British Journal of Educational Technology *Vol 48 No 3 2017* 860–877
doi:10.1111/bjet.12458

The effect of surprising events in a serious game on learning mathematics

Pieter Wouters, Herre van Oostendorp, Judith ter Vrugte, Sylke vanderCruysse, Ton de Jong and Jan Elen

Pieter Wouters holds a PhD in Instructional Design and works at the Department of Information and Computing Sciences (Interaction Technology group), Utrecht University, Utrecht 3508 TB, Netherlands. Herre van Oostendorp is associated professor Human Media Interaction, at Utrecht University, Department of Information and Computing Sciences (Interaction Technology group), Utrecht University, Utrecht 3508 TB, Netherlands. Judith ter Vrugte is PhD student at the Department of Instructional Technology, Faculty of Behavioral Sciences, University of Twente, Enschede 7500 AE, Netherlands. Sylke Vandercruysse holds a PhD in Education Sciences and works at the Faculty of Psychology and Educational Sciences @ Kulak, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium. Ton de Jong is full professor of Instructional Technology at the University of Twente, Enschede 7500 AE, Netherlands. Jan Elen is full professor at the CIP&T, Center for Instructional Psychology and Technology, KU Leuven, Dekenstraat 2, 3000 Leuven, Belgium. Address for correspondence: Dr Pieter Wouters, Department of Information and Computing Sciences (Interaction Technology group), Utrecht University, P.O. Box 80.089, 3508 TB Utrecht, The Netherlands. Tel: +31 (30) 253 6355; Fax: +31 (30) 253 2804; Email: p.j.m.wouters@uu.nl

Abstract

The challenge in serious games is to improve the effectiveness of learning by stimulating relevant cognitive processes. In this paper, we investigate the potential of surprise in two experiments with prevocational students in the domain of proportional reasoning. Surprise involves an emotional reaction, but it also serves a cognitive goal as it directs attention to explain why the surprise occurred and can play a key role in learning. In our experiments, surprises were triggered by a surprising event, ie, a nonplaying character who suddenly appeared and changed characteristics of a problem. In Experiment 1—comparing a surprise condition with a control condition—we found no overall differences, but the results suggested that surprise may be beneficial for higher level students. In Experiment 2, we combined Expectancy strength (Strong vs. Weak) with Surprise (Present vs. Absent) using higher level students. We found a marginal overall effect of surprising events on learning indicating that students who experienced surprises learned more than students who were not exposed to these surprises but we found a stronger effect of surprise when we included existing proportional reasoning skill as factor. These results provide some evidence that a narrative technique as surprise can be used in game-based learning for the purpose of learning.

Despite the increasing popularity of serious games or game-based learning (GBL), recent meta-analytic reviews have shown that GBL is only moderately more effective and not more motivating than traditional instruction (Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013; Clark, Tanner-Smith, & Killingsworth, 2015). For example, in their meta-analysis Wouters *et al.* (2013) found only a (significant) moderate effect size (.29) for learning in favor of GBL. Likewise, they found a moderate, but statistically nonsignificant, effect for motivation in favor of GBL.

GBL influences learning in two ways, directly by changing the cognitive processes and indirectly by affecting the motivation (Wouters *et al.*, 2013). Preferably both sources should be used to maximize learning. A potential problem with GBL is that the outcomes of players' actions in the

Practitioner Notes

What is already known about this topic

- There is only little empirical evidence about the effect of surprise in serious games.
- That a surprising event can lead to a better comprehension of text.

What this paper adds

- Two controlled empirical experiments investigating the effects of serious game design on learning.
- Indications that surprises in a serious game yield higher learning gain in mathematics, in particular for learners with sufficient (meta)cognitive skills.

Implications for practice and/or policy

- Narrative techniques such as embedding surprising events can be used in a serious game to improve learning.
- Research in serious game design is relevant because it can lead to more effective serious games.

game are directly reflected in the game world which may lead to a kind of intuitive learning: players know how to apply knowledge, but they cannot explicate it. In other words: they do not necessarily acquire the underlying rules (Leemkuil & de Jong, 2011). It is possible that studies, therefore, find no relation between success in the game and success on a knowledge test. The articulation of knowledge and underlying rules is important, because it triggers students to *organize* new information and *integrate* it with their prior knowledge (Mayer, 2011; Wouters, Paas & van Merriënboer, 2008) and thus construct a mental model that is more broadly applicable. This implies that genuine learning in GBL requires additional features in the game that will provoke the player to engage in the process of knowledge articulation. Our review on the impact of instructional support, however, shows that such support often fails to effectively facilitate the process of knowledge articulation (Wouters & van Oostendorp, 2013). In learning environments, knowledge articulation is often prompted by explicitly asking students to reflect on their actions and thoughts, eg, by means of self-explanations (cf., Chi, de Leeuw, Chiu, & La Vancher, 1994). In complex GBL environments, such an explicit intervention may compromise the motivating quality of the game because it disturbs the flow of the game or can be so cognitively demanding that learning will not take place (cf., ter Vrugte *et al.*, 2015).

The question raised in this paper is how we can stimulate players to engage in relevant cognitive processes such as organizing and integrating knowledge that foster learning without jeopardizing the motivational appeal of the game. A promising technique is the generation of manageable cognitive conflicts by introducing surprises. In this study, we make a distinction between a *surprising event* and *surprise*. This distinction is also made by other scholars (eg, Adler, 2008; Foster & Keane, 2015a). We define surprise as a disruption of an active expectation. Surprise involves an emotional reaction, but it also serves a cognitive goal as it directs attention to explain why the surprise occurred and can play a key role in learning (Foster & Keane, 2015b; Howard-Jones & Demetriou, 2009; Ranganath & Rainer, 2003). In our view, a surprise is triggered by a surprising event, ie, the occurrence of an unexpected event that disrupts the coherence or logical sequence of a series of events resulting in an urgent representational updating process (Itti & Baldi, 2009; Maguire, Maguire, & Keane, 2011). This notion of a surprising event aligns with definitions in other fields, although it is sometimes mentioned slightly different. In story discourse, eg, Brewer

and Lichtenstein (1982) describe a surprise event structure in which information is omitted in the beginning of a story and inserted later in such a way that the coherence of the story discourse is disrupted. There is also a relation with psychological theories regarding learning such as the unexpected event hypothesis which regards unexpected events as discrepancies at the behavioral level that occur when learners perform a task in a way that deviates from the performance they expected (Rünger & French, 2008). Finally, also educational theories use this term. For example, in their study on conceptual change Clement and Steinberg (2002) use the term discrepant events to refer to events that disrupt the coherence of a mental model. In addition, they state that these discrepant events produce reactions of surprise and are eventually followed by model revisions. In the remainder of this study, we will use the term surprising event when we *explicitly* refer to the intervention in the learning environment. When we use the term surprise we refer to the situation that a surprising event emotionally and cognitively affects the student.

In the domain of narratives and text comprehension, it has been shown that surprise has a beneficial effect on learning. Hoeken and van Vliet (2000) found that surprise improved text comprehension and appreciation more than other techniques such as curiosity and suspense. Likewise, O'Brien and Myers (1985) confronted participants with a word that was either predictable or unpredictable from a preceding context and observed that the texts that preceded unpredictable words were better recalled. In learning a medical procedure with a serious game, van der Spek, van Oostendorp, and Meyer (2013) demonstrated that surprise yielded superior knowledge structures, indicating that it fosters deep learning.

Readers understand a story because they construct a situation model in which dimensions such as the protagonist, time, space, causality and intentionality are related (Zwaan, Langston, & Graesser, 1995). Likewise, in computer games players construct a mental model and/or situation model based on the story line, the events and the underlying rules of the game (van der Spek *et al.*, 2013). The situation or mental model makes new events plausible (although such events may cause adaptations in the model) and is the starting point for expectations of the reader or player. A surprising event on the other hand is unexpected and not logically follows from the situation/mental model. Readers/Players will be surprised and wonder what they have missed and start to re-evaluate preceding events. In this process, the mental model will be activated, retrieved and updated, thereby enhancing learning (van der Spek *et al.*, 2013).

The assumption of the current study is that surprise also pertains to problem solving in serious games. Ideally, the mental model will enable the student to recognize specific characteristics of a problem and determine how the problem can be solved. Because our aim is to integrate the instructional technique (ie, the introduction of the surprising events) with the learning content (Habgood & Ainsworth, 2011), the surprises have to be focused on what has to be learned, ie, the mental model of proportional reasoning problems and methods to solve them. For this reason, the surprising events change some of the problem characteristics, and the solution method previously applied, is no longer easily applicable and the player has to re-evaluate the situation and decide which problem characteristics are now relevant and which solution method is now most appropriate (this can be the one used earlier, but also another one). We expect that surprise has a positive effect on learning because it involves relevant cognitive processes such as organizing and integrating information (Mayer, 2011) without compromising the motivational appeal of computer games.

In this study, we investigate the impact of surprise on learning and how this impact is moderated by the expectancy of the student (in the second study). We used the GBL environment "*Zeldenrust*" that was specifically developed for learning proportional reasoning in secondary prevocational education (see Vandercruysse *et al.*, 2015) (see also http://www.projects.science.uu.nl/mathgame/zeldenrust/). Proportional reasoning was chosen because it is a relevant and well-defined domain and existing methods for proportional reasoning are often ineffective (Rick, Bejan,

Roche, & Weinberger, 2012). To qualify learning, we look separately to the skill to solve problems that were comparable with the problems in the game (proportional reasoning items) and to the skill to solve problem in different contexts in which the proportional reasoning characteristics were more difficult to discern (transfer items).

Experiment 1

In Experiment 1, a group of students playing the game with surprising events (surprise group) embedded in the game was compared with a group without these surprising events (control group). We expected that the group with surprising events would experience surprise and, therefore, learn more than the control group.

Method

Participants and design

The participants were 71 students from 2nd-year prevocational education with a mean age of 14.1 ($SD = .61$) recruited from four classes of two schools. We adopted a pretest-posttest design with a control condition ($N = 36$) and a surprise condition ($N = 35$). Participants were randomly assigned to the conditions. Dependent variables were proportional reasoning skill, transfer skill and game performance.

Materials

Domain. The domain of proportional reasoning comprises three problem types: comparison problems, missing value problems and transformation problems (cf., Tourniaire & Pulos, 1985). In comparison problems, students have to find out whether one ratio is "more than," "lesser-than" or "equal to" another ratio. These problems can be classified in difficulty levels ranging from equal values of ratios (eg, 2/11 and 3/11), ratios with simple multiplication (eg, 11/20 and 22/36) or complete calculation. In missing value problems, one value in one of two ratios is missing. Students have to find this "missing value" to ensure that both ratios are equal. Transformation problems involve two ratios as well and all values are known, but the ratios are not equal. Students have to find out how much has to be added to one or more of the ratios to make both ratios equal (for a more extensive description see Vandercruysse *et al.*, 2015). Both missing value and transformation problems can be classified in one of four difficulty levels based on the integrity of the ratio within (comparing the same term of two ratios) or between (comparing the different terms of the ratios) two ratios. For example, a problem with 1/2 and 3/6 can be classified as Level 1 (easy) because both the ratio "within" (in this Case 1 and 3 or 2 and 6) and the "between" ratio (in this Case 1 and 2 or 3 and 6) are whole integers.

Game environment. *Zeldenrust*, a cartoonlike 2D game developed in Flash/ActionScript 3, can best be characterized as a combination of a simulation game and a role-playing game. Players have a summer job in a hotel and can earn money for a holiday destination by doing different tasks in the game: the more money they earn, the further they can travel.

During the game the player is accompanied by the manager, a nonplaying character, who provides information about the task and gives feedback regarding the performance on the task. The game comprises a base game and several subgames. The base game provides the structure from which the subgames can be started. After selecting an avatar, the players receive an introduction animation in which the context of the game is presented and finally enter the "Student room" from which the player can control the game (eg, by choosing a specific subgame). Each task is implemented as a subgame and covers a specific problem type in the domain of proportional reasoning. The tasks are directly related to proportional reasoning (eg, mixing two drinks to make a cocktail according to a particular ratio directly involves proportional reasoning skills).

In addition, mental operations with respect to proportional reasoning are connected with the game mechanics (eg, to get the correct amount of bottles in the refrigerator the player has to drag the correct number of bottles in the refrigerator). Table 1 shows the subgames and the distribution of difficulty levels across the game levels.

Although the subgames cover different problem types, they have several common elements. The actual assignment is described on a *whiteboard*. With drag-and-drop or clicking the player can accomplish the assignment, but the specific action depends on the subgame. To further motivate the player, a "*geldmeter*" (*money meter*) is implemented which visualizes the amount of money that the player will receive after an assignment. Correct and incorrect actions during an assignment are directly reflected in the money meter. For example, if the player breaks a bottle, the money meter will decrease (and the color becomes redder); if the player places bottles in the refrigerator the money meter will increase (and becomes greener). The money meter also shows the (accumulated) amount of money that the player has earned. The player can use a built-in *calculator*, but using it will cost some money. Depending on the subgame, the player has to perform a typical action (eg, closing the door of the refrigerator) to receive *verbal feedback* from the manager of the hotel who tells whether the answer is correct or not (eg, "Excellent" or "You have too much Cola in relation to Fanta"). If the answer is correct the money meter will be increased.

In the *control* condition, all assignments were presented in an identical way: all information required to perform the assignment was available. The whiteboard described what the player had to do. In the Refrigerator subgame, bottles or crates of Cola with a caption indicating the number of bottles they represented were placed aside the refrigerator. The Blender subgame contained bottles of yoghurt and juice with a caption indicating the number of units they represented. The Serving subgame contained two jugs with a caption indicating the ratio of yoghurt and juice in these jugs. Based on this information, the player could decide and act.

The *surprises* condition was different from the control condition in two ways. To start with, this condition involved a nonplaying niece character in the introduction animation who tells she is bored and that she sometimes will make it difficult to carry out the task. When a surprising event occurred the niece character popped up and told that she had changed something. This change involved specific characteristics of the task whereby the player has to reconsider the original solution method. Eight surprising events were equally divided over the Refrigerator and Blender subgames (the Jugs subgame was not suitable for the surprising events). The moment that the surprising events occurred was predefined by the designers but for the players these events occurred randomly. Also, the triggering of the surprising event during the task was unpredictable. In one task, the event could start after the player had dragged 12 or more bottles to the refrigerator, while in another task this could happen after dragging 2 bottles. Figure 1 gives an example of the occurrence of a surprising event. The assumption is that the event incites a surprise (an emotional and cognitive response).

Figure 1a depicts the starting situation. The given number of Fanta bottles is 24. The player has to figure out how many bottles of Cola have to be put into the refrigerator in a ratio of 9–12. The player can solve the problem by looking at the ratio "within": the number of Fanta in the refrigerator is twice as much as the number of Fanta in the desired ratio (12 Fanta) as $12 \times 2 = 24$, so the number of Cola also has to be doubled ($9 \times 2 = 18$ Cola). The easiest way to compute this is to use the so-called "within" ratio method. When the player is implementing the solution the surprising event occurs while the player had already put in 10 bottles of Cola (Figure 1b). When the niece character has disappeared it appears that characteristics of the task are modified (Figure 1c); ie, the desired ratio changed by the niece character is now 5 Cola per 10 Fanta. In this case, the ratio "within" that the player used is not applicable anymore and the player can better use a method based on the ratio "between" method (the desired proportion is 5 Cola/10 Fanta, so the

Table 1: Description of the subgame and their level structure

Subgame	*Problem type*	*Example of problem*	*Game level: Difficulty of proportional problem*
Jugs	Comparison	"There are two jugs of juice on the counter. A customer asks for the sweetest juice mix. Which juice mix will you give to the customer?" The ratio of milk/fruit is presented on the jugs. The student has to draw the correct jug to a tray to solve the problem and receive feedback.	1: contains Level 1 problems 2: contains Level 2 problems 3: contains Level 3 problems 4: contains a mix of all levels
Fridge	Missing value	"This is the reception desk refrigerator. This refrigerator always contains 3 bottles of Fanta for every bottle of Cola. It already contains 9 bottles of Fanta. Fill the refrigerator so it will contain the right amount of Cola." The given ratio of 3/1 is presented next to the ratio with the missing value 9/?. The student has to answer the question by clicking the juice bottles into the refrigerator. Clicking on the refrigerator door evaluates the answer of the student.	1: contains Level 1 problems 2:contains Levels 2 and 3 problems 3: contains Level 4 problems 4: contains a mix of all levels
Blender	Transformation	"A fruit cocktail contains 28 units of juice for every 24 units of yoghurt. According the recipe this should be 6 units of juice and 9 units of yoghurt. How many units of juice and/or yoghurt should you add to the blender so that the mix is in line with the mix?" The student can move the bottles to the blender and stir units into the blender. The evaluation of the answer follows after clicking the blender button.	1: contains Level 1 problems 2:contains Levels 2 and 3 problems 3: contains Level 4 problems 4: contains a mix of all levels

Source: Adapted from ter Vrugte *et al.* (submitted).

number of Cola in the refrigerator should also be half the number of the given bottles of Fanta, 12/24). In total, the players received eight surprising events (four in both the missing value and the transformation subgames). Please note that players can use other methods as well. It is possible that they use a method before the surprising event that can also be used for the problem after the surprising event. However, the surprising event still incites the player to think about the question whether the chosen method is still applicable in the new situation.

Tests. The arithmetic tempo test, the *Tempo Test Rekenen* (TTR), measures the degree of fluency in basic arithmetic operations, ie, addition, subtraction, multiplication and division (de Vos, 1992). For each operation, there is a column with 40 arithmetic problems (so four columns). A fifth

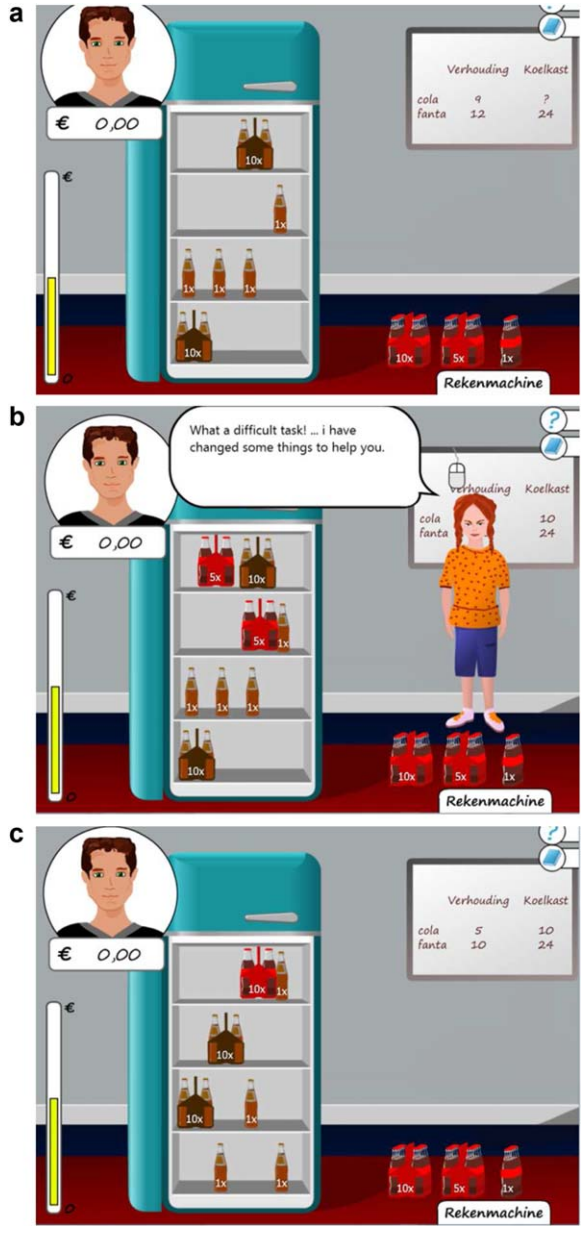

Figure 1: (a) Starting situation in a task with a surprising event in the game Zeldenrust. (b) Occurrence of the surprising event. (c) Task characteristics have been modified

column contains problems with mixed operations. The students have 1 minute per column to solve as many arithmetic problems as possible.

Proportional reasoning skill was measured with a test consisting of 12 open questions: 4 questions for each problem type. An example (missing value) is:

"For a banana milkshake you have to use 28 bananas and 48 units of ice. How many units of ice do you need if you are going to use 56 bananas and you want to remain the same proportion?"

The questions were comparable with the assignments in the game.

Transfer skill was measured with four additional items in which the contexts were different from those presented in the game tasks. In addition, in these items it was more difficult to discern the relevant problem characteristics. An example item is:

"A hiker walks uphill. It takes 60 seconds for 80 meters. How much time does it take to walk 120 meters uphill?"

There were two versions of the tests. The structure of these versions was the same, but the numbers were different. The comparability of both versions was tested in pilot study.

Procedure

The experiment was run on the computers of the schools. The experiment took 150 minutes divided into three sessions of 50 minutes. In the first session, the experiment was introduced and the pretest was administered (40 minutes). When participants had finished the pretest they could do their homework. In the second session, a week later, the participants played the game (40 minutes). At the beginning of the session, the participants were seated at a designated computer and received a login code. All actions of the players during playing the game were logged. The posttest was administered in the third session (40 minutes, a week after playing the game). One version was used in the pretest, the other version in the posttest.

Scoring

TTR. The TTR score is calculated as the sum of correct answers in the five columns. The range of possible scores is 0–200.

Skill test. Each answer in the pretest and posttest (both proportional reasoning skill and transfer skill) was coded as 0 (*wrong answer or no answer*) or 1 (*correct answer*).

Game performance. Due to technological problems during logging, the data of six participants was removed from the dataset. Two variables were calculated for each participant:

1. The total time they spent in a subgame to perform the assignments.
2. The number of assignments they correctly solved in a subgame.

Results and conclusion

For all statistical tests, a significance level of .05 was applied. To test whether playing the game yields learning, we separately tested the effect on the proportional reasoning items and the transfer items. To test the effect of surprise on learning, we used the combined score of the proportional reasoning items of the two problem types in which surprise was applied (missing value—Refrigerator subgame; transformation—Blender subgame). Table 2 shows the results for each condition.

An independent *T* test with TTR showed no difference in computational fluency between both conditions ($t(69) = .17$, $p > .05$). Both conditions did not differ in prior knowledge (proportional reasoning items: $t(69) = .11$, $p > .05$; transfer items: $t(69) = .08$, $p > .05$).

A paired samples *T* test reveals that playing the game ($t(70) = 2.73$, $p = .008$, $d = 0.29$) improves proportional reasoning skills. Playing the game, however, does not yield improvement in solving the transfer items ($t(70) = 1.35$, $p > .05$). An independent *T* test with overall learning gain (posttest score–pretest score) as dependent variable shows no difference between control and surprises condition ($t(69) = .07$, $p > .05$) regarding proportional reasoning items or the transfer items ($t(69) = .94$, $p > .05$). In addition, we evaluated if game performance predicted posttest performance on the proportional reasoning items and the transfer items with a hierarchical regression analysis. The first block consisted of the pretest score and the TTR score. From Table 3 can be concluded that pretest and TTR predict 41% of the posttest variance. In the second block, correct

Table 2: Mean scores and standard deviations on the dependent variable for all conditions of Experiment 1

	Control		*Surprises*	
TTR	*76 (12)*		*75 (20)*	
	Pre	*Post*	*Pre*	*Post*
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
All items [0–12]	4.32 (2.33)	5.07 (2.65)	4.38 (2.01)	5.02 (2.56)
Surprise items [0–8]	2.30 (2.87)	2.88 (2.34)	2.06 (1.71)	2.66 (2.31)
Comparison [0–4]	2.02 (.99)	2.20 (.91)	2.32 (1.83)	2.36 (1.18)
Transfer items [0–4]	.72 (.84)	.48 (.67)	.71 (.69)	.68 (.79)

Notes: Range of scores between []. All items mean all proportional reasoning skill items. Surprise items are missing value + transformation items.

Table 3: Hierarchical regression on posttest performance in Experiment 1

	Proportional reasoning			*Transfer*		
	B	*SE B*	*β*	*B*	*SE B*	*β*
Step 1: Constant	1.35	1.27		−.79	.44	
Pretest score	.84	.12	.69****	.15	.11	.16
TTR	.001	.02	.003	.02	.01	.34**
Step 2: Constant	1.92	1.53		−.66	.57	
Pretest score	.69	.13	.57****	.18	.13	.20
TTR	−.01	.02	−.04	.01	.006	.35**
Correct task	.11	.04	.37***	−.01	.02	−.11
Time-on-task	−.001	.001	−.16	.000	.000	−.008

Notes: *Proportional reasoning:* $R^2 = .46$ for Step 1, $\Delta R^2 = .06$ for Step 2.
Transfer: $R^2 = .15$ for step 1, $\Delta R^2 = .01$ for Step 2.
*$p < .05$, **$p < .01$, ***$p < .005$, ****$p < .001$.

assignments and time on task were entered stepwise. Using two blocks, the effect of the pretest score and the TTR score on the posttest score can be isolated.

As shown in Table 3, game performance only partly predicts posttest performance on the proportional reasoning items. When the variance caused by the pretest and the TTR score is accounted for (Block 1), only the number of correct tasks is predictive for posttest performance, but it explains only 6% additional variance. For the transfer items, neither the number of correct game tasks nor the time spent on the game were predictive for posttest performance.

There are two plausible explanations for the finding that the surprise condition did not perform better than the control condition. To start with, a surprise requires a certain level of cognitive flexibility and (meta)cognitive skills. Students must perceive and understand that the changes in the problem situation of the game are not superficial but that some deeper characteristics of the problem have been altered, see that the changes may have consequences for the chosen solution method and consider whether another method is more appropriate. For students who do not possess these skills sufficiently, surprise can be confusing or even frustrating because their solution method is thwarted. The students in this experiment were recruited from three educational levels

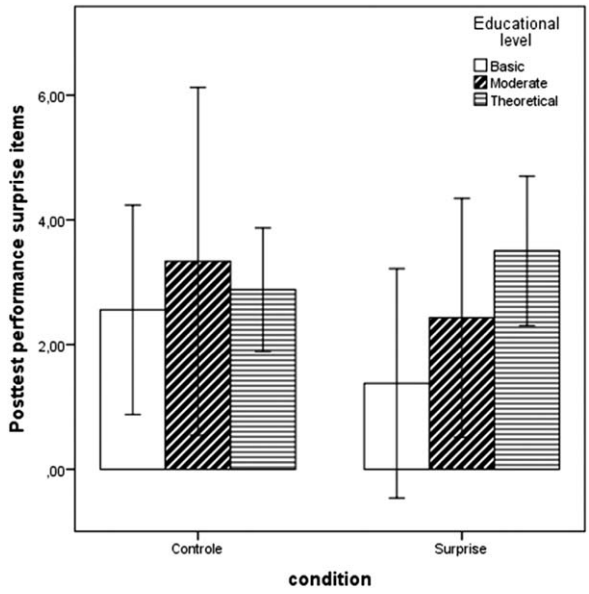

Figure 2: Posttest performance score surprise items (missing value + transformation items) for basic, moderate and theoretical level students in Experiment 1

in prevocational education. Prevocational education in the Netherlands lasts 4 years and students are divided into an educational level when they enter prevocational education. Each educational level gives an indication regarding cognitive flexibility and (meta)cognitive skills. Students that follow the *basic* level have a practical orientation, ie, they have difficulty with learning. The focus in the curriculum is to have students learn the theory by doing things. In the *moderate* level students, still like to do things, but they have less difficulty with learning. In the *theoretical* level, students have no problem with learning, can concentrate well and have no problem with the theory. Whereas the first two educational levels prepare for vocational education, the last level also prepares for senior general secondary education. Our expectation is that students in the theoretical level have higher (meta)cognitive skills than those in the basic and moderate level. Figure 2 shows the posttest scores on the surprise items for the three educational levels in each condition.

Although ANCOVAs (posttest score surprise items as dependent variable; condition and educational level as independent variables and TTR and pretest score surprise items as covariates) showed no significant effects (main effect condition and educational level $F < 1$; interaction condition and level $F(1,67) = 1.36$, $p > .05$), theoretical level students seem to benefit more from surprises. Although refuted by the test results, Figure 2 suggests that theoretical level students benefit from surprise, whereas surprise has less added value for low level students.

The second explanation concerns the order in which problems with different characteristics were presented to the player. Often the characteristic of a new problem was different from the preceding one which may have thwarted the emergence of a strong expectation. If this is true, the potential beneficial effect of surprise may not have been fully realized. In that case, the inclination of the student to retrieve and update the mental model will be weak.

Experiment 2

In Experiment 2, we investigated the two possible explanations discussed in Experiment 1. First, participants were recruited only from the theoretical educational level. Second, we introduced a second independent variable in which expectancy was manipulated. We tested three hypotheses:

1. Playing the game will improve learning.
2. We expect a main effect of surprise indicating that surprise will increase proportional reasoning skill more because the surprising events trigger students to interpret the changes in the problem characteristics and the consequences for the solution process.
3. In addition, we hypothesize an interaction between surprise and expectancy strength indicating that a surprising event after multiple problems with the same characteristics (Strong expectancy) will have the largest effect on proportional reasoning skill because the strong unexpectedness of the surprising event will incite students more (or deeper) to think about the changes in the characteristics of the problem characteristics and the consequences for the solution process.

Method

Participants and design

The participants were 94 students from 2nd-year prevocational education with a mean age of 13.9 ($SD = .81$) recruited from five classes of one school. We adopted a pretest-posttest design with the independent variables *Surprises* (Yes or No) and *Expectancy strength* (Strong or Weak) resulting in four conditions: Surprises and Strong expectancy ($N = 22$), Surprises and Weak expectancy ($N = 23$), No surprises and Strong expectancy ($N = 26$) and No surprises and Weak Expectancy ($N = 23$). Participants were randomly assigned to conditions. Dependent variables were proportional reasoning skill, transfer skill and game performance.

Materials

Domain. The same domain was used as in Experiment 1.

Game environment. The game environment (*Zeldenrust*) was the same as in Experiment 1 with a modification in the surprise conditions. In Experiment 1, the niece character appeared immediately while in Experiment 2 the screen first became brighter (to make the surprise more salient) and then dimmed again before the niece character appeared.

In Experiment 1, the characteristics of a problem changed often with every new problem within a game level. This means that the characteristics of the current problem could be different from the previous and the next problem. In this way, problems with only integer intern ratios, only integer extern ratios or a combination of both were possible in one level. This may have prevented players to create strong expectations. For the implementation of the factor *Expectancy Strength*, we presented the assignments differently. To facilitate the occurrence of strong expectancy, we classified the problems according to their characteristics into three groups

1. Problems with an integer intern ratio but not an integer extern ratio
2. Problems with an integer extern ratio but not an integer intern ratio
3. Problems have neither an integer extern ratio nor an integer intern ratio

Strong expectancy was defined as a series of problems with the same characteristics (eg, three consecutive problems from Group 1). Weak expectancy was defined as a series of problems in which the characteristics of each problem varied. In all conditions, the players received three levels with five problems in each level, but the distribution of problems with specific problem characteristics was different. All together this resulted in the following conditions:

Surprise with Strong expectancy. The first level consisted of only problems with an integer intern ratio. When the player started solving the third consecutive problem with that problem characteristic (meant to create strong expectancy) a surprising event occurred that changed the problem into a problem with different characteristics. During the fifth problem, another surprising event occurred. The same procedure was applicable to the second level which consisted of

problems with an integer extern ratio. The problems in the third level had neither an integer extern ratio nor an integer intern ratio and no surprising events occurred. In total, the players received 8 surprising events (four in both the missing value and the transformation subgames).

Surprises with Weak expectancy. In all levels problems with different characteristics were presented in random order. For example, Level 1 started with a problem with an integer extern ratio followed by a problem with neither an integer extern ratio nor an integer intern ratio and then again a problem with an integer extern ratio. In this way, the player did not know what to expect with each new problem (Weak expectancy). The surprising event also occurred during the third and fifth problem in the first two levels (in total again 8 surprising events).

No surprises with Strong expectancy. In each level only problems with the same characteristics were presented (Level 1: integer intern ratio, Level 2: integer intern ratio, Level 3: no intern and no extern ratio). During the third and fifth problem, the characteristics were not modified by a surprising event.

No surprises with Weak expectancy. In all levels, problems with different characteristics were presented in random order. During the third and fifth problem, the characteristics were not modified by a surprising event.

The tests, procedure and scoring were the same as in Experiment 1.

Results and conclusion

For the analysis of the results the same procedure as in Experiment 1 was used. Table 4 shows the results for each condition on TTR and the dependent variables.

ANOVA's with post hoc comparisons on TTR and prior proportional reasoning skill revealed no differences between the conditions (TTR: $F = 1.25$, $p > .05$; proportional reasoning items and transfer items $F < 1$). To test hypothesis 1, we conducted a paired-samples T test. The results show that playing the game improves proportional reasoning skills ($t(93) = 2.54$, $p = .013$, $d = .25$). Playing the game, however, does not yield improvement in solving the transfer items ($t(93) = 1.72$, $p > .05$).

This is partly corroborated by the results of the hierarchical regression analysis (see Table 5): for proportional reasoning items the TTR and pretest score predict 33% of posttest variance, while the number of correct tasks in the game explains an additional 10% of the posttest variance. For the transfer items, these numbers are 22 and 6%.

Hypotheses 2 and 3 were tested with a 2 × 2 ANCOVA with Surprises and Expectancy strength as independent variables, posttest score on the surprise items (missing value and transformation problem types) as dependent variable and TTR and pretest score on the surprise items as covariates. For the surprise items, we found a marginally significant main effect for Surprises ($F(1, 90) = 3.161$, $p = .079$). The main effect for Expectancy strength and the Surprises × Expectancy strength interaction were not significant (both $F(1, 90) < 1$). For the comparison items, we neither found main nor interaction effects (all $F < 1$).

Although all participants were selected from the same educational level, the population is still very heterogeneous which is reflected in the large *SD* (in each condition there are large differences). We assumed that better performing students would possess the (meta)cognitive skills to deal with the surprising events and benefit from the cognitive processes that they trigger. We divided the sample in low and high skill level students based on the median score of 6 on the pretest. We ran an ANCOVA with the posttest score on the surprise items as dependent variable; surprises, expectancy strength and skill level as fixed factors and TTR as covariate. We expected to see an interaction between surprises and skill level, indicating that high level students would benefit more from surprises than low level students. The results, however, only show significant main

Table 4: Mean scores and standard deviations on the dependent variable for all conditions of Experiment 2

	Surprises				*No surprises*			
	Strong expectancy		*Weak expectancy*		*Strong expectancy*		*Weak expectancy*	
TTR	*118 (30)*		*130 (20)*		*121 (27)*		*124 (31)*	
	Pre *M (SD)*	*Post* *M (SD)*	*Pre* *M (SD)*	*Post* *M (SD)*	*Pre* *M (SD)*	*Post* *M (SD)*	*Pre* *M (SD)*	*Post* *M (SD)*
All items [0–12]	5.71 (2.34)	6.90 (3.18)	6.00 (3.05)	6.95 (3.29)	5.07 (2.41)	5.53 (3.46)	5.56 (1.61)	5.86 (3.16)
Surprise items [0–8]	3.04 (1.88)	4.28 (2.23)	4.00 (2.41)	4.79 (2.53)	3.11 (2.41)	3.57 (2.14)	3.21 (1.44)	3.47 (2.44)
Comparison [0–4]	2.66 (1.06)	2.61 (.97)	2.00 (1.06)	2.17 (1.20)	1.96 (.95)	1.96 (1.18)	2.35 (1.07)	2.39 (1.40)
Transfer items [0–4]	.81 (.87)	.85 (1.31)	1.08 (1.34)	.83 (1.27)	.73 (.96)	.53 (1.10)	1.13 (1.17)	.69 (1.18)

Notes: Range of scores between []. All items mean all proportional reasoning skill items. Surprise items are missing value + transformation items.

effects for Surprises ($F(1, 85) = 4.120$, $p = .046$) and Skill Level ($F(1, 85) = 18.980$, $p = .000$), but all other main or interaction effects were not significant (Expectancy × Surprises: $F(1, 85) = 1.057$, $p > .05$; all other effects $F < 1$) (Figure 3).

General discussion

In two experiments, we found that proportional reasoning skills improve by playing the game. This corroborates earlier findings regarding serious games in general (cf., Wouters *et al.*, 2013) and other studies with the game *Zeldenrust* (ter Vrugte *et al.*, 2015; Vandercruysse *et al.*, submitted; Wouters *et al.*, 2015). There is some evidence that engaging in a game-based environment can improve transfer skills (eg, Barab *et al.*, 2009). However, we did not find evidence that proportional reasoning skills can be transferred to problems with different contexts in which it is more difficult to find the relevant characteristics of the problem. If that was the case it should be reflected in better performance in the transfer items. This implies that the game we used supports students to solve problems that are similar to those that they practiced with, but that it does not help them to acquire a deeper understanding and solve problems with a different context. Although motivation issues cannot be discarded (in both experiments students scored lower in the posttest of the transfer items compared to the pretest), a potential explanation can be found in the structure of the game. The repetitive character of the game *Zeldenrust* may support students in the automation of proportional reasoning skills, but it does not facilitate the acquisition of new insights. Maybe transfer skills can be further developed when students also are exposed to tasks in which the characteristics that define a proportional reasoning problem are presented in a different context. In both experiments, we also found that effective game play (the number of correct game tasks) is predictive for posttest performance (proportional reasoning items and to some degree for the transfer items).

In Experiment 1, we failed to find a clear beneficial effect of surprise although high educational level students did benefit. We provided two arguments for this finding. The students did not possess sufficient (meta)cognitive skills and the expectancy factor which is important for surprise was not optimally utilized. In Experiment 2, we operationalized these new demands by focusing on higher cognitive level students and by manipulating the strength of expectancy. We found a marginal effect of surprise on learning indicating that students who experienced surprise learned more than students who did not experience surprise but we found a stronger effect of surprise when we included existing proportional reasoning skill as factor. These results also imply that

Table 5: Hierarchical regression on posttest performance in Experiment 2

	Proportional reasoning			*Transfer*		
	B	*SE B*	*β*	*B*	*SE B*	*β*
Step 1: Constant	.25	1.50		.10	.61	
Pretest score	.67	.13	.50****	.49	.11	.45****
TTR	.02	.01	.15	.002	.005	.04
Step 2: Constant	.07	1.54		.24	.66	
Pretest score	.55	.13	.41****	.45	.11	.41****
TTR	.01	.01	.09	−.001	.005	−.027
Correct task	.10	.03	.34***	.034	.013	.31*
Time-on-task	−.001	.03	−.005	−.01	.01	−.14

Notes: *Proportional reasoning*: $R^2 = .33$ for Step 1, $\Delta R^2 = .10$ for Step 2.
Transfer: $R^2 = .22$ for Step 1, $\Delta R^2 = .06$ for Step 2.
*$p < .05$, **$p < .01$, ***$p < .005$, ****$p < .001$.

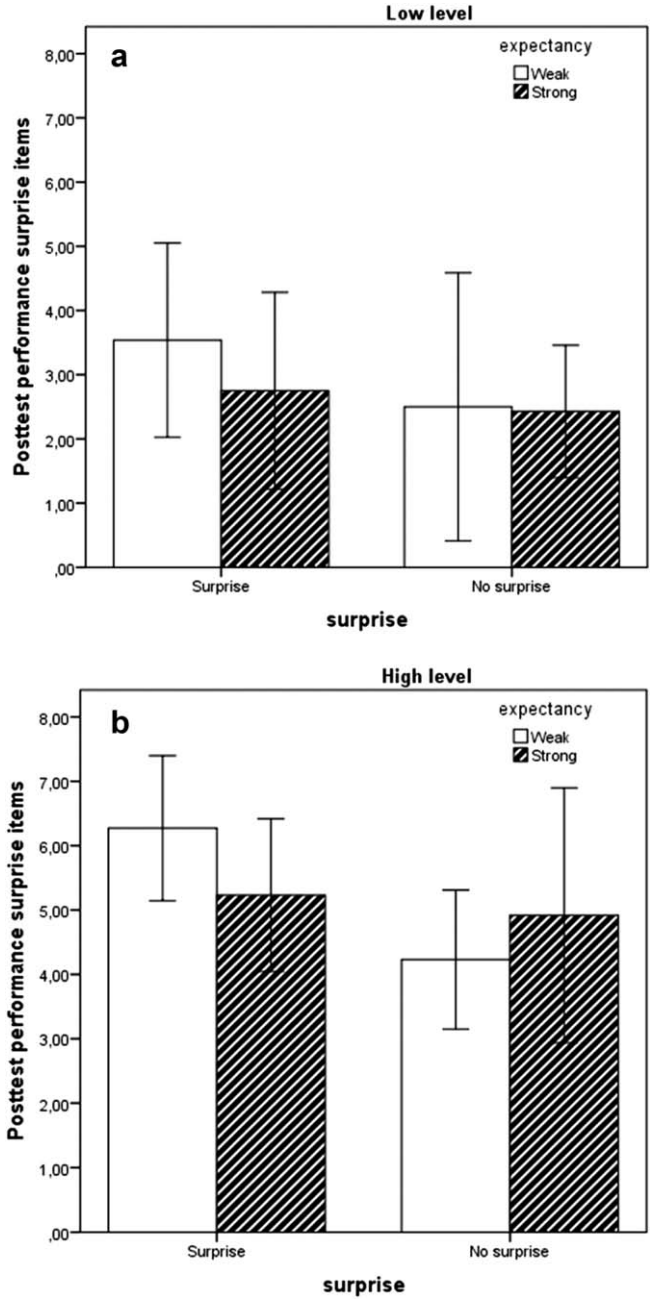

Figure 3: Posttest performance on surprise items (missing value and transformation) for low level students (a) and high level students (b)

instructional techniques like embedding surprising events should be applied with care. An important precondition for effective surprise is that players have sufficient cognitive flexibility and (meta)cognitive skills to orientate on the task, to re-evaluate the results and to reflect on the performed actions. This implies that designers of GBL should not present surprising events in order to enhance learning when students do not possess these skills. As mentioned before, surprising

events trigger a state of surprise. In complex learning environments such as computer games, both the events and the state of surprise may introduce additional cognitive demands that easily overwhelm students that do not possess sufficient (meta)cognitive capabilities. More research is required to investigate the robustness of the surprise effect and the underlying cognitive mechanisms. In particular, we propose further research that explores how different levels of metacognitive skills deal with surprise. These promising results connect with other studies that find positive cognitive effects of narrative techniques (eg, surprise, curiosity, suspense) in games (van der Spek *et al.*, 2013; Wouters, van Oostendorp, Boonekamp, & van der Spek, 2011).

The assumption in this study is that surprise will trigger students to think about the problem solving process which will improve their proportional reasoning skills. The surprise effectuated that a problem that first could be solved with a specific solution method changed in such a way that it became easier to use another method. Take, eg, the following situation just before the surprise: the desired proportion is 9 Cola/12 Fanta, the refrigerator contains 0 Cola/24 Fanta. In this case, a solution method based on intern ratio is most appropriate: $12 \times 2 = 24$, so also $9 \times 2 = 18$ Cola. After the surprise, the desired proportion has become 6 Cola/12 Fanta, the refrigerator contains 8 Cola/30 Fanta because the surprise is triggered after entering 8 Cola and the niece character added 6 additional Fanta bottles. The original solution method has to be replaced by a more effective method. In this case, a method based on the extern ratio: the number of Fanta is twice the number of Cola (6 Cola $\times$ 2 = 12 Fanta), the number of Fanta in the refrigerator should also be twice the number of Cola (so 15 bottles of Cola). As there are already eight bottles in the refrigerator, seven extra bottles of Cola have to be added. Although we did not determine whether students used another solution method after the occurrence of the surprising event, observations made during the experiment showed that many students solved the problems, regardless of the characteristics, most of the times with a single method (convert to 1). This may have weakened the intended effect of surprise: reconsidering a solution method and, if necessary, choose a more appropriate method. Possibly, the effect of surprise can be increased by offering students instructional support during the problems before the surprise intervention occurs which may help them to select an appropriate method for a problem. One could think of exercises that help them to automatize part-tasks such as multiplication tables so that they can more easily identify intern or extern ratios and/or worked examples in which solution methods for specific types of problems are modeled. We also propose replication studies in which the problem solving method that is used before and after the surprising event is administered so that can be examined whether students choose another method because of a change in problem characteristics.

Four other lines of research can also be interesting. First, there is some evidence that (meta)cognitive skills in math improve with small differences in age (van der Stel, Veenman, Deelen, & Haenen, 2010). The students in the current study came from the 2nd-year class (mean age of 13.9–14.1 years) and the (meta)cognitive skills of some students may have been insufficiently developed. Another point is that the students come from the least advanced of three Dutch educational tracks in which students are prepared for intermediate vocational education. It would be interesting to replicate this study with older students in the same educational level (3rd or 4th year class) or students from a higher educational track. A second research avenue pertains to the characteristics of the game. The game *Zeldenrust* has a repetitive character, students engage in the same type of tasks which require similar actions. It is not unlikely that students finally will expect that the niece character—the embodiment of the surprise—will reappear and modify the nature of the task. In that case, they may anticipate these events and thus undermine the potential effect of surprise. If that is the case more variation in surprise can perhaps further increase their effectiveness. A third interesting prospect is a further investigation of the role of different characteristics of students. We found that the possession of (meta)cognitive skills is such a

characteristic. It is not excluded that other student characteristics play a role as well (eg, self-efficacy and attitude toward the domain). A final topic of further research concerns the role of motivation. We assumed that narrative techniques as embedding surprising events can also have a positive effect on motivation, and because of that also on learning. In our studies we focused on learning so we cannot confirm the indirect positive effect of surprise on learning. Future studies should look closer into the role of motivation triggered by these techniques.

Acknowledgement
This research is funded by the Netherlands Organization for Scientific Research (project Number No. 411-00-003).

Statements on open data, ethics and conflicts of interest
We have no institute respository. Requests for the data (together with a short description of the purpose of the request) can be addressed to the corresponding author. Approvement has to be given by the project partners and NWO (Netherlands Organization for Scientific Research).

The experiments were conducted at several schools with the approval of the school authority and teachers. Students were made anonymous by using id's instead of names.

There is no conflict of interest.

References

Adler, J. E. (2008). Surprise. *Educational Theory*, *58*(2), 149–173.

Barab, S. A., Scott, B., Siyahhan, S., Goldstone, R., Ingram-Goble, A., Zuiker, S. J. *et al.* (2009). Transformational play as a curricular scaffold: using videogames to support science education. *Journal of Science Education and Technology*, *18*, 305–320.

Brewer, W. F., & Lichtenstein, E. H. (1982). Stories are to entertain: a structural-affect theory of stories. *Journal of Pragmatics*, *6*(5), 473–486.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & La Vancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2015). Digital games, design, and learning a systematic review and meta-analysis. *Review of Educational Research*. doi: 10.3102/0034654315582065.

Clement, J. J., & Steinberg, M. S. (2002). Step-wise evolution of mental models of electric circuits: a "learning-aloud" case study. *The Journal of the Learning Sciences*, *11*(4), 389–452.

de Vos, T. (1992). *Tempo test rekenen handleiding en verantwoording [Tempo test arithmetic manual and justification]*. Amsterdam: Pearson.

Foster, M. I., & Keane, M. T. (2015a). Predicting surprise judgments from explanation graphs. In *International Conference on Cognitive Modeling (ICCM)*. Groningen, The Netherlands: Groningen University.

Foster, M. I., & Keane, M. T. (2015b). Why some surprises are more surprising than others: surprise as a metacognitive sense of explanatory difficulty. *Cognitive Psychology*, *81*, 74–116.

Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: exploring the value of intrinsic integration in educational games. *Journal of the Learning Sciences*, *20*, 169–206.

Hoeken, H., & van Vliet, M. (2000). Suspense, curiosity, and surprise: how discourse structure influences the affective and cognitive processing of a story. *Poetics*, *26*, 277–286.

Howard-Jones, P., & Demetriou, S. (2009). Uncertainty and engagement with learning games. *Instructional Science*, *37*(6), 519–536.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*, 1295–1306.

Leemkuil, H., & de Jong, T. (2011). Instructional support in games. In S. Tobias, & D. Fletcher (Eds), *Computer games and instruction* (pp. 353–369). Charlotte, NC: Information Age Publishing.

Maguire, R., Maguire, P., & Keane, M. T. (2011). Making sense of surprise: an investigation of the factors influencing surprise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 176–186.

Mayer, R. E. (2011). Multimedia learning and games. In S. Tobias, & J. D. Fletcher (Eds), *Computer games and instruction* (pp. 281–305). Greenwich, CT: Information Age Publishing.

O'Brien, E. J., & Myers, J. L. (1985). When comprehension difficulty improves memory for text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 12–21.

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, *4*, 193–203.

Rick, J., Bejan, A., Roche, C., & Weinberger, A. (2012). Proportion: learning proportional reasoning together. In A. Ravenscroft, S. Lindstaedt, C. D. Kloos, & D. Hernández-Leo (Eds), *Lecture notes in computer science: volume 7563, 21st Century learning for 21st century skills* (pp. 513–518). Berlin, Germany: Springer.

Rünger, D., & Frensch, P. A. (2008). How incidental sequence learning creates reportable knowledge: the role of unexpected events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1011–1026.

ter Vrugte, J., de Jong, T., Wouters, P., Vandercruysse, S., Elen, J., & van Oostendorp, H. (2015). When a game supports prevocational math education but integrated reflection does not. *Journal of Computer Assisted Learning*, *31*(5), 462–480.

ter Vrugte, J., de Jong, T., Wouters, P., Vandercruysse, S., Elen, J., & van Oostendorp, H. (submitted). How heterogeneous collaboration and competition interact in prevocational game-based math education.

Tourniaire, F., & Pulos, S. (1985). Proportional reasoning: a review of the literature. *Educational Studies in Mathematics*, *16*, 181–204.

Vandercruysse, S., ter Vrugte, J., de Jong, T., Wouters, P., van Oostendorp, H., & Elen, J. (2015). 'Zeldenrust': a mathematical game-based learning environment for vocational students. In J. Torbeyns, E. Lehtinen, & J. Elen (Eds), *Describing and studying domain-specific serious games* (pp. 63–81). New York, NY: Springer.

Vandercruysse, S., ter Vrugte, J., de Jong, T., Wouters, P., van Oostendorp, H., & Elen, J. (submitted). Content integration as a factor in math-game effectiveness.

van der Spek, E. D., van Oostendorp, H., & Meyer, J.-J. Ch. (2013). Introducing surprising events can stimulate deep learning in a serious game. *British Journal of Educational Technology*, *44*, 156–169.

van der Stel, M., Veenman, M. V., Deelen, K., & Haenen, J. (2010). The increasing role of metacognitive skills in math: a cross-sectional study from a developmental perspective. *ZDM - The International Journal on Mathematics Education*, *42*(2), 219–229.

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.

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

Wouters, P., van Oostendorp, H., Boonekamp, R., & van der Spek, E. (2011). The role of Game Discourse Analysis and curiosity in creating engaging and effective serious games by implementing a back story and foreshadowing. *Interacting with Computers*, *23*(4), 329–336.

Wouters, P., van Oostendorp, H., ter Vrugte, J., Vandercruysse, S., de Jong, T., & Elen, J. (2015). The role of curiosity-triggering events in game-based learning for mathematics. In J. Torbeyns, E. Lehtinen, & J. Elen (Eds), *Describing and studying domain-specific serious games* (pp. 191–207). New York, NY: Springer.

Wouters, P., Paas, F., & van Merriënboer, J. J. M. (2008). How to optimize learning from animated models: a review of guidelines base on cognitive load. *Review of Educational Research*, *78*, 645–675.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: an event-indexing model. *Psychological Science*, *6*(5), 292–297.