoning without these games.¹⁴

3.2. Effects of the interventions on the three aspects of multiplicative reasoning ability

To investigate the effects on the three different aspects of multiplicative reasoning ability (Research question 2), we used the path model displayed in Fig. 3, specified for each aspect separately: number fact *knowledge*, operation *skills*, and *insight* in multiplicative number relations. The model results are displayed in Table 8 (note that the effects of the separate Grade 2 and Grade 3 interventions are in line with the pairwise comparison results in Table 7). As is shown, we found effects of the games in enhancing skills and insight, but not knowledge. In particular, in the $E_{\text{home-school}}$ condition the games affected both skills and insight, and both the Grade 2 intervention and the combined Grade 2–3 intervention were effective. Significant effects were the effect of the $E_{\text{home-school}}$ intervention in Grade 2 on insight ($d = 0.29$), and the total effect of the

$E_{\text{home-school}}$ intervention in Grade 2–3 on both skills ($d = 0.26$) and insight ($d = 0.22$). In the E_{school} condition the games only affected insight, and only the Grade 2 intervention was effective ($d = 0.35$). No significant effects were found in the E_{home} condition ($p > .05$).¹⁵

The significant effects we found for the $E_{\text{home-school}}$ and E_{school} interventions are of small size (ds from 0.22 to 0.35). Yet, the effect sizes are similar to the average effect size found in Wouters et al.'s (2013) meta-analysis ($d = 0.29$). Though small, the effects we found can be viewed as reasonable given the fact that the mini-games interventions were compared to the regular educational program for multiplicative reasoning (instead of comparing to a control group in which there is no instruction on the particular educational topic involved), and were carried out in the real educational practice (as compared to a more controlled research setting).

As is clear from the above results, effects of the games were primarily found for the Grade 2 intervention (in $E_{\text{home-school}}$ effects were also found for the combined Grade 2–3 intervention, but we did not find effects of the Grade 3 intervention alone). This may be explained by the stage of the students' learning process. In Grade 2, students are at the beginning of learning multiplicative reasoning, which may imply that there is more room for improvement than in Grade 3. Another possible explanation is the occurrence of a novelty effect (e.g., Li & Ma, 2010). Students, as well as teachers, may be more motivated to put attention into the games when they are new for them. This explanation is supported by our finding that the games were played more in Grade 2 than in Grade 3.

The above results suggest that the games were most effective in enhancing students' multiplicative insight. This is in line with the finding from Ke's (2009) review that games seem to promote higher-order thinking more than factual knowledge acquisition. Computer games may be especially useful for the teaching of insight when they allow for free exploration and experimentation, as was the case in our study.

¹³ Correlations between all variables in the path models are available as an online supplement to this paper.

¹⁴ The analysis with inverse probability weights (see Section 2.7.4) led to similar results: a significant total effect was found for $E_{\text{home-school}}$ ($\beta_{\text{ps}} = 0.24$, $p < .05$, $d = 0.24$).

¹⁵ When employing inverse probability weights (see Section 2.7.4), similar results were found (direct effect $E_{\text{home-school}}$ Grade 2 on insight: $\beta_{\text{ps}} = 0.45$, $p < .01$, $d = 0.41$; total effect $E_{\text{home-school}}$ on skills: $\beta_{\text{ps}} = 0.20$, $p < .10$, $d = 0.20$; total effect $E_{\text{home-school}}$ on insight: $\beta_{\text{ps}} = 0.24$, $p < .05$, $d = 0.24$; direct effect E_{school} Grade 2 on insight: $\beta_{\text{ps}} = 0.35$, $p < .10$, $d = 0.31$).

Table 9Paired comparisons between the E_{school} , E_{home} , and $E_{\text{home-school}}$ condition of effects on overall multiplicative reasoning ability.^a

Effect	Comparison					
	$E_{\text{school}}-E_{\text{home}}$		$E_{\text{home-school}}-E_{\text{home}}$		$E_{\text{home-school}}-E_{\text{school}}$	
	$\Delta\beta_{\text{ps}} (=d)$	SE	$\Delta\beta_{\text{ps}} (=d)$	SE	$\Delta\beta_{\text{ps}} (=d)$	SE
Effect of Grade 2 intervention	0.18	0.17	0.19	0.15	0.01	0.19
Effect of Grade 3 intervention	-0.11	0.11	0.10	0.09	0.21*	0.11
Total effect of Grade 2–3 intervention	0.00	0.17	0.21**	0.09	0.22	0.17

Note. E_{school} : $n = 112$; E_{home} : $n = 202$; $E_{\text{home-school}}$: $n = 78$. SkillsTest1, Female, AgeDelayed, ParEdLow, NonDutch, and GMath were included as covariates (see Fig. 3). Two-tailed significance tests were used for the $E_{\text{school}}-E_{\text{home}}$ comparison; one-tailed tests were used for the $E_{\text{home-school}}-E_{\text{home}}$ and the $E_{\text{home-school}}-E_{\text{school}}$ comparison (because of our directional hypothesis). $\Delta\beta_{\text{ps}}$ = partially standardized difference between coefficients.

^a Averaged over the three aspects of multiplicative reasoning ability (standardized paths of the path models of the three aspects constrained to be equal).

* $p < .05$. One-tailed.

** $p < .01$. One-tailed.

3.3. Comparisons between the three game-playing settings

To statistically test the difference between the three experimental interventions in their effectiveness as compared to the control group (Research question 3), we compared the path coefficients of the three condition variables in the joint model of overall multiplicative reasoning ability (effects averaged over the three aspects of multiplicative ability). We used Wald χ^2 tests (comparable to one-way ANOVAs), and pairwise comparisons between the path coefficients. None of the Wald test results were significant ($p > .10$), but some of the paired comparisons were, as is shown in Table 9. We found a significant difference between $E_{\text{home-school}}$ and E_{school} , in favor of $E_{\text{home-school}}$, for the effect of the Grade 3 intervention ($d = 0.21$). Furthermore, we found that the total effect of the Grade 2–3 intervention was significantly higher in $E_{\text{home-school}}$ than in E_{home} ($d = 0.21$). Also when we looked at the three aspects of multiplicative ability separately (see Appendix B), we found several significant differences indicating that the $E_{\text{home-school}}$ intervention was more effective than the E_{school} and the E_{home} intervention, while there were no differences between the E_{school} and the E_{home} intervention.

The higher effectiveness of the $E_{\text{home-school}}$ intervention was as expected and can be explained by this condition having the advantage of playing at home (extended learning time, more learner control) as well as the advantage of an in-class intervention (the debriefing sessions). Another explanation of the higher effectiveness of the $E_{\text{home-school}}$ intervention as compared to the E_{home} intervention may lie in the amount of time spent on the mini-games, which was higher in $E_{\text{home-school}}$ than in E_{home} . This means that, apart from having a reflective role, the debriefing sessions in the $E_{\text{home-school}}$ condition may also have functioned as an encouragement for the students to play the games at home.

3.4. Relations between gameplay behavior and learning outcomes

To examine the relations between students' gameplay behavior (time and effort spent on the games) and the effects of the interventions (Research question 4), we added to the path model in Fig. 3 interactions between the condition variables and the variables Gplay2 and Gplay3 (as defined in Section 2.3.4), as predictors of Grade 2 score and Grade 3 score, respectively.¹⁶ In this way we could measure for each of the experimental conditions the influence of gameplay behavior on the learning effects of the mini-games. The paths from the Gplay * Condition interactions to the test scores for the three aspects of multiplicative reasoning ability are presented in Table 10.¹⁷

Regarding knowledge and skills, we found that in Grade 2, gameplay in E_{school} was significantly related to learning outcomes

in multiplicative fact knowledge ($r = .25$), and in $E_{\text{home-school}}$ to learning outcomes in multiplicative skills ($r = .21$). These findings are as expected and are in line with Jansen et al.'s (2013) finding that more practice leads to higher automatization. However, it is unclear why the other relations between gameplay behavior and learning effects on number fact knowledge and skills, for example in Grade 3, were not significant.

Regarding multiplicative insight, we did not necessarily expect a relation between gameplay behavior and learning effects, as once the learning concepts in a game are discovered, more gameplay may not lead to more learning (Kolovou et al., 2013). Yet, in contrast with the finding of Kolovou et al. (2013), we did find a positive influence of gameplay behavior on the learning effect on multiplicative insight in the E_{school} condition in Grade 2 ($r = .17$), and in the $E_{\text{home-school}}$ condition in Grade 3 ($r = .12$). This may be explained by the fact that in our study more gameplay not only meant that more time and effort was spent on particular games, but also that more different games were played, which differs from Kolovou et al.'s study, in which there was only one mini-game. As in different games different concepts were embedded, playing more different games may have led to more development of insight.

In the E_{home} condition we did not find significant relations between gameplay and learning effects. Possibly, playing the games alone was not sufficient for promoting learning, but reflection, for example in the form of a debriefing session, was necessary. This corresponds to the importance of debriefing as indicated by several researchers (e.g., Egenfeldt-Nielsen, 2005; Garris et al., 2002; Klawe, 1998).

3.5. Generalizability issues and limitations of our study

It should be noted that our findings apply only to the use of multiplicative mini-games in Grade 2 and Grade 3 of primary school, in three specific instructional settings. Results can, in principle, not be generalized to other grade levels, other mathematics domains, other instructional settings, other games, or other countries. Another issue regarding generalizability is the fact that our analysis sample was not fully representative of the Dutch population of primary schools and students. The selective dropout of schools and students (as mentioned in Section 2.2) caused the analysis sample to contain students with more favorable characteristics (higher average level of parental education and higher average mathematics ability) than was the case for the representative sample that was initially recruited. This means that our results can, essentially, only be generalized to (schools with) students with similarly favorable characteristics.

In addition to the above generalizability issues, some further limitations of our study should be noted. Most of these limitations are a natural consequence of the fact that we performed a large-scale experiment in the real school practice. First of all, as is common in such an experiment, in our study the interventions were conducted

¹⁶ The possible influence of Gplay2 on Grade 3 test scores was controlled for in the model.

¹⁷ Correlations between all variables in the model are available as an online supplement to this paper.

Table 10
Interactions of gameplay with condition variables predicting Grade 2 and Grade 3 test scores on the three aspects of multiplicative reasoning ability.

Interaction	Aspect of multiplicative reasoning ability ^a								
	Knowledge			Skills			Insight		
	β	SE	<i>r</i>	β	SE	<i>r</i>	β	SE	<i>r</i>
<i>Condition E_{school}</i>									
Gplay2 * <i>E_{school}</i> → Grade 2 score	.39*	.20	.25	.20	.18	.10	.35*	.22	.17
Gplay3 * <i>E_{school}</i> → Grade 3 score	.03	.14	.02	.07	.10	.04	-.06	.16	-.03
<i>Condition E_{home}</i>									
Gplay2 * <i>E_{home}</i> → Grade 2 score	.01	.07	.01	.09†	.05	.10	.07†	.05	.08
Gplay3 * <i>E_{home}</i> → Grade 3 score	.02	.05	.01	-.06	.08	-.03	-.06	.07	-.03
<i>Condition E_{home-school}</i>									
Gplay2 * <i>E_{home-school}</i> → Grade 2 score	.03	.04	.09	.09***	.03	.21	.01	.02	.02
Gplay3 * <i>E_{home-school}</i> → Grade 3 score	.05†	.04	.15	.01	.05	.02	.06*	.03	.12

Note. *N* = 719. SkillsTest1, Female, AgeDelayed, ParEdLow, NonDutch, and GMath were included as covariates (see Fig. 3). Because of the large differences in gameplay behavior between conditions, *r* effect sizes were computed using per-condition standard deviations.

^a The model was separately specified for each of the three aspects of multiplicative reasoning ability.

† *p* < .10. One-tailed.

* *p* < .05. One-tailed.

*** *p* < .001. One-tailed.

by the regular class teachers. The teachers might have interpreted our instructions in their own way, as is generally the case when teachers use instructional materials. Although the teacher logbooks and gameplay log data informed us on how many of the games were treated by the teachers, and despite the fact that we took several measures to prevent the intervention from being implemented other than intended (e.g. providing precise guidelines and organizing information meetings), we cannot be sure about the actual in-class activities that have contributed to the effectiveness of the games. In fact, the micro-level of instruction needs further research.

Another issue related to doing research in the school practice is that, beyond the mini-games in our intervention, other educational software or games could have been used. As we wanted to examine the effects of our mini-games in an educational situation as realistic as possible, we did not forbid teachers or students to work with other educational software (as explained in Section 2.3.5). This means that our results should be interpreted as the effects of implementing the mini-games as part of the regular curriculum for multiplication and division, as compared to such curriculum without these mini-games but with possibly some other educational software related to multiplicative reasoning.

A further point is that the intervention that was found to be most effective – playing the games at home with debriefing at school (*E_{home-school}*) – seemed to be hard to maintain for some of the participating teachers (several of the *E_{home-school}* classes did not meet the intervention fidelity criterion). Possibly, this had something to do with decreasing enthusiasm of the students for playing the games at home, for example, due to a decreasing novelty effect. Teachers may have skipped debriefing sessions when they noticed that only a few students had played the games. This means that, possibly, the effect we found for this intervention primarily counts for classes in which students are sufficiently motivated to keep playing the games at home, or, alternatively, for classes in which teachers are willing to hold debriefing sessions regardless of whether students have played the games. The specific requirements for successfully implementing an intervention including playing at home with debriefing at school should be further investigated.

Finally, although the interventions employed in this study were based on several theoretical notions (e.g., experiential learning, different aspects of multiplicative reasoning ability, immediate feedback), the study did not intend to prove these theories. Finding support for effects of some of the interventions on some of the aspects of multiplicative ability suggests the relevance of these

theoretical notions. Yet, we cannot be sure which exact aspects of the interventions caused the learning effects. Further research should make more fine-grained comparisons between, for example, different versions of a game. In this way, theories on the effectiveness of certain characteristics of game-based learning can be tested, and results can be used in informing the design and selection of educational games.

3.6. Conclusions

Our findings give evidence for the possibility of increasing primary school students' multiplicative reasoning ability through an intervention in which multiplicative mini-games are played at home and debriefed at school. When utilized in this way, mini-games were found to promote students' multiplicative operation skills (procedural knowledge) as well as their insight in multiplicative number relations (conceptual knowledge), and both an intervention in Grade 2 and a combined Grade 2–3 intervention were effective. When the mini-games were played at school, only a limited effect was found: this intervention only enhanced multiplicative insight and only in Grade 2. Playing the games at home without attention at school was not effective, indicating the importance of debriefing sessions when playing at home. Our findings further show that more gameplay was in some cases related to more learning, but this relation was not always present, indicating that there was not always a one-to-one relation between learning time and learning outcomes.

In the course of our research project, it appeared that a large-scale study situated in school practice is hard to carry out. Because of teachers' busy schedules it was hard to find teachers willing to participate in a long-term study, and to motivate teachers in subsequent grades to continue the study. However, we think that conducting this research in real school settings to collect evidence for the effectiveness of mathematics games in primary education was worth the effort. It provided us with knowledge of when and for what learning objectives mathematics mini-games are useful. Moreover, as the interventions were delivered by the teachers themselves, our results are directly applicable to the school practice.

Acknowledgments

This research was carried out with a grant from the "OnderwijsBewijs" program of the Ministry of Education in the

Netherlands (Project number: ODB 08007). We would like to thank all teachers and students who participated in this study. Furthermore we thank our colleagues Sylvia van Borkulo and

Hanneke Loomans for their contribution to the execution of the research project. Finally, we thank Anne Teppo for language editing this article.

Appendix A

Descriptions of the mini-games in the four game periods.

Mini-game	Description
<i>Game period 1</i>	
1. Catching	Catching ladybugs in equal-sized groups to be able to easily count them
2. Making groups	Making rectangular groups of smiley faces and determining the number of faces in each group (see Fig. 2a)
3. Stamps	Making multiplication problems (tables of 2, 5, and 10), connected to a representation of a number of equal-valued stamps on an envelope
4. Easy problem	Making multiplication problems (from 1×1 to 5×5) in a grid in which known neighbor problems can be used to find answers to unknown problems
5. Clothesline	Counting back and forth with steps of 2, 5, or 10
6. Quick problems	Quickly finding a total amount, represented as a collection of equal-valued coins (coins of 2, 5, or 10)
7. Which of three?	Choosing from three numbers the number that belongs to a certain multiplication table (tables of 2 and 5)
8. Three in a row	Selecting, in a grid of multiplication problems from 1×1 to 5×5 , a problem that has a given outcome. Subsequent selections should form a row of three in the grid as quickly as possible
<i>Game period 2</i>	
1. Choosing money	Choosing from two sets of coins or bank notes the largest amount of money. Each set contains multiple coins or bank notes of only one or two types, and is represented in a structured way
2. Making groups 2	Making rectangular groups of smiley faces and determining the number of faces in each group. Linking each rectangular group to the corresponding multiplication problem
3. Frog	Entering a known multiplication problem with the answer, then answering a related multiplication problem (see Fig. 2b)
4. Quick problems 2	Quickly finding a total amount, represented as a collection of equal-valued coins (coins of 2, 3, 4, 5, or 10)
5. Falling problems	Quickly deciding whether a falling multiplication problem has an outcome below or above 25 (problems with outcomes up to 50)
6. Wall of numbers	Finding combinations of two or more numbers that together multiply to a given target number (target numbers 12, 16, 18, 20, 24, and 36 are included)
7. Number factory	Combining numbers using addition, subtraction, and multiplication, to come as close as possible to a target number. For each target number at least one multiplication is needed to come close
8. Four in a row	Selecting, in a grid of multiplication problems from 1×1 to 10×10 , a problem that has a given outcome. Subsequent selections should form a row of four in the grid as quickly as possible
<i>Game period 3</i>	
1. Which of three? 2	Choosing from three numbers the number that belongs to a certain multiplication table (tables of 2 to 9)
2. Falling problems 2	Quickly deciding whether a falling multiplication problem has an outcome below or above 50 (problems with outcomes up to 100)
3. Art floor	Determining the surface area of differently sized tiles on a floor, based on comparison to tiles with known areas
4. Magic book	Combine four given numbers using addition, subtraction, and multiplication, to obtain a certain target number. For each target number at least one multiplication is needed
5. Money problems	Solving multiplication problems with money, such as $6 \times \text{€}12$. The student can request a representation of the problem with bank notes and coins, which encourages strategies such as first calculating $6 \times \text{€}10$
6. Fair sharing	Selecting a number of children among whom X bags with Y marbles each can be evenly divided. The numbers X and Y cannot be selected. An animation is shown in which the marbles are divided over the selected number of children one by one
7. Pay the exact amount	Paying a certain amount of money using only one type of coin or bank note. The student has to select the type of coin or bank note and decide how many of those are needed. Different solutions for the same problem are encouraged
8. Enlargement	Deciding how many times a small picture fits into an enlarged version of the picture. The length and width of the picture and the enlargement are given. The enlargement is displayed on squared paper; the small picture can be moved over the enlargement
<i>Game period 4</i>	
1. Four in a row	^a
2. Choosing money 2	Choosing from two sets of coins or bank notes the largest amount of money. Each set contains multiple coins or bank notes of only one or two types, and is represented in a structured way. More difficult sets

(continued on next page)

Appendix A (continued)

Mini-game	Description
3. Wall of numbers 2	are included as compared to “Choosing money” Finding combinations of two or more numbers that together multiply to a given target number (target numbers 24, 36, 48, 54, 60, 72, and 120 are included)
4. Number factory 2	Combining numbers using addition, subtraction, multiplication, and division to come as close as possible to a target number. For each target number at least one multiplication is needed to come close
5. Frog	^b
6. Pay the exact amount 2	Paying a certain amount of money using only one type of coin or bank note. The student has to select the type of coin or bank note and decide how many of those are needed. Different solutions for the same problem are encouraged. The amounts to be paid are higher and/or more difficult than in “Paying the exact amount”
7. Magic book	^c
8. Falling problems 3	Quickly deciding whether a falling division problem has an outcome below or above 5 (problems with outcomes up to 10)

Note. Mini-games with a 2 or 3 behind the name are new versions of earlier presented mini-games, with a higher difficulty level (e.g., including more difficult multiplication problems or including division).

^a Same as the eighth game in Game period 2.

^b Same as the third game in Game period 2.

^c Same as the fourth game in Game period 3.

Appendix B

Paired comparisons between the E_{school} , E_{home} , and $E_{\text{home-school}}$ condition of effects on the three aspects of multiplicative reasoning ability.

Effect	Aspect of multiplicative reasoning ability ^a					
	Knowledge		Skills		Insight	
	$\Delta\beta_{\text{ps}} (=d)$	SE	$\Delta\beta_{\text{ps}} (=d)$	SE	$\Delta\beta_{\text{ps}} (=d)$	SE
<i>Comparison $E_{\text{school}}-E_{\text{home}}$</i>						
Effect of Grade 2 intervention	0.18	0.21	0.10	0.15	0.14	0.15
Effect of Grade 3 intervention	-0.25	0.17	-0.03	0.11	0.06	0.09
Total effect of Grade 2–3 intervention	-0.15	0.23	0.05	0.13	0.13	0.15
<i>Comparison $E_{\text{home-school}}-E_{\text{home}}$</i>						
Effect of Grade 2 intervention	0.24	0.24	0.18 [†]	0.11	0.09	0.12
Effect of Grade 3 intervention	0.06	0.13	0.06	0.11	0.16 [*]	0.09
Total effect of Grade 2–3 intervention	0.21 [†]	0.14	0.19 [*]	0.11	0.21 [*]	0.09
<i>Comparison $E_{\text{home-school}}-E_{\text{school}}$</i>						
Effect of Grade 2 intervention	0.06	0.24	0.08	0.15	-0.05	0.18
Effect of Grade 3 intervention	0.31 [*]	0.18	0.08	0.10	0.10	0.09
Total effect of Grade 2–3 intervention	0.35 [*]	0.21	0.14	0.13	0.08	0.14

Note. E_{school} : $n = 112$; E_{home} : $n = 202$; $E_{\text{home-school}}$: $n = 78$. SkillsTest1, Female, AgeDelayed, ParEdLow, NonDutch, and GMath were included as covariates (see Fig. 3). Two-tailed significance tests were used for the $E_{\text{school}}-E_{\text{home}}$ comparison; one-tailed tests were used for the $E_{\text{home-school}}-E_{\text{home}}$ and the $E_{\text{home-school}}-E_{\text{school}}$ comparison (because of our directional hypothesis). $\Delta\beta_{\text{ps}}$ = partially standardized difference between coefficients.

^a The model was separately specified for each of the three aspects of multiplicative reasoning ability.

[†] $p < .10$. One-tailed.

^{*} $p < .05$. One-tailed.

Appendix C. Supplementary material

Supplementary data associated with this article can be found in the online version, at <http://dx.doi.org/10.1016/j.cedpsych.2014.09.001>.

References

- Anghileri, J. (2006). *Teaching number sense* (2nd ed.). London: Continuum.
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.
- Ault, M., Adams, D., Rowland, A., & Tiemann, G. (2010). Targeted educational games: Fun and so much more! Paper presented at the annual conference of the international society for technology in education, Denver, Colorado.
- Bai, H., Pan, W., Hirumi, A., & Kebritchi, M. (2012). Assessing the effectiveness of a 3-D instructional game on improving mathematics achievement and motivation of middle school students. *British Journal of Educational Technology*, 43, 993–1003.
- Bakker, M., Van den Heuvel-Panhuizen, M., Van Borkulo, S., & Robitzsch, A. (2012). Effects of mini-games for enhancing multiplicative abilities: A first exploration. In: S. De Wannemacker, S. Vandercruyse, & G. Clarebout (Vol. Eds.), *Communications in computer and information science: Vol. 280. Serious games: The challenge* (pp. 53–57). Berlin: Springer. <http://dx.doi.org/10.1007/978-3-642-33814-4_7>.
- Bakker, M., Van den Heuvel-Panhuizen, M., Van Borkulo, S., & Robitzsch, A. (2013). Effecten van online mini-games op multiplicative vaardigheden van leerlingen in groep 4 [Effects of online mini-games on Grade 2 students' multiplicative abilities]. *Pedagogische Studiën*, 90(3), 21–36.
- CBS (Central Bureau of Statistics) (2012). *StatLine Basisonderwijs; leerlingen in het basis- en speciaal basisonderwijs (StatLine primary education; students in primary and special primary education) (Data file)*. <<http://statline.cbs.nl/StatWeb/publication/?DM=SLNL&PA=37846sol&D1=a&D2=1&D3=a&D4=18&HDR=G2,G3,G1&STB=T&VW=T>>.
- CFI (Central Financial Institutions) (2011). *Gegevens Nederlandse basisscholen schooljaar 2009–2010 (Data Dutch primary schools in school year 2009–2010) (Data file)*. Received by e-mail on 02.12.11.

- Chang, K.-E., Sung, Y.-T., Chen, Y.-L., & Huang, L.-H. (2008). Learning multiplication through computer-assisted learning activities. *Computers in Human Behavior*, 24, 2904–2916.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cordova, D. L., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88, 715–730.
- De Vos, T. (2010). *TempoTest Automatiseren (Tempotest automatization)*. Amsterdam: Boom test uitgevers.
- Department for Education UK (2011). *Primary national curriculum until 2014. Mathematics: Key stage 1*. <<http://www.education.gov.uk/schools/teachingandlearning/curriculum/primary/b00199044/mathematics/ks1>> Retrieved 27.05.13.
- Egenfeldt-Nielsen, S. (2005). *Beyond edutainment. Exploring the educational potential of computer games*. Doctoral dissertation, IT-University of Copenhagen, Denmark.
- Frazier, A., Argles, D., & Wills, G. (2007). Assessing the usefulness of mini-games as educational resources. *Paper presented at the annual conference of the association for learning technology, Nottingham, UK*.
- Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, 33, 441–467.
- Gee, J. P. (2005). Good video games and good learning. *Phi Kappa Phi Forum*, 85(2), 33–37.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576.
- Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *The Journal of the Learning Sciences*, 20, 169–206.
- Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Baltimore, MD: Johns Hopkins University Press.
- Honey, M. A., & Hilton, M. (Eds.). (2011). *Learning science through computer games and simulations*. Washington, DC: National Academies Press.
- Jansen, B. R., Louwerse, J., Straatemeier, M., Van der Ven, S. H., Klinkenberg, S., & Van der Maas, H. L. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, 24, 190–197.
- Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8 (Scientific justification of the LOVS mathematics tests for Grades 1 to 6)*. Arnhem, Netherlands: Cito.
- Jonker, V., Wijers, M., & Van Galen, F. (2009). The motivational power of mini-games for the learning of mathematics. *Paper presented at the European conference on game based learning, Graz, Austria*.
- Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. In R. E. Ferdig (Ed.), *Handbook of research on effective electronic gaming in education* (pp. 1–32). Hershey, PA: IGI Global.
- Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & Education*, 55, 427–443.
- Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson.
- Kirriemuir, J. (2002). *The relevance of video games and gaming consoles to the higher and further education learning experience*. Techwatch report. <<http://tecnologiaedu.us.es/nweb/html/pdf/301.pdf>> Retrieved 10.12.12.
- Klawe, M. M. (1998). When does the use of computer games and other interactive multimedia software help students learn mathematics? *Paper presented at the technology and NCTM standards 2000 conference, Arlington, VA*.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.
- Kolovou, A., Van den Heuvel-Panhuizen, M., & Köller, O. (2013). An intervention including an online game to improve Grade 6 students' performance in early algebra. *Journal for Research in Mathematics Education*, 44, 510–549.
- Leemkuil, H., & De Jong, T. (2004). Games en gaming (Games and gaming). In P. Kirschner (Ed.), *ICT in het onderwijs: The next generation. Katern bij onderwijskundig lexicon, uitbreiding editie III* (pp. 41–63). Alphen aan de Rijn, Netherlands: Kluwer.
- Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review*, 22, 215–243.
- Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, 4, 333–369.
- Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning, and instruction: Cognitive and affective process analyses* (Vol. 3, pp. 223–253). Hillsdale, NJ: Lawrence Erlbaum.
- Mayer, R. E. (2011). Multimedia learning and games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 281–305). Charlotte, NC: Information Age.
- Miller, S. P., & Hudson, P. J. (2007). Using evidence-based practices to build mathematics competence related to conceptual, procedural, and declarative knowledge. *Learning Disabilities Research & Practice*, 22, 47–57.
- 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, 850–864.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.
- Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles: Author.
- NCTM (National Council of Teachers of Mathematics) (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Reston, VA: Author.
- Nunes, T., Bryant, P., Barros, R., & Sylva, K. (2012). The relative importance of two different mathematical abilities to mathematical achievement. *British Journal of Educational Psychology*, 82, 136–156.
- Panagiotakopoulos, C. T. (2011). Applying a conceptual mini game for supporting simple mathematical calculation skills: Students' perceptions and considerations. *World Journal of Education*, 1(1), 3–14.
- Panagiotakopoulos, C. T., Sarris, M. E., & Koleza, E. G. (2013). Playing with numbers: Development issues and evaluation results of a computer game for primary school students. In T. Sobh, & K. Elleithy (Eds.), *Emerging trends in computing, informatics, systems sciences, and engineering* (Vol. 151, pp. 263–275). Lecture notes in electrical engineering. <http://dx.doi.org/10.1007/978-1-4614-3558-7_22>.
- Parsons, H. M. (1974). What happened at Hawthorne? *Science*, 183, 922–932.
- Prensky, M. (2001). *Digital game-based learning*. New York: McGraw-Hill.
- Rosas, R., Nussbaum, M., Cumsille, P., Marianov, V., Correa, M., Flores, P., et al. (2003). Beyond Nintendo: Design and assessment of educational video games for first and second grade students. *Computers & Education*, 40, 71–94.
- Sandberg, J., Maris, M., & De Geus, K. (2011). Mobile English learning: An evidence-based study with fifth graders. *Computers & Education*, 47, 1334–1347.
- Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22, 278–295.
- Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research*, 78, 427–515.
- Tobias, S., Fletcher, J. D., Dai, D. Y., & Wind, A. P. (2011). Review of research on computer games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 127–222). Charlotte, NC: Information Age.
- Van Borkulo, S., Van den Heuvel-Panhuizen, M., Bakker, M., & Loomans, H. (2012). One mini-game is not like the other: Different opportunities to learn multiplication tables. In: S. De Wannemacker, S. Vandercruyssen, & G. Clarebout (Vol. Eds.), *Communications in computer and information science: Vol. 280. Serious games: The challenge* (pp. 61–64). Berlin: Springer. <http://dx.doi.org/10.1007/978-3-642-33814-4_9>.
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- Van den Heuvel-Panhuizen, M. (Ed.). (2008). *Children learn mathematics: A learning-teaching trajectory with intermediate attainment targets for calculation with whole numbers in primary school*. Rotterdam, Netherlands: Sense.
- Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*, 34, 229–243.
- Williamson, B. (2009). *Computer games, schools, and young people: A report for educators on using games for learning*. Bristol, UK: Futurelab.
- 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, 249–265.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER ConQuest (version 2.0) (Computer Software)*. Camberwell, Australia: Australian Council for Educational Research.
- Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., et al. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research*, 82(1), 61–89.