Worked Examples in Game-based Learning for Primary Education

Laura M. H. van Oostrom

Supervisor Henny H. Leemkuil

Keywords: learning, feedback, fun, mental effort, prior knowledge

Abstract

The current study focusses on which combination of worked example (worked examples with self-explanation, worked examples with given explanation, and no worked examples) as feedback and level of prior knowledge increases learning without compromising the fun the learners are having or increasing the mental effort it takes to complete the lesson. A game with worked examples as feedback with self-explanation, given explanations, or no worked example was used. The pre- and post-test scores of the different conditions are compared with each other and with the having fun and mental effort scores. Unfortunately, no fourth grade participants volunteered to take part in the experiment and the main question had to be limited to the worked examples without the combination with the level of prior knowledge. The results showed that all three conditions performed similarly on learning gain, having fun and mental effort. Which indicates that using games without worked examples are most efficient for learning, because those use less time and smaller activities.

Table of content

Introduction

Would it not be perfect if we could combine learning with having fun? Baltra (1990) states that playing is a form of learning especially often used with young children. In other words, there is a way to combine learning and having fun, namely game-based learning. Similar terms, like serious games and educational games, all have in common that educational goals are met through playing a game, but what identifies a game? Juul (2003) states that games have six characteristics based on which they are called a game: a system of rules, quantifiable outcome, different outcomes have different values, player feels attached to the outcome, and consequences of the activity are negotiable (Juul, 2003). Salen and Zimmerman (2003) confirm in their definition that a system of rules and a quantifiable outcome are important in defining games, but also stress the importance of the presence of an artificial conflict. So, games need to have at least a system of rules and a quantifiable outcome. By playing the game and following the rules, certain goals can be accomplished. According to Wouters, Van Nimwegen, Van Oostendorp, & Van der Spek (2013) the goal of serious games is, first and foremost, to have fun. However, according to Michael and Chen (2006) the goal is education. For the current study, both goals will be seen as equally important.

The problem with game-based learning is twofold: Literature is inconclusive and mixed about the effectiveness for learning (De Freitas, 2006) and often the more the players are learning, the less they are having fun, and vice versa (Kim, Park, & Baek, 2009). The use of instructional support, like providing feedback or scaffolding, in game-based learning can improve learning (Wouters & Van Oostendorp, 2012) because the support helps the learners use their cognitive capacity more optimally (Mayer & Moreno, 2003). On the other hand, scaffolds to improve the effectiveness for learning seem to reduce the fun the learners are having while playing (Broza & Barzilai, 2011).

Game-based learning increases in popularity (Mortara, Catalano, Bellotti, Fiucci, Houry-Panchetti, & Petridis, 2014), but still has a twofold problem that shows that game-based learning has not reached its full potential. Doing more research to lessen this problem seems more important now than ever. As described above, the problem of game-based learning is a combination of ineffectiveness for learning (De Freitas, 2006) and improving it is likely to compromise the fun the learners are having (Wouters et al., 2013). Worked examples have

already been proven effective as part of an instruction as well as part of feedback (Paas, 1992; Paas & Merriënboer, 1994), but how about in a game-environment? By testing findings of studies about learning support, like worked examples, and guidance in serious games on effectiveness for learning and the fun learners are having, scientists can increase the knowledge about how to design more effective serious games, both effectiveness for learning as well as having fun. The more effective serious games that can be designed and published because of this, are beneficial for practitioners like game designers and school teachers. Game designers can improve their design choices and implement them on existing and new serious games. School teachers can implement serious games in their teaching, or when they already have, can replace the serious games they are already using with more effective serious games.

An example of a game-like environment is Rekentuin (published by Oefenweb). Rekentuin is actively used by more than 400.000 Dutch primary school children (Brinkhuis, Savi, Hofman, Coomans, Van der Maas, & Maris, 2018). Interestingly enough, it uses very little learning support like feedback. The learners merely receive a statement of correctness and speed of answering (Meijer & Karssen, 2013), but nothing about the calculation of the answer or the process behind it (Klinkenberg, Straatemeier, & Van der Maas, 2009). An example of instructional support that can be added are worked examples. These reduce the extraneous load and leave more working memory capacity to process information (Sweller, 1988). Furthermore, worked examples seem very effective, both in instruction as in feedback (Paas, 1992; Paas & Merriënboer, 1994). Example-problem pairs (worked example as instruction) are most beneficial for learners with little prior knowledge and problem-example pairs (worked example as feedback) are most beneficial for learners with a lot of prior knowledge (Reisslein, Atkinson, Seeling, & Reisslein, 2006).

Mayer and Wittrock (1996) state that worked examples explain about the problem, the goals and the steps in between, but adding in the explanation of the process behind the steps is necessary for transfer of the solution to similar problems (Mayer & Wittrock, 1996). According to Chi, Bassok, Lewis, Reimann, and Glaser (1989), the most useful form of explanation for transfer for students with high prior knowledge is self-explanation. Johnson and Mayer (2010) studied self-explanation in a computer-based game-like environment. They found that *selection self-explanation* (choosing a reason out of a list of nine reasons), is more effective than

generation self-explanation (writing down a reason) or no self-explanation at all. Also, the selection of a reason provokes reflection without influencing the flow while playing (Johnson & Mayer, 2010). Self-explanation is only effective if the learners are able to self-explain well and have high prior knowledge. For learners with low prior knowledge a given explanation of the process behind the steps of the worked example is useful (Chi et al., 1989).

The current study extends the studies mentioned above by testing the combination of the level of prior knowledge and three conditions (worked examples with self-explanation, worked examples with given explanation, and no worked examples) in the context of a game while taking into account having fun and mental effort. The goal of this study is to add to the knowledge base about instructional support in serious games and to help designers improve serious games.

Theoretical framework

The worked example effect can be explained by the cognitive load theory. This theory states that there are three types of loads: The intrinsic load is caused by the complexity of the task itself and depends on the amount of prior knowledge of the learner. The higher the prior knowledge is, the more relevant schemas exist in long-term memory and the lower the intrinsic load is. The extraneous and germane load are caused by the design of the instruction and can be divided into germane load which is useful for learning and extraneous which is not (Sweller, Van Merriënboer, & Paas, 1998). Worked examples lower the extraneous load relative to the load of problem solving exercises because the learner can focus more working memory capacity on schema construction and automation and storing it in the long-term memory (Sweller, 1988). The germane load can be increased by the way the worked examples get presented to the learner, for example by increasing variability (Paas & Van Merriënboer, 1994).

Worked examples can be presented in the form of videos, which are motivating and help with memorising and visualizing of content (Mateer, 2011). Also, videos avoid extra cognitive load by avoiding the need to read lengthy written texts (Mayer, 2014). The videos should be kept short in order to not overload the learner with information (Van der Meij & Van der Meij, 2013), should use spoken words instead of written words, and associated words and pictures should be placed as close as possible to and in sync with each other (Mayer, 2014).

In the current study the worked examples are used as feedback instead of instruction. Feedback is knowledge about the gap between where a student is now in its learning and where it should be according to the learning goals, and this knowledge is used to alter this gap (Ramaprasad, 1983). Formative feedback is knowledge intended for the learners to improve their thinking and/or behaviour in order to improve learning (Shute, 2008). Formative feedback can help to lower the cognitive load, because it can support a learner through an overwhelming learning assignment, as stated in a study using worked examples (Sweller et al., 1998) and learners with low prior knowledge (Moreno, 2004). In addition, a benefit from using a worked example as feedback is that learners have already tried a problem in which they experienced their shortcomings and so they might be more motivated to find answers in the example (Reisslein et al, 2006).

The current study will not only measure the effect of worked examples as feedback in the context of a serious game in learning gains, but also in mental effort and having fun. It is interesting to know how much effort it takes learners to learn something, because if there seem to be two ways which are equally effective for learning, it is more learner-friendly to use the way with the least effort to get the same learning outcome (Paas & Van Gog, 2006). Practice and more automated schemas of a learner, reduce the cognitive load in the working memory, and therefore cost less mental effort to solve the problem (Paas & Van Merriënboer, 1993).

As said before, having fun is an important goal of a serious game (Wouters et al., 2013). Fun is a combination of provoking engagement or captivation and provoking new or unusual emotions (Carroll, 2004) which are pleasurable (Sim, MacFarlane, & Read, 2006). Scaffolds to improve the effectiveness of the game in terms of learning seem to reduce the fun the learners are having while playing (Broza & Barzilai, 2011).

Research questions

Main question: Which combination of type of worked example as feedback and level of prior knowledge has the biggest influence on learning, compromises the fun the students are having the least and influences the mental effort the students have to put in the most? The main and three sub questions will be checked for between groups intelligence differences by adding the grade the participant is in as a covariate in the analysis.

The first sub question is ‘does the addition of the worked examples to the game compromise the fun the learners are having?’. It is hypothesized that the addition of the worked example to the game will compromise the fun the learners are having, because the worked examples will be interrupting the game (Barzilai & Blau, 2014).

The second sub question is ‘does the addition of the worked examples influence the mental effort needed to play the game?’. Worked examples reduce the extraneous load and help improve a learners schemas (Sweller, 1988), so it is expected that the worked example conditions need less mental effort for the same amount of learning than the condition without the worked examples.

The third sub question is ‘do the worked examples have different effects on learners with high and low prior knowledge?’. It is hypothesized that the worked example with the self-explanation is best in combination with high prior knowledge and the worked example with the given explanation with low prior knowledge (Chi et al., 1989).

Method

Research design

The current study used a quasi-experimental design. The participants were gathered with a convenient sample and randomly divided over the conditions per class instead of in the whole sample. The conditions include two types of worked example conditions and a control condition without worked examples. A pre-test post-test design with six conditions, see Table 1, was used in combination with two surveys. This design fits the research questions well, because the best performing condition(s) can be recognized by comparing the conditions on the several factors.

Table 1

Overview of the Six Conditions: Type of Worked Example x Level of Prior Knowledge.

	High prior knowledge	Low prior knowledge
WE with self-explanation	A	D
WE with given explanation	B	E
No WE	C	F

Respondents

The participants were supposed to be learners from third and fourth grade of Dutch primary schools (in Dutch: *groep 5* and *groep 6*). The participants were recruited by asking schools to let their classes participate and offering them the free game-environment and a class overview. Unfortunately, no fourth grade classes volunteered to participate in the current study. The sample of $N = 86$ consists of 41 boys (47.7%) and 45 girls (52.3%) of 8 to 11 years of age ($M = 8.65$, $SD = 0.58$) from two different schools. All participants were able to take both a pre- and a post-test and fill in the fun survey. The participants were evenly divided over the three conditions: Worked examples with self-explanation ($n = 29$), worked examples with given explanation ($n = 29$), and no worked examples ($n = 28$). The participants were also evenly divided over the two orders of the tests: pre-test A and post-test B ($n = 43$) and pre-test B and post-test A ($n = 43$). Unfortunately, the mental effort scores were incomplete and ten participants were excluded from the sample for research questions with one or more mental effort variables. The remaining participants were a combination of participants with data from all mental effort scores and participants who are missing one or more of the six mental effort scores, in both cases an average mental effort score was computed. In the analyses the participants with missing data were excluded listwise. The reasons for the missing data are threefold: First, the participants could not have answered the survey questions, so data was missing. Secondly, due to software problems not all answers were saved and data was lost. Thirdly, due to software problems participants were able to click on multiple answers. Which were coded as missing data, because there was no way of knowing which answer the participant meant to give.

Instruments

Content. The subject of the game is solving problems about nature and technology for which the learners need to multiply and divide. Core goal 42 (TULE, n.d.) is about researching phenomena and materials on the subject of nature and technology. Part of core goal 27 (TULE, n.d.) for third and fourth grade (in Dutch: *groep 5 and 6*) describes multiplying with often used and useful numbers, for example 2 x 50 and 4 x 15. In the game, there can be multiplications from 1 to 8 from the marks and 5, 10, 15, .. 95 from the weights of the persons and objects. So, the level of the game seems compatible with the goals for multiplying in third and fourth grade. Furthermore, these core goals are chosen because probably the majority of the learners in third

grade will have low prior knowledge on this subject and the majority of the learners in fourth grade will have high prior knowledge.

Pre- and post-test. To test the participants' (prior) knowledge, two different versions of a test were used. The participants did both versions of the tests once, either test A as pre-test and test B as post-test or the other way around. This is done to counterbalance the possible difficulty difference between the tests and to prevent boosting test scores because of familiarity with the test questions.

The game used three types of questions: 'balance me', 'what will happen?', and 'what is the mass?'. The tests contained four questions of each of the three types. Also, test items with multiple weights were included in the tests (50% of the questions equally divided over the three types of questions) to be more certain that the tests were difficult enough to best prevent the ceiling effect (Austin & Brunner, 2003) and all learners are able to show their learning gains.

The reliability analysis of test A showed a Cronbach's Alpha of .661, which is quite high. The seventh question of test A can be excluded in order to increase the reliability (Cronbach's Alpha if item deleted = .672), but the seventh question is not excluded based on the validity of the test. The reliability analysis of test B showed a Cronbach's Alpha of .756, which is high. No questions could be excluded to increase the reliability.

At the end of the last test, the fun survey was used to measure the amount of fun the participants were having while playing the game on the Funometer scale (Read, MacFarlane, & Casey, 2002). The Funometer seems most useful for the current study, because it matches best the target group. See Appendix for the overview of the questions. The reliability of the fun survey is not tested, because an existing scale was used which is already proven to be reliable.

Game. The game from the Phet lab *Balancing Act* was used, instead of the Rekentuin mentioned above. This game is similar to the Rekentuin games, but has a smaller chance of participants being familiar with it and thus reduces unwanted influences on the results of the current study. The game meets the two characteristics of a game where the two studies (Juul, 2003; Salen & Zimmerman, 2003) mentioned above agreed upon; a system of rules and a quantifiable outcome. Also, the characteristic of the player feeling attached to the outcome was met. A few participants even asked if they could retake a level of the game to obtain a better score.

The game had four levels with increasing difficulty. The game is identical for all three conditions, just the worked examples differ. Between level 1 and 2 and between level 4 and 4 again, a worked example is used as feedback to an exercise question for all three types of questions. Level 4 is played twice instead of level 3 and 4, because the difference between level 2 and 3 is neglectable and in level 4 exercises with multiple weights are introduced. The worked examples with self-explanation let the participants choose from a list of options instead of writing the explanation themselves, like Johnson and Mayer's (2010) study. This study uses nine options to choose from. In the current study this is reduced to three options in order to lessen the amount of reading and lessen the difficulty of the task. Just like the studies of Paas (1992) and Paas and Merriënboer (1994), the current study has worked examples identical to the problems.

A one-question survey after each worked example was used to measure the mental effort the participants have experienced. Both the question as the scale are copied from Paas's (1992) study. Paas's (1992) 9-point mental effort rating scale is stated to be very accurate at measuring the cognitive load brought to the learner by the task (Paas, Tuovinen, Tabbers, & Van Gerven, 2003). This survey contains one question 'In solving or studying the preceding problem I invested..', which is asked directly after the worked examples. For this experiment, the question is translated to Dutch. The reliability of the mental effort survey is not tested, because an existing scale and survey question were used which are already proven to be reliable.

Pilot

For the game-environment and the pre- and post-tests, a pilot was held to test the clarity, quality and difficulty. This was done with some volunteers, who tested the materials by giving data on their performance on the tests, mention what they did not understand about the tests, and mention what they thought of the tests (MacFarlane, Read, Höysniemi, & Markopoulos, 2003).

Based on the volunteers of the pilot of the digital environment a few small changes were made to the learning environment. First, the problem of data loss due to software problems became clear and it was decided to take the tests on paper instead of in the learning environment to decrease the amount of missing data. The second change was based on a strong fourth grade child who mentioned and showed that the content of the game and tests were pretty hard for the target group. In order to lessen this problem without having to make changes to the existing game, the worked examples were improved and expanded so much that the choice was made to

present them in the form of a set of videos. The last small change that was made based on the information gathered from the pilot was a change in the time planning of the experiment. The initial plan was to take the pre-test for half an hour, work in the learning environment for 45 minutes to 1 hour, and then take the post-test for half an hour. According to the volunteers, the half an hour for the tests was very long and 45 minutes for the learning environment was too short. The new time planning was adjusted to 15 minutes for the tests and at least an hour to work in the learning environment.

Procedure

Before conducting the experiment, the study was explained to the parents of the participants and consent was asked from the parents. Only the consent of the parents of the participants was asked by using a consent form, because the participants are under 12 years of age.

The experiment started with the pre-test, half of the participants took test A and the other half test B. The pre-test output is a score which was used in the analyses to form groups based on prior knowledge. The participating classes were split randomly in one third receiving worked examples with self-explanation, one third receiving worked examples with given explanations, and the last third receiving no worked examples. The participants continued the experiment in rounds behind a computer. The digital learning environment started with an instruction video about the game. The video was followed by a demographic variables questionnaire, which asked for the age, grade and gender of the participant. Then the participants all played the four levels of the game: level 1, 2, 4, and 4 again. Between level 1 and 2 and between level 4 and 4 again three practice questions with or without worked examples were presented to the participants. After each practice question with or without worked example, the participants were asked how much effort it took them to answer using the one-question mental effort survey. The participants took the post-test, accordingly version A or B, accompanied by the survey about having fun. Within a week after the experiment was conducted, the teachers received the participants tests scores and access to the three versions of the game-environment. The teachers also received (the summary of) the research article.

Video

The worked examples were presented in the form of a set of videos, because videos are

motivating, and help with memorising and visualizing the content of the game (Mateer, 2011). Also, and this is the main reason why videos were used, videos avoid extra cognitive load by avoiding reading lengthy written texts (Mayer, 2014). To guarantee a higher quality of the videos, the guidelines of Van der Meij and Van der Meij (2013) and the principles of Mayer (2014) were used while making the videos. A few important ones will be explained below.

Van der Meij and Van der Meij's (2013) guideline of keeping the videos short is met, because the longest video is around one minute long. Mayer's (2014) principles of modality and split attention are met. The first is met by using spoken words in the videos to share information to the participants instead of written words. The second is met by synchronizing the narration with the animation. The animation consisted of screenshots of the game and the practice questions while using highlighters and pointers to help focus the attention on the important parts of the video. This all was in sync with the spoken words, according to the temporal contiguity principle.

Data analysis

Due to the lack of fourth grade students, the original design for the main question cannot include the high (mostly fourth graders) and low (mostly third graders) prior knowledge anymore. Therefore, the design is altered to no longer include the variable of prior knowledge. The new main question is: 'Which type of worked example as feedback has the biggest influence on learning, compromises the fun the students are having the least and influences the mental effort the students have to put in the most?'. This question fits with a MANOVA. This test is conducted with the three conditions as independent variable and the mental effort score, the having fun score, and the learning gains as the dependent variables.

Before conducting this test, the assumptions were checked. The assumption of independence is met, because the participants are all participating only once in the experiment. It is assumed that the participants did not collaborate or share answers, because the participants were not allowed to touch someone else's laptop and worked in silence. The assumption of cell size is also met, because each cell has at least 28 participants. The assumption of normality is partly violated. The learning gains and mental effort scores are normally distributed, but the fun scores are not (worked examples with self-explanation $p = .002$, worked examples with given explanation $p = .014$, and no worked examples $p = .003$). ANOVA is quite robust against mild

violations of the assumption of normality, so the test can still be conducted. The assumption of multicollinearity is partly violated. The learning gains and fun scores are correlated, $r(74) = .02$, $p = .428$. The other two combinations were not correlated. The Maximum Mahalanobis Distance is only 12.085, which is smaller than the 16.266 for three degrees of freedom, and so indicates the absence of multivariate outliers. The assumption of linearity is met because all three combinations of variables seemed roughly linear. The assumption of homogeneity of variance-covariance is not violated. Both the Box's test of equality of covariance matrices ($p = .547$) and the Levene's test of equality of error variances (worked examples with self-explanation i = .376, worked examples with given explanation $p = .812$, and no worked examples $p = .788$) are not significant.

The first sub question about the effect of the worked examples on having fun is answered with a one-way between groups ANOVA with the three conditions as the independent variable and the having fun score as the dependent variable. First, the assumptions were checked. The assumption of scale of measurement is met because the dependent variable fun score is interval or ratio data. The assumption of independence is also met, each participant is participating only once in the experiment. The assumptions of normality and normality of difference scores are not violated. The Skewness and Kurtosis statistics suggests that the differences between fun score 1 and fun score 2 are approximately normal because the scores are fairly close to zero. Also, the relevant histograms seems normally distributed after a visual inspection. The assumption of homogeneity of variances is not violated, because the Levene's statistic is not significant at $\alpha = .05$ ($F = .129$, $p = .879$).

The second sub question about the effect of worked examples on mental effort is answered with a one-way between groups ANOVA with the three conditions as the independent variable and the mental effort score as the dependent variable. First, the assumptions were checked. The assumption of scale of measurement is met because the dependent variable is interval or ratio data. The assumption of independence is also met, each participant is participating only once in the experiment. The assumptions of normality and normality of difference scores are not violated. The Skewness and Kurtosis statistics suggests that the average mental effort scores are approximately normal because the scores are fairly close to zero. Also, the Shapiro-Wilk statistic suggests that the data is normally distributed, because the statistics are

not significant for all groups. The assumption of homogeneity of variances is not violated, because the Levene's statistic is not significant at $\alpha = .05$ ($F = .239$, $p = .788$).

The third sub question cannot be answered because the design included fourth grade participants, which did not volunteered to participate in the current study. In order to still answer the third sub question, the prior knowledge variable in the question is changed from high (mostly fourth grade participants) and low (mostly third grade participants) to three levels within the group of third grade participants. There is chosen for three levels instead of two, because both the most frequently obtained score and the median score of the pre-test are at three correctly answered questions. In order to prevent very unevenly divided cell sizes, a third level of prior knowledge is added into the conditions, namely the medium level prior knowledge. So, the three levels of prior knowledge were organized based on the pre-test scores and the cut scores are located between two and three and between three and four correctly answered questions to make three reasonably sized subgroups (low $n = 31$, medium $n = 23$, and high $n = 32$ prior knowledge). For the conditions of worked examples in combination with level of prior knowledge the conditions are optimally sized. The smallest prior knowledge subgroup (medium prior knowledge $n = 23$) is most equally divided into two cells of eight and one cell of seven participants. The rest of the cells have up to 12 participants. Even though the cells are optimally sized, they are very limited and the answer on this research question can only be an estimation. The cell sizes are rather small and a discussion of the worthiness of mentioning the estimation in the current paper is justly. The reason why this estimation is still mentioned in this paper is because of the interestingness of the outcome of the estimation and its potential for future research.

The third sub question is answered with a factorial between groups ANOVA with the three conditions and level of prior knowledge as independent variables and the learning gains as the dependent variable. The difference score between the pre-test and post-test (called learning gain) is used for analysing this research question instead of the repeated measures (pre- and post-test scores), because of the convenience for analysing due to the amount of variables. Before conducting this test, the assumptions were checked. The assumption of scale of measurement is not violated, because the dependent variable learning gain is interval or ratio data. The assumption of independence is also not violated, because each participant participates only once in the experiment. The assumption of normality is partly violated, because all groups of the 3x3

design are not significant on the Shapiro-Wilk statistic except the worked examples with given explanation in combination with low prior knowledge (p = .034). ANOVA is quite robust against mild violations of the assumption of normality, so the factorial between groups ANOVA is still conducted. The assumption of homogeneity of variance is not violated, because the Levene's test showed a not significant result at α = .05, $F(8, 77) = 0.99$, p = .451.

The control question about the difference between the fun score of the seesaw-game and educational games in general is tested with a paired sample t test. Before conducting this test, the assumptions were checked. The assumptions of normality and normality of difference scores are not violated. The Skewness and Kurtosis statistics suggests that the differences between fun score 1 and 2 are approximately normal because the scores are fairly close to zero. Also, the relevant histograms seems normally distributed after a visual inspection.

Results

Main question

A multivariate analysis of variance (MANOVA) was performed to examine the influence of the three conditions (worked examples with self-explanation, worked examples with given explanation, and no worked examples) on learning gain, having fun and mental effort (N = 86). The MANOVA was not significant, F (6, 144) = 0.551, p = .769, $\eta 2$ = .022, which indicates that there are no differences between the three worked example conditions. The means for each group on each variable are shown in Table 2.

Table 2

Descriptive Statistics for Worked Examples with Self-Explanation, Worked Examples with Given Explanation, and No Worked Examples on Each Dependent Variable and the Pre- and Post-test.

Dependent variable	Type of worked example	*n*	*M*	*SD*
Learning gain *	WE with self-explanation	29	1.48	3.17
(12 questions)	WE with given explanation	29	1.79	2.64
	No WE	28	1.68	2.65
Having fun	WE with self-explanation	29	3.72	1.28
(5-point scale)	WE with given explanation	29	3.69	1.17
	No WE	28	3.46	1.37
Mental effort	WE with self-explanation	25	4.86	2.06
(9-point scale)	WE with given explanation	26	4.08	1.69
	No WE	25	4.56	1.88
Pre-test **	WE with self-explanation	29	3.07	1.79
	WE with given explanation	29	3.41	1.52
	No WE	28	2.86	1.65
Post-test **	WE with self-explanation	29	4.55	2.94
	WE with given explanation	29	5.21	2.98
	No WE	28	4.54	2.06

Notes. * Learning gain is measured by computing the difference between the amount of correctly answered questions on the pre- and post-test, both with 12 questions.
** The pre- and post-test scores are the amount of correctly answered questions out of the 12 questions per test in total.

Sub question 1

A one-way between groups ANOVA was performed to investigate whether the different conditions compromise the fun the players are having while playing the game. For this analysis the dependent variable FunDiff is used, which is computed by subtracting the fun score of the seesaw-game from the fun scores of educational games in general. The ANOVA was not

significant, $F(2, 83) = 1.55$, $p = .219$, indicating that there is no difference in perceived fun the participants were having between the three conditions. The means of the fun scores are shown in Table 3 for each type of worked example.

Table 3

Descriptive Statistics for Worked Examples with Self-Explanation, Worked Examples with Given Explanation, and No Worked Examples on the Dependent Variable Having Fun Measured on a 5-point Likert Scale.

		Fun1 general		Fun2 seesaw		FunDiff	
Type of worked example	*n*	*M*	*SD*	*M*	*SD*	*M*	*SD*
WE with self-explanation	29	3.97	1.05	3.72	1.28	0.24	1.15
WE with given explanation	29	4.00	0.96	3.69	1.17	0.31	1.14
No WE	28	4.18	1.09	3.46	1.37	0.71	0.98

Sub question 2

A one-way between groups ANOVA was performed to investigate whether the worked examples influenced the mental effort the players put in while playing the game. For this analysis the dependent variable MEavg is used, which is the computed average mental effort score of a participant throughout the experiment. The ANOVA was not significant, $F(2, 73) = 1.11$, $p = .336$, indicating that there is no difference in mental effort the participants had to put in between the three conditions. The means of the average mental effort scores are shown in Table 2 for each type of worked example.

Sub question 3

A factorial between groups ANOVA was performed to investigate whether there is a difference in learning gain between the three conditions in combination with the level of prior knowledge of the participants. Although the sample size of this experiment is not large enough to make a meaningful conclusion about the differences between the groups, the sample sizes per condition are similar and big enough to make an estimation.

The factorial between groups ANOVA revealed no significant interaction effect between prior knowledge and type of worked example, $F(4, 77) = 0.87$, $p = .484$. This indicates that there is not a specific combination of level of prior knowledge and one of the three conditions (worked examples with self-explanation, worked examples with given explanation, and no worked

examples) that leads to higher learning gain. Which is interesting about these results, even though they are not significant differences, is that the learning gain means, see Table 4, show the exact opposite of what was expected: It was expected that students with low prior knowledge would benefit most from the worked examples with the given explanations and students with high prior knowledge would benefit most from worked examples with self-explanation, but these combinations scored a very low average learning gain for the students with a similar level of prior knowledge.

Table 4

Average Learning Gains on a Scale of -12 to 12 Correctly Answered Questions on the Post-test compared to the Pre-test for the Nine Conditions.*

	Worked examples with self-explanation	Worked examples with given explanation	No worked examples
Low prior knowledge	$M = 3.10$ ($SD = 3.14$, $n = 10$)	$M = 2.56$ ($SD = 1.81$, $n = 9$)	$M = 3.50$ ($SD = 2.36$, $n = 12$)
Medium prior knowledge	$M = 1.13$ ($SD = 1.55$, $n = 8$)	$M = 1.25$ ($SD = 2.71$, $n = 8$)	$M = 1.29$ ($SD = 2.43$, $n = 7$)
High prior knowledge	$M = 0.27$ ($SD = 3.64$, $n = 11$)	$M = 1.58$ ($SD = 3.15$, $n = 12$)	$M = -0.44$ ($SD = 1.24$, $n = 9$)

Notes. * For example, a score of 2 means two correctly answered questions more on the post-test than on the pre-test and a score of -1 means one correctly answered question less on the post-test than on the pre-test.

The main effect of worked examples on learning gain is not significant, $F(2, 77) = 0.14$, $p = .867$. This indicates that all three conditions of worked examples have similar learning gains and no worked example is more effective for learning. This is confirmed by the results of the main question. The main effect of level of prior knowledge on learning gain is significant, $F(2, 77) = 7.80$, $p = .001$, $\eta 2 = .168$. This indicates that one of the three levels of prior knowledge scored significantly better or worse than the others. The learning gains of students with low prior knowledge ($M = 3.10$, $SD = 2.45$, $n = 31$) are significantly higher than for students with medium

prior knowledge (M = 1.22, SD = 2.17, n = 23) and high prior knowledge (M = 0.56, SD = 2.98, n = 32). This is an interesting result, because apparently the game is very accommodating for the weak third graders even though it's difficulty level was supposed to be high enough to enable strong fourth graders to show learning gain. Nevertheless, these are just estimations and future research is needed to confirm or reject these statements.

Control question

A paired samples t test with an α of .05 was performed to compare how much fun the participants were having while playing the seesaw-game (M = 3.63, SD = 1.27) compared to educational games in general (M = 4.05, SD = 1.03). On average, the participants liked educational games in general 0.42 point better, 95% CI [0.183, 0.655], than the seesaw-game, on a 5-point Likert scale. This difference was significant, $t(85) = 3.53$, p = .001, and large, d = 3.40. This result indicates that the seesaw-game is liked less than educational games in general.

Discussion

In this study three conditions (worked examples with self-explanation, worked examples with given explanation, and no worked examples) were compared on learning gain, having fun and mental effort within game-based learning with 86 third grade (in Dutch: groep 5) students. The main research question is: 'Which type of worked example as feedback has the biggest influence on learning, compromises the fun the students are having the least and influences the mental effort the students have to put in the most?'. Based on the results of the analysis, the answer seems to be none; the MANOVA was not significant. This means that all three conditions (worked examples with self-explanation, worked examples with given explanation, and no worked examples) used in the current study were equally effective for learning, as well as equal in preserving the fun and equal in decreasing the mental effort. Based on this statement, it seems that not using worked examples is the most efficient way of learning because it has the same effects on learning gain, experienced fun and mental effort but using less materials and time to do so than with one of the two types of worked examples added to the game. The unexpected results will be explained in more detail per sub question.

The first sub question is 'does the addition of the worked example to the game compromise the fun the learners are having?'. Based on the results of the analysis, it seems that it does not. This means that the game is not liked less by students who had worked examples added

into the game than by students without the worked examples. Based on this statement, it seems that worked examples can safely be added into a game. This is an unexpected outcome; according to Broza and Barzilai (2011) the fun of the students was expected to be compromised in the seesaw-game because of the addition of the worked examples and their interruption into the game. Possibly, the worked examples did not compromise the fun the students were having because the worked examples were no disruption of the game. According to Barzilai and Blau (2014) scaffolds only cause disruption of the game when they are internal. External scaffolds do not interrupt and thus do not reduce the fun the students are having (Barzilai & Blau, 2014). The worked examples were placed between the levels instead of in the middle of playing a level. This might be external enough to not disrupt the game and not compromise the fun.

The second sub questions is ‘does the addition of the worked examples influence the mental effort needed to play the game?’. Based on the results of the analysis, it seems that worked examples do not reduce the mental effort needed to play the game, as was expected, compared to the game without the worked examples. This means that it costs the students as much mental effort to play the game in each condition (worked examples with self-explanation, worked examples with given explanation, and no worked examples). This is an unexpected outcome because according to Sweller (1988) worked examples reduce the extraneous load of a student and according to Paas and Merriënboer (1993) reducing the cognitive load in the working memory also reduces the mental effort it takes to solve a problem. The unexpected outcome might have been affected by the focus of the cognitive load theory. This theory is focussed on the effects of instruction on the three types of cognitive load. Especially the extraneous and germane load, because those are caused by design choices for the instruction (Sweller, Van Merriënboer, & Paas, 1998). With worked examples as feedback the focus of the cognitive load theory can no longer be on direct instruction anymore, because the focus is now on giving feedback. However, worked examples are proven to be effective for learning in the form of instruction as well as in the form of feedback (Paas, 1992; Paas & Merriënboer, 1994) and in the context of game-based learning (Ter Vrugte, 2016).

There is one difference between the current study and the existing studies mentioned above which might have influenced the results, namely the average age of the target groups of the studies. The target group of the current study is young children (approximately 8 to 9 years of

age) instead of secondary school students of 12 to 15 years of age in Ter Vrugte's (2016) study, 16 to 18 years of age in Paas' (1992) study and 19 to 23 years of age in Paas and Merrienboer's (1994) study. The unexpected outcomes of the current study might be explained by the suitability of the mental effort survey with this age group. It is possible that it is difficult for the young students to judge their own mental effort. Another explanation might be the development of arithmetic strategic performance, which is needed to solve problems about the subject of the game. Imbo and Vandierendonck (2008) tested the arithmetic strategic performance of second (approximately age 7), fourth (approximately age 9) and sixth (approximately age 11) grade elementary school students and found a significant increase with age. The worked examples in the current study were expected to increase learning gains by optimising the cognitive load on the working memory of the students, but based on Imbo and Vandierendonck's (2008) study this might not the case because the target group of the current study might has underdeveloped memory retrieval and execution of procedural strategies, which are important variables in arithmetic strategic performance.

The third sub question is 'do the worked examples have different effects on learners with high and low prior knowledge?'. The estimation of the answer seems to be that the effect of the worked examples is not different for the levels of prior knowledge. Even though it is just an estimation, it is unexpected. According to Chi et al. (1989) worked examples with self-explanation would be more effective for learning for students with high prior knowledge and worked examples with given explanation would be more effective for learning for students with low prior knowledge. Possibly, the estimation is influenced by the lack of diversity in the sample and the results would be different if a larger range of prior knowledge would be examined.

A strong point of the current study is that the sample was very evenly divided over the conditions, either 28 or 29 participants per cell. Also, the order in which the participants received the versions of the test was divided evenly over the sample: 43 participants had test order AB and 43 had test order BA. Another strong point of the current study is that the participants are from two different schools. Two classes belonged to the same school but were located in different villages and communication between the teachers seemed limited. The different schools help to lessen the effect of the variable of the school the class belongs to on the results.

On the other hand, a limitation of the current study is that the game itself is also a variable

while answering how fun it was. The FunDiff variable is the difference between the fun score of the seesaw-game and the fun score of educational games in general. The difference between the fun score of the seesaw-game and the fun score of educational games in general is not only the addition of the worked examples but also the seesaw-game itself. Maybe, the game itself was liked less than educational games in general and the FunDiff was this big with or without the addition of the worked examples. Therefore, the comparison between the fun score of the seesaw-game and the educational games in general might be unfair when trying to assess the difference between the conditions. This is not the case when only the fun score of the seesaw-game is used, because then the only difference between the conditions is the type of worked examples added into the game. Sub question 1 is answered with the FunDiff variable as the dependent variable. The analysis is repeated with only the fun score of the seesaw-game as dependent variable but the outcome was also not significant.

Another limitation related to the fun score is the timing of the fun score survey. Possibly, the timing of the survey influenced the scores the participants gave on the fun score of educational games in general, because the question was asked after playing the seesaw-game. When the question was asked before playing the seesaw-game, the scores might have been different. Lastly, a limitation of the current study are the choices which are made about the worked examples. The results might have been different if a different amount of worked examples were used in the experiment. Throughout the game, the two conditions with worked examples had six questions with worked examples and twenty-four without. While the condition without worked examples had thirty questions without worked examples. Possibly, the amount of worked examples in the two conditions was too small to show the effect on the dependent variables, if there is any. The results might also have been influenced by the quality of the worked examples. Possibly, the results would be different with, for example, differently formulated or differently visualised worked examples.

Suggestions for future research could include different types of worked examples, like for example other types of worked examples or worked examples with another amount of exercises or other types of exercises than a multiple choice question, such as filling in the blanks or joining the dots exercises. Suggestions for future research might also include testing if worked examples in the form of videos, written texts or any other form is more effective for learning, possibly even

in combination with the lengthiness of the worked examples and the age or reading comprehension of the learners. Another suggestion for future research, as mentioned above, could experiment with the age of the target group, as it might have effects on the mental effort it takes for students to complete the activity.

References

Austin, P. C., & Brunner, L. J. (2003). Type I error inflation in the presence of a ceiling effect. *The American Statistician, 57*(2), 97-104. https://doi.org/10.1198/0003130031450

Baltra, A. (1990). Language learning through computer adventure games. *Simulation & Gaming, 21*(4), 445-452. https://doi.org/10.1177/104687819002100408

Barzilai, S., & Blau, I. (2014). Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences. Computers & Education, 70, 65-79. https://doi.org/10.1016/j.compedu.2013.08.003

Bielaczyc, K., & Recker, M. M. (1991). Learning to learn: The implications of strategy instruction in computer programming. In L. Birnbaum (Ed.), The International Conference on the Learning Sciences (pp. 39-44). Charlottesville, VA: Association for the Advancement of Computing in Education.

Brinkhuis, M. J., Savi, A. O., Hofman, A. D., Coomans, F., van der Maas, H. L., & Maris, G. (2018). Learning as it happens: A decade of analyzing and shaping a large-scale online learning system. *Journal of Learning Analytics, 5(2),* 29-46. http://dx.doi.org/10.18608/jla.2018.52.3

Broza, O., & Barzilai, S. (2011). When the mathematics of life meets school mathematics: Playing and learning on the "my money" website. In Y. Eshet-Alkalai, A. Caspi, S. Eden, N. Geri & Y. Yair (Eds.), Learning in the technological era: Proceedings of the sixth chais conference on instructional technologies research 2011 (pp. 92-100). Ra'anana, Israel: The Open University of Israel.

Carroll, J. M. (2004). Beyond fun. *Interactions, 11*(5), 38–40. doi: 10.1145/1015530.1015547

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145-182. https://doi.org/10.1207/s15516709cog1302_1

De Freitas, S. I. (2006). Using games and simulations for supporting learning. *Learning, Media, &Technology, 31,* 343-358. https://doi-org.ezproxy2.utwente.nl/10.1080/17439880601021967

Imbo, I., & Vandierendonck, A. (2008). Effects of problem size, operation, and working-memory

span on simple-arithmetic strategies: differences between children and adults? *Psychological research*, *72*(3), 331-346. https://doi.org/10.1007/s00426-007-0112-8

Johnson, C. I., & Mayer, R. E. (2010). Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior*, *26*, 1246-1252. https://doi.org/10.1016/j.chb.2010.03.025

Juul, J. (2003). "The game, the player, the world: Looking for heart of gameness." Paper presented at the Level Up-Digital games Research Conference, Utrecht. http://dx.doi.org/10.29378/plurais.2447-9373.2010.v1.n2.%25p

Kim, B., Park, H., & Baek, Y. (2009). Not just fun, but serious strategies: Using meta-cognitive strategies in game-based learning. *Computers & Education*, *52*(4), 800-810. https://doi.org/10.1016/j.compedu.2008.12.004

Klinkenberg, S., Straatemeier, M., & Van der Maas, H. (2009). Serious gaming in de Rekentuin – spelenderwijs oefenen en meten. Utrecht: FIsme, Universiteit Utrecht.

MacFarlane, S. J., Read, J. C., Höysniemi, J., & Markopoulos, P. (2003). Evaluating interactive products for and with children. Interact 2003, Zurich, SU: IOS Press.

Mayer, R. E. (2014). *The Cambridge handbook of multimedia learning.* United Kingdom, Cambridge: Cambridge University Press.

Mayer, R., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*, 43-52. https://doi.org/10.1207/S15326985EP3801_6

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. *Handbook of educational psychology*, 47-62. Retrieved from: https://books.google.nl/books?hl=nl&lr=&id=TjDIqrzfYaMC&oi=fnd&pg=PA47&dq=Problemsolving+transfer&ots=AAtPqZPAlB&sig=GjSCZ4NWQMFljbfE5zV29z8j8yU#v=onepage&q=Problem-solving%20transfer&f=false

Meijer, J., & Karssen, M. (2013). Effecten van het oefenen met Rekentuin. Retrieved from: https://www.leraar24.nl/app/uploads/TechnischEindRapportRekentuin.pdf

Michael, D. and Chen, S. (2006). *Serious games: games that educate, train and inform.* United States, Boston: Thomson Course Technology.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. Instructional Science, 32, 99–113. https://doi.org/10.1023/b:truc.0000021811.66966.1d

Mortara, M., Catalano, C. E., Bellotti, F., Fiucci, G., Houry-Panchetti, M., & Petridis, P. (2014). Learning cultural heritage by serious games. *Journal of Cultural Heritage*, *15*(3), 318-325. https://doi.org/10.1016/j.culher.2013.04.004

Oefenweb. (n.d.). Spellen Rekentuin. Retreived at March 3, 2019 from: https://www.oefenweb.nl/wp-content/uploads/2018/10/Spellen-Rekentuin.pdf

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology, 84,* 429-434. Retrieved from: https://www.researchgate.net/profile/Fred_Paas/publication/232523259_Training_Strategies_for_Attaining_Transfer_of_Problem-Solving_Skill_in_Statistics_A_Cognitive-Load_Approach/links/5599931008ae21086d25b10c/Training-Strategies-for-Attaining-Transfer-of-Problem-Solving-Skill-in-Statistics-A-Cognitive-Load-Approach.pdf

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38,* 63-71. https://doi.org/10/1207/S15326985EP3801_8

Paas, F., & Van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction, 16,* 87-91. https://doi.org/10.1016/j.learninstruc.2006.02.004

Paas, F. & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental-effort and performance measures. *Human Factors, 35,* 737-743. https://doi.org/10.1177/001872089303500412

Paas, F., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86,* 122-133. Retrieved from: https://pdfs.semanticscholar.org/1512/a1cc2b8199c1e3258b1bf26bc402d42ee88f.pdf

Phet Lab Balancing Act. https://www.golabz.eu/lab/balancing-act.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioural Science, 28*(1), 4–13.

https://doi.org/10.1002/bs.3830280103

Read, J. C., MacFarlane, S. J., & Casey, C. (2002). Endurability, engagement and expectations: Measuring children's fun. *Proceedings of interaction design and children, 2*, 1-23. Retrieved from: https://www.researchgate.net/profile/Janet_Read/publication/228870976_Endurability_Engagement_and_Expectations_Measuring_Childrenaposs_Fun/links/0deec518618d0828ce000000.pdf

Reisslein, J., Atkinson, R. K., Seeling, P., & Reisslein, M. (2006). Encountering the expertise reversal effect with a computer-based environment on electrical circuit analysis. *Learning and Instruction, 16*, 92-103. https://doi.org/10.1016/j.learninstruc.2006.02.008

Salen, K., & Zimmerman, E. (2004). *Rules of Play: Game Design Fundamentals.* London: MIT Press.

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research, 78*(1), 153 189. https://doi.org/10.3102%2F0034654307313795

Sim, G., MacFarlane, S., & Read, J. (2006). All work and no play: Measuring fun, usability, and learning in software for children. *Computers & Education, 46*(3), 235-248. doi: 10.1016/j.compedu.2005.11.021

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257-285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-295. https://doi.org/10.1023/A:1022193728205

Ter Vrugte, J. (2016). Serious support for serious gaming. (Doctoral dissertation) University of Twente, The Netherlands. doi: 10.3990/1.9789036541060

TULE. (n.d.). TULE inhouden & activiteiten. Retrieved from: http://tule.slo.nl/

Van der Meij, H., & Van der Meij, J. (2013). Eight guidelines for the design of instructional videos for software training. *Technical communication, 60*(3), 205-228. Retrieved from: https://core.ac.uk/download/pdf/18296149.pdf

Wouters, P., & Van Oostendorp, H. (2012). A meta-analytic review of the role of instructional support in game-based learning. *Computer & Education, 60*(1), 412-425.

doi: 10.1016/j.compedu.2012.07.018

Wouters, P. J. M., Van der Spek, E. D., & Oostendorp, H. van (2008). Serious games for crisis management: What can we learn from research on animations? In A. Maes & S. Ainsworth (Eds.), Exploiting the Opportunities: Learning with Textual, Graphical and Multimodal Representations (pp. 162-165). Tilburg: Earli.

Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van Der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of educational psychology*, *105*(2), 249. doi: 10.1037/a0031311

Appendix: Overview of Funometer questions

Question 1: How do you like to play games in school from which you learn? (In Dutch: Hoe leuk vind je het om op school een spel te doen om iets te leren?)
Question 2: How do you like to play this game from which you learn about the seesaw? (In Dutch: Hoe leuk vond je dit spel om te leren over de wipwap?)

The Funometer (Read, MacFarlane, & Casey, 2002) scale:

Figure 1. Funometer Smileys