Using video games to combine learning and assessment in mathematics education

Kristian Kiili[1*], Keith Devlin[2*], Arttu Perttula[1], Pauliina Tuomi[1], Antero Lindstedt[1]
[1]*Tampere University of Technology,*
{kristian.kiili, arttu.perttula, pauliina.tuomi, antero.lindstedt}@tut.fi
[2]*Stanford University, devlin@stanford.edu*

Abstract

One problem with most education systems is that learning and (summative) assessment are generally treated as quite separate things in schools. We argue that video games can provide an opportunity to combine these processes in an engaging and effective way. The present study focuses on investigating the effectiveness and the assessment power of two different mathematics video games, Semideus and Wuzzit Trouble. In the current study, we validated the Semideus game as a rational number test instrument. We used it as a pre- and a post-test for a three-hour intervention in which we studied the effectiveness of Wuzzit Trouble, a game built on whole number arithmetic and designed to enhance mathematical thinking and problem solving skills. The results showed that (1) games can be used to assess mathematical knowledge validly, and (2) even short game-based interventions can be very effective. Based on the results, we argue that game-based assessment can create a more complete picture of mathematical knowledge than simply measuring students' accuracy, providing indicators of student misconceptions and conceptual change processes.

Keywords: *Game, Mathematics, Learning, Assessment;*

1. Introduction

The acquisition of mathematical skills is crucial for life in today's society. For example, Parsons and Bynner [1] have stated that at an individual level, insufficient mathematical competencies may be even more harmful to career prospects than reading or spelling deficiencies. The growing demands of mathematical skills has led to an increased need for developing effective and engaging ways to learn and assess mathematical competencies. The growth, in recent years, of the number of digital mathematics applications and video games has helped to meet this need.

One problem with most education systems is that learning and summative assessment are generally treated as quite separate things in schools. We argue that video games can provide us an opportunity to combine these processes in an engaging and effective way. Thus, it is reasonable to consider games not only as tools for learning, but also as assessment solutions.

In the case of learning, scholars have increasingly argued that the sense-making practices that occur when people engage with digital games constitute a form of literacy that is potentially better suited to address the needs of learners in the 21st century [2, 3, 4].

Research has indicated that educational games can support learning (e.g. [5, 6, 7, 8]) and engage learners (e.g. [9, 10, 11]). However, much of the evidence about the effectiveness of game-based learning is based on unsystematic studies involving inappropriate methods and designs [12, 13]. The need for better research is evidenced by several studies showing that game playing is not always an effective learning method. Successful use of video games may depend on how well the game is designed to meet the learning goals and player characteristics (e.g. [14, 15]). Research has shown that games can be used to support the learning of mathematics [13, 16, 17] — *provided they are properly designed*.

1.1 Games as learning assessment tools

Generally, assessment refers to the process of gathering information about the competencies or other attributes of a person. Shute and Kim [18] have argued that too often student assessment is used for purposes of grading, promotion, and placement, but rarely to enhance learning. Assessment that aims to improve learning is called formative assessment. In contrast, summative assessment refers to the traditional approach that is used to assess educational outcomes for purposes such as grading, certification, etc. For both forms of educational assessment, summative and formative, an important goal is to minimise uncertainty or error. Validity and reliability are key aspects of assessment quality [18]. Validity refers to the extent to which the assessment accurately measures what it is supposed to measure, and the accuracy of any inferences made from the test results regarding underlying competencies; reliability refers to the consistency of assessment results.

Bellotti et al. [19] have conducted a comprehensive literature review of the educational effectiveness of serious games. As they have stated, assessing learning within a simulation or a serious game is not a simple matter, and further work and studies are required, especially in game-based assessment. They suggest two major directions for future research on serious games: detailed characterization of players' activities and better integration of assessment in games.

External assessment methods such as multiple-choice questions are likely to disturb flow in immersive games [20]. To counter this, Shute et al have emphasised the adoption of stealth (embedded) assessments that are less disruptive to flow experience in the game. According to Shute and Kim [18], stealth assessment is an evidence-based approach to assessment where the tasks that a student performs are highly interactive and engaging. The aim of stealth assessment is to blur the distinction between assessment and learning.

Shute and Kim [18] have emphasised that evidence-centered assessment design (ECD) is the key element for creating coherent assessment solutions for games. The conceptual assessment framework of ECD includes competency, evidence, and task models that define what variables are measured and how they relate to targeted competencies and knowledge. The assembly model specifies how the competency, evidence, and task models work together to form a valid assessment. One of us (Kiilli) utilized the ECD approach for designing embedded assessment into the *Semideus* game. In the present study we used *Semideus* for summative assessment, but at the same time we studied ways to modify the game so that it would also support formative assessment.

1.2 Current study domain

The US National Research Council [21] has stated that mathematical proficiency consists of five components: 1) conceptual understanding, 2) procedural fluency, 3) strategic competence, 4) adaptive reasoning, and 5) productive disposition. Devlin [4] has argued that well designed digital video games could support numerical development and mathematical proficiency. Nevertheless, many published mathematics games support mainly procedural fluency, a focus that tends to be found in classroom practices as well.

In terms of mathematical proficiency, the games used in this study address conceptual understanding and strategic competence. Although the *Semideus* game is about rational numbers and the *Wuzzit Trouble* game deals with natural numbers, both games support conceptual understanding of numbers (number sense[1]) and the development of strategies that support processing of numbers on the number line. The integrated theory of numerical development supports this argument by emphasizing that, in addition to differences between natural and rational numbers, natural numbers and rational numbers also have important commonalities [23]. In both, students need to learn how to interpret number symbols in terms of the magnitudes to which they refer, and this understanding of magnitude is central to general mathematical competence [24]. Moreover, it has been shown that a good understanding of natural numbers is highly predictive for overall mathematics achievement, and in particular for performance on rational number tasks [25].

Because *Semideus* is used as a pre- and post-test in the study's intervention, a brief discussion is in order concerning common learning difficulties related to rational numbers and methods frequently used to study this domain.

[1]In recent years, number sense has been identified as a key (arguably the most important key) to mastery of numerical mathematics. For an introductory discussion of the concept, see Section 2,3 of the paper by Pope and Mangram [22].

Understanding of rational numbers is crucial for mathematics learning, in particular because it is predictive of students' mathematical achievement years later [26, 27]. Nevertheless, there is a great deal of evidence that children find understanding of rational numbers very difficult, and even after considerable mathematics instruction many children fail to perform adequately in simple fraction tasks [23, 28, 29]. Mathematics education researchers suggest that most of the students' difficulties with rational numbers can be attributed to inadequate instruction [30]. This may be, at least in part, due to the fact that recent advances in modeling numerical development have not been incorporated into the practices of teachers, and present-day instruction tends to emphasize procedural instead of conceptual knowledge [31].

According to conceptual change theories, children form an initial conception of numbers as counting units before they encounter fractions, and later they draw heavily on this initial understanding to make sense of rational numbers [29, 32]. Misconceptions about rational numbers tend to originate in children's false belief that all properties of whole numbers can be applied to rational numbers. This phenomenon is often referred as a *whole number bias* [33] or *rational number bias* [34].

According to Siegler, Thompson & Schneider [23], significant conceptual change is required to understand that fractions and decimals, like whole numbers, represent magnitudes that can be located on number lines. Recent findings have suggested that instructional interventions that aim to support conceptual change should target learners' interpretation of rational numbers as magnitudes by reasoning about them as points on number lines [24]. This approach is reasonable, because there is clear evidence that conceptual fraction knowledge also promotes fraction arithmetic skills (e.g. [31, 35]). Moreover, research has shown that number line estimation tasks do not purely measure an underlying representation of number magnitude [36], but appear to measure also the ability to apply strategies that allow students to accurately place a number on the line [37].

It is important that teachers and researchers are able to accurately assess students' knowledge and identify misconceptions, in order to be able to support those students' numerical development. Past research has focused on identifying misconceptions in mathematics through individual interviews, written assignments, multiple-choice questions, or by classifying errors on collective assessments (e.g. [30, 38]). For example, Durkin and Rittle-Johnson [38] identified whole-number-bias misconceptions in a fraction magnitude context by dividing the number of misconception errors made by the total number of items on which that type of misconception error was possible. This approach is questionable, because it does not take into account the accuracy of tasks in which misconceptions are not possible. In general, most of the methods used to identify misconceptions require considerable effort, and finding more convenient and engaging methods to study the phenomenon would be a welcome advance. Game-based assessment could be one answer, and in this paper we explore how a game can be used to assess students' conceptual fraction knowledge.

1.3 The present study

In this paper we report the results of a recent pilot study undertaken in schools in California and Finland, in which we used one mathematics learning game (*Semideus*, a rational number game) to provide pre- and post-tests to determine the learning outcomes resulting from an intervention with another mathematics learning game (*Wuzzit Trouble*, an integer arithmetic game). An earlier study, directed by Jo Boaler and carried out by Pope and Mangram (this special issue) at Stanford University found that a 120-minute intervention with free *Wuzzit Trouble* play spread over one month led to significant learning outcomes, as assessed by a written pre- and post-test dealing with whole number operations. Our study explores this same phenomenon. In particular, the aim of this paper is to investigate if we can use another mathematics game (*Semideus*) as a testing (assessment) instrument, and moreover, does this game-based test instrument produce similar results as the paper-based test did about the effectiveness of the *Wuzzit Trouble* game.

We hypothesized that *Semideus* playing performance provides similar results to a paper-based rational number test, and thus can be used to assess students' conceptual rational number knowledge (Hypothesis 1a). Second, in line with recent results about the positive relationship between understanding of rational number magnitudes and rational number arithmetic performance (e.g. [31, 35]), we hypothesized that the *Semideus* playing performance is positively related to rational number arithmetic performance (Hypothesis 1b). Third, based on the earlier learning-outcome study of *Wuzzit Trouble* [this volume] and the positive relationship between understanding of whole numbers and performance on rational number tasks [25], we hypothesized

that a free 120 minute *Wuzzit Trouble* play would enhance students' conceptual rational number knowledge, assessed with the *Semideus* game (Hypothesis 2).

2. Method

2.1 Design and participants

A pre-test/post-test quasi-experiment design involving a treatment group and a control group was used. The *Semideus* rational number game was used as a pre-test and post-test and the *Wuzzit Trouble* game as a treatment. Table 1 shows the overall design of the study. The length of the whole study was approximately 2 months, with students participating in the study for 40 minutes each week.

Table 1. Design of the study

	Pre-test	**Treatment**	**Post-test**
Treatment group only		*Wuzzit Trouble* game (3*40 minutes)	Playing experience questionnaire (Wuzzit Trouble)
Treatment & control groups	*Semideus* game (40 minutes)		*Semideus* game (40 minutes)
	Playing experience questionnaire (Semideus)		
	Paper based RNT*		

* RNT = Rational Number Test

Two Finnish and two US sixth grade classes, each pair taught by the same teacher, participated in the study. In the execution, the research protocol followed in the two countries was different, due to some unexpected constraints on the teacher of the Finnish sample. Thus, only the intervention data from the US sample is analyzed in this paper. However, the Finnish sample is included in the playing experience analyses and the analyses that relate to validation of *Semideus* as a rational number knowledge assessment tool. Henceforth, we refer to these two analyses as the Intervention Study (US only) and the Validation Study (US and Finland).

Validation study. The final sample that was used in the validation study was 66, of which 30 were Finnish students (18 females) and 36 US students (24 females). The mean age of the Finnish students was 12.1 and the mean age of the US students was 11.43. The main aim of this study was to validate the *Semideus* game as a test instrument.

Intervention study. The *Wuzzit Trouble* treatment was assigned to one class in each country and the other classes were treated as control groups. The final US sample that was used in the intervention analysis was 25. The final sample is small, because US participants who did not complete both the *Semideus* pre- and post-tests were excluded. Additionally, two participants from the control group were excluded as outliers because of their very weak post-test game behavior (log files revealed impulsive behavior). As a result, the sample of the US treatment group numbered 11 (6 females), and the sample of the control group numbered 14 (10 females). The main aim of this study was to study the mathematical effectiveness of the *Wuzzit Trouble* game.

2.2 Description of the games

Semideus is designed to support the development of rational number conceptual knowledge. *Wuzzit Trouble* is built on whole number arithmetic, and is designed to enhance mathematical thinking and problem solving skills. Both games are designed to work also as assessment tools.

2.2.1 Description of the Semideus game

The thematic setting and visual appearance of the *Semideus* game relate to the mythology of ancient Greece. In the story, Semideus, a son of Zeus, is tasked to seek golden coins that Kobolas the goblin has stolen from Zeus. Kobolas has hidden the coins, as well as traps, along the trails of Mount Olympos. Semideus has discovered the locations of the coins, encrypted in mathematical symbols, and must race the goblin to retrieve the coins. While collecting the coins, Semideus

climbs up to reach the mountaintop (end of the level – each level is represented as a mountain), where Zeus is eagerly awaiting his coins.

Semideus is designed to support the development of rational number conceptual knowledge. In particular, the game addresses the development of two conceptual sub-concepts necessary for a complete mathematical understanding of rational numbers: 1) representations of the magnitudes of rational numbers and 2) the density of the rational numbers. The idea of the game is grounded in the integrated theory of numerical development [23], and is designed to expand understanding of the connection between different number representations and magnitudes. As the recent research has suggested, *Semideus* relies on an interpretation of rational numbers as magnitudes situated on a number line.

The game has four different task types. In magnitude estimation tasks, the player has to estimate a spot, hidden coin cache, on a number line that matches the numerical value given. In addition there can be values telling the positions of traps that the player must avoid. Values can be represented as fraction or decimal numbers. In the tasks included in this study we used number lines from zero to one and from zero to five. The player was rewarded with coins (score) if his or her estimation was accurate enough (over 98% = 500 coins; 95% - 98% = 300 coins; under 92% - 94% = 100 coins). If the estimation accuracy was under 92%, Semideus lost energy and he had to progress to the next platform (i.e., faced a new task).

In magnitude comparison, magnitude ordering, and density tasks, Semideus faces stones annotated with rational number symbols (Figure 1). The task of the player is to organize the stones in ascending order before the goblin steals all the gold coins from the upper platform. In comparison tasks, the player has to organise two stones and both in ordering and density tasks three stones. If the magnitudes of the stones are equal they must be stacked in a pile. In this way, the magnitude ordering and comparison tasks can also address the equivalency of rational numbers. Basically the density task is similar to the ordering task, but the fractions or decimals of the task are selected to create cognitive conflicts between whole number and rational number properties. (e.g. Items to be ordered could be 2/5, 3/5 and ½.) Regarding whole number properties, the player might first think there cannot be numbers between 2/5 and 3/5, because 2 (nominator of 2/5) is followed by 3 nominator of 3/5) in terms of whole number ordering. However, when the player has to order the stones, he or she might realize that ½ has to be between the other two fractions, which may lead to conceptual change and understanding the concept of ratio. If the stones are organized in the wrong order, Semideus loses energy and has to progress to the next platform (i.e., faces a new task).

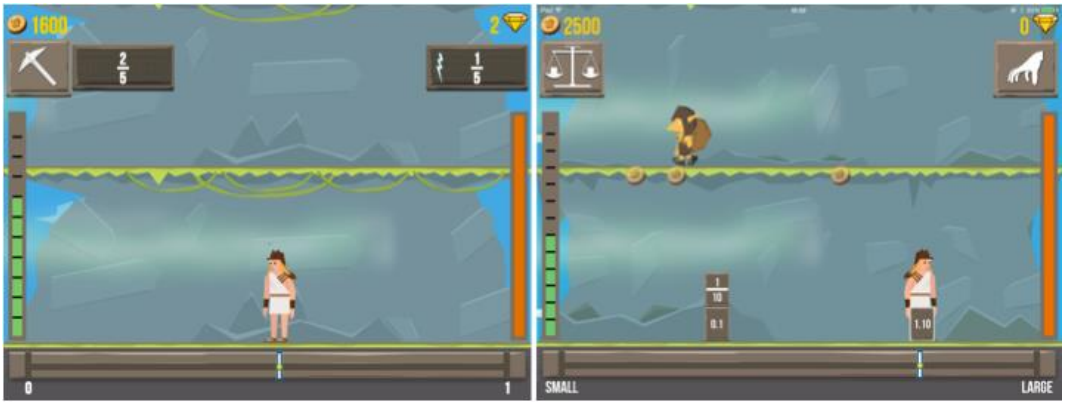

Figure 1. Examples of an estimation task (left) and ordering task (right). The orange bar is an energy indicator and the green bar expresses the progress towards the mountaintop.

The player can control the Semideus character by tilting the tablet device. When tilting the tablet right, Semideus walks right on the platform (number line), referring to an increasing the number magnitude. When tilting the tablet left, Semideus walks left, referring to decreasing number magnitude. In terms of Pouw et al. [39], the tilting utilizes embedded embodied interactions, an approach that can reduce cognitive load. In general, the game character and stone tablets can be considered as visual manipulatives that help players to concretize the mental number line. The more detailed description of the game mechanics and related theories can be can be found from www.gamestoschools.com.

In the present study, the game was balanced in the way that the player always manages to climb to the top of the mountain. In other words the player could not loose all energy before reaching the mountaintop. This was done because we wanted comparable data from each player – each player played all seven levels once. However, in order to engage players, we tried to create an illusion that Semideus could run out of energy. On the top of each mountain, the player's overall performance was evaluated. Bonus coins were provided according to the player's energy level (maximum 1000 coins). Furthermore, the player could earn 1-3 stars. The number of stars achieved was determined based on the number of collected coins and the energy level. In this study, we did not investigate how well stars and collected coins express player's understanding of rational numbers, rather the assessment was based entirely on the player's absolute accuracy recorded in the secured server. (More details in section 2.3 - Instruments and measures.)

2.2.2 Description of the Wuzzit Trouble game

Wuzzit Trouble's UI is a representation of certain kinds of integer-arithmetic problems (integer partitions—the expression of a whole number as a sum of other whole numbers—and Diophantine equations) equivalent, but alternative, to the familiar symbolic algebra representation (trading in a static, spatial configuration of symbols for a dynamic interaction with a digital gears mechanism).
The player rotates one or more of up to four small drive cogs to rotate a large gear-wheel (Figure 2). The object it to bring the keys in line with a fixed marker at the top, which causes the key to move from the wheel to a collection slot at the top of the screen. Collecting all the keys in this manner releases the Wuzzit from the trap. A small cog may be wound-up to rotate up to five times with a single player-action. More stars are obtained by releasing the Wuzzit with the fewest number of rotation actions, making optimization a key objective.

Figure 2. Example level of Wuzzit Trouble.

But there is a twist. If the mechanism as outlined above were all there were, the game would simply provide valuable practice with the basic skills of arithmetic. The insight behind the unique feature of *Wuzzit Trouble* that makes it a *powerful representation of complex mathematical performance tasks* is the addition of optional bonus items that are also situated on the wheel. Collecting bonus items yields extra points. While bonus items are common in video games, the way this feature is implemented in *Wuzzit Trouble* adds a whole new layer of complexity, requiring deep mathematical reasoning to find optimal solutions. The critical requirement is that bonus items have to be collected before the last key drops. With this restriction, optimizing the score can require sophisticated algorithmic reasoning. The specific requirements for one-, two-, and three-star solutions are given when the puzzle is selected. A three star solution requires collection of all keys and bonus items in at most the number of moves specified with the puzzle. (The possible solution paths number just over 2 trillion, so quite clearly trial-and-error is not a viable approach. Players must develop multi-step algorithms to solve the puzzles optimally.)

Wuzzit Trouble permits use in many different ways to achieve different learning outcomes. At the most basic level it provides valuable practice that reinforces number sense. Play beyond the early game levels involves complex performance tasks, requiring a range of powerful problem solving techniques, including algebraic and algorithmic thinking. These features make the game suitable for use with both K-12 students, as well as having appeal for adult puzzle lovers, facilitating family use.

2.3 Instruments and measures

User experience was measured in terms of flow experience [40] and playability. Flow experience was measured with a 9-item questionnaire developed by the authors. The items included were derived from the flow scale used in a recent serious games study [9]. The dimensions included were: challenge-skill balance, clear goals, concentration, autotelic experience, loss of self-consciousness, sense of control, and action awareness merging. Playability was measured with a two-item questionnaire developed by the authors. The playability dimensions included were intuitiveness of the user interface and controlling accuracy. A 6-point Likert-type response format was used for all dimensions.

The Rational Number Test (RNT) was used to collect data about participants' rational number conceptual knowledge. The tasks of the RNT used in this study was derived from validated RNT developed by McMullen et al. [41]. The original RNT was shortened (open-ended questions and whole number tasks removed) and some tasks were changed to better correspond the difficulty level of the *Semideus* game. For example, the elements of one ordering task (6/12, 5/7, 2/6) were changed to 6/12, 5/7, 4/6. Moreover, we did not include any magnitude estimation tasks in which numerical hints were shown on the number line. The conceptual part of the RNT consisted of eight estimation tasks, eleven comparison tasks, six ordering tasks, and four multiple-choice density tasks. Additionally, six procedural fraction tasks were included from which two were the same asin the original RNT.

Semideus Log files. The version of the *Semideus* game that was used in this study included seven levels. The aim of the first level was to introduce the user interface to the players. It consisted of 10 whole number magnitude estimation tasks that were not included in the final analyses. Levels 2-7 consisted of 30 estimation tasks on 0-1 number line, 12 estimation tasks on 0-5 number line, 12 magnitude comparison tasks, and 18 ordering tasks (including four density tasks). Each level could be played only once. The *Semideus* game continuously logged detailed playing behavior on a secured server.

The assessment of player's rational number conceptual knowledge was designed using the Evidence-Centered Design (ECD) framework [18]. Rational number magnitude estimation, magnitude comparison and magnitude ordering are the core competences that describe player's conceptual rational number understanding in the game. The competence model describes these core competencies in more details. For example, magnitude ordering competence consists of proper fractions, improper fractions, decimal numbers, whole number bias risk, digit count bias risk, common nominators, common denominators, whole number ratio, equivalency, and density properties. The task/action model expresses situations that evoke evidence for the competencies, and game's evidence model describes what playing behavior can provide evidence of certain competencies. For example, the task/action model may indicate that a player faces a task related to ordering competence with density and fraction properties. In this case the evidence model determines that if this is the player's first answer (after the level has been started) to the task, then the density accuracy variable of the ordering competence should be updated based on the player's answering success. In addition, related variables such as task duration are recorded. Finally the the assembly model specifies how the valid assessment is determined based on the other models.

We have defined semantics for the task model that is used to describe all the tasks of the game. Based on the semantics, each task is tagged with keywords that describe the task in terms of rational number competencies. For example, the comparison task: 2/1 vs. 3/5 would be tagged as: 'COMPARISON'; 'WN_O_IC' (refers to inconsistency of whole number ordering – both components of the 3/5 fraction are bigger than components of 2/1, but the magnitude of the 3/5 is smaller); 'IMPROPER_INC' (refers to inclusion of improper fraction 2/1), 'WN_RATIO_INC' (Whole Number Ratio Included, 2/1 = 2). We have developed a data-analyzing tool that can fetch data based on the tags. For example, we might want to search for game data related only to tasks containing fraction numbers that are susceptible to whole number bias misunderstanding. The same interface can be used to create realtime learning analytics and assessment reports.

Wuzzit Trouble playing performance. The overall points score that participants obtained in *Wuzzit Trouble* was used to assess problem solving abilities. The maximum possible point counts for each of the three rooms that the puzzles appear in (25 to each room) are 54,000, 61,500, and 85,500, respectively.) To obtain a high points score on a puzzle, the player has to (1) solve the puzzle (free the Wuzzit), which means collecting all the keys on the large wheel, (2) collect as many of the bonus items as s/he can (all of them for a maximum score), and (3) do so in as few moves as possible (at or below a stated threshold that varies from puzzle to puzzle). Since the puzzle ends when the final key has dropped, players must sequence their moves carefully to avoid dropping the final key before collecting all the bonus items. This requires that players develop good solution strategies. The depth of strategies required may be seen by representing a solution using algebraic notation, which is possible because the puzzle is simply a mechanical representation of a system of simultaneous linear equations (Figure 3).

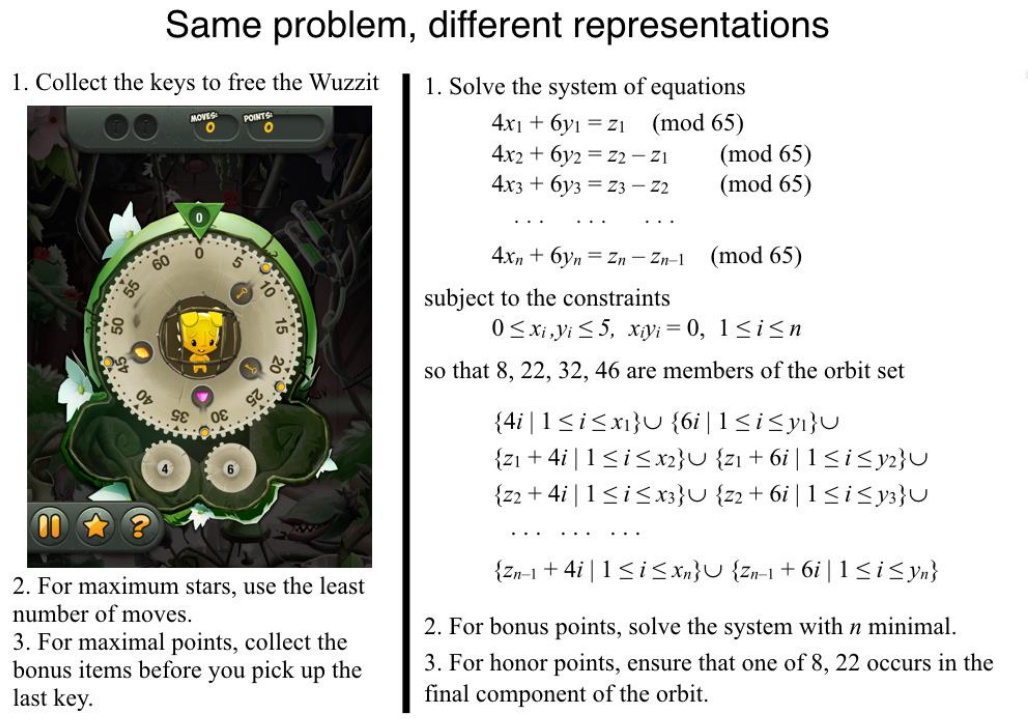

Figure 3. The system of equations on the right is the sequence of moves the player must make to solve the puzzle on the left. Because the space of solution paths is so large, a player cannot obtain a high score without implicitly finding an efficient solution to such an equation system. The fact that young children can readily solve the puzzle demonstrates the power of an efficient representation, and provides an illustration of Devlin's observation that much of the difficulty people have learning mathematics arises because of the Symbol Barrier [4]

2.4 Procedure

The length of the whole study was approximately two months, with students participating in the study for 40 minutes each week. Each student had his or her own iPad during the study. At the beginning of the study, students were given a username that was used to link different data sources of the study: questionnaires and playing behavior logs. The progress of the study was as follows:

- Week 1: Students filled in the demographics questionnaire and the paper-based rational number test (RNT).
- Week 2: The researchers introduced the game and the aim of the study to students. After that, students played the *Semideus* pre-test game (40 minutes playing session). Students were not allowed to discuss the game tasks with other students during the testing session. After the pre-test, students filled in the *Semideus* playing experience questionnaire.
- Weeks 3-5: *Wuzzit Trouble* was played three times (3x40min) a week within the three-week period. Additionally, students were allowed to play *Wuzzit Trouble* at home if they

wished to. Within this three-week period, the US control group played games that did not directly include mathematical content (3x40min). After the intervention, students filled in the *Wuzzit Trouble* playing experience questionnaire.

- Week 6: Students conducted a Math Puzzle Test (not analyzed in this paper).
- Week 7. Students played the *Semideus* post-test game (40 minutes playing session). The post-test was the same game as the pre-test.

The teachers of the classes administered all the sessions except the first playing session (week 2). Support for the teachers was provided whenever needed.

2.5 Analyses

Related to playing experience analyses, a flow construct was computed (a mean of nine flow dimensions). Similarly, a playability construct was computed (a mean of two dimensions). The analysis of the flow questionnaire indicated that the flow construct was internally consistent in both game contexts (*Semideus* game: $\alpha = .79$, *Wuzzit Trouble* game: $\alpha = .92$), which indicates that all nine dimensions measured the same phenomenon, flow construct. However, the internal consistency of the playability construct was weak in the *Semideus* context ($\alpha = 0.5$), but good in *Wuzzit Trouble* ($\alpha = .86$).

Several variables related to playing performance in the *Semideus* game were computed. Estimation tasks accuracy was computed as

100*abs (correct value – estimated value)/numerical range of the number line.

In the fraction magnitude comparison, ordering, and density tasks, the percentage of correctly solved tasks was computed. In some analyses, we used *Overall playing performance*, which is an average of estimation accuracy, comparison accuracy, ordering accuracy, and density accuracy.

A conceptual rational number knowledge accuracy was computed in a paper-based RNT test, similarly to the Overall playing performance, and included accuracy of estimation, comparing, and ordering tasks. Additionally, the procedural fraction accuracy was calculated.

In the context of the intervention study, the gain scores were computed by subtracting the *Semideus* pre-test scores from the post-test scores for both treatment and control groups. A Shapiro-Wilk's test ($p>.05$) [42] and a visual inspection of histograms, normal Q-Q plots, and box plots showed that the gain scores were approximately normally distributed for both the treatment group and the control group, with a skewness of 0.267 (SE = 0.66) and kurtosis of -0.34 (SE = 1.28) for the treatment group and a skewness of -0.89 (SE = 0.60) and kurtosis of 4.14 (SE = 1.15) for the control group [43]. Thus, a t-test was used to compare gain scores of treatment and control groups. The effect size (Cohen's d) was calculated based on t-test values, to evaluate the effectiveness of the treatment.

Moreover, Cronbach's alpha coefficient was used to evaluate the internal consistency of measures, and correlation analyses were used to study relationship between different variables.

3. Results

The results are presented in two parts. First, we present the results of the validation study in which we investigated the usefulness of *Semideus* as an assessment tool. Second, we present the results of the intervention study in which we studied the effectiveness of *Wuzzit Trouble* in terms of rational number conceptual knowledge.

3.1 Validation study

The content validity of the tasks included in the *Semideus* game was evaluated before data collection. Content validity was determined through a panel of experts [44]. Three mathematics experts provided feedback on the included tasks. The feedback indicated that the task types of *Semideus* were relevant for measuring rational number conceptual knowledge. However, the experts pointed out that the number of fraction number tasks was larger than the number of decimal number tasks. We were aware of this imbalance, but we could not add more decimal number tasks because of time restrictions set by the participating schools (possible length of the test sessions). On the other hand, having a greater number of fraction tasks is defensible because fraction numbers are more complex and children usually find them harder to understand.

One of our main aims was to study how well the *Semideus* game performance describes students' rational number conceptual knowledge. In order to further shed light on content validity we compared students' overall accuracy in the paper-based RNT to students' overall accuracy in the *Semideus game*. The correlation analysis indicated that the content validity was good, with a significant relationship between paper-based RNT accuracy (M = 74.13%, SD = 14.73%) and accuracy in *Semideus* (M = 70.19%, SD = 16.24%), r = .79, p < .001. Moreover, the test-retest reliability coefficient within control group (r = .76, p < .001) indicated that *Semideus* provided consistent results over time (There was a five weeks delay between the pre- and post-tests.).
Both the paper based RNT and the *Semideus* game consisted of rational number estimation, comparing, ordering, and density task types. The correlation analysis indicated that there were statistically significant correlations between paper-based task types and game-based task types (see table 2). The correlations for each task type ranged from .38 to .60 and indicated some consistency in conceptual rational number knowledge between paper based test and game test. It was no surprise that the smallest correlation was between paper-based density tasks (M = 37.08%, SD = 25.46%) and game-based density tasks (M = 59.4%, SD = 33.8) r = .38, p < .05, because these tasks were implemented totally differently. Paper-based RNT measured understanding of the density concept with multiple-choice questions and in *Semideus* the density tasks were based on ordering tasks (e.g. items to be ordered 2/5, 3/5 and ½). In fact, the density tasks of *Semideus* are designed to create cognitive conflicts – first the player may think that there cannot be numbers between 2/5 and 3/5 (previous example), but when they have to order the stones they might realize that ½ has to be between the other two fractions. We assume that such cognitive conflicts may be one reason why the mean value of the game-based density tasks was much higher than the mean of the paper-based multiple-choice density tasks. On the other hand the paper-based RNT was conducted before the *Semideus* playing session, and the multiple-choice tasks may have triggered reflective processes in participants before they played *Semideus*. Finally, as we can see, the relationship between the overall accuracy in both tests was stronger than the correlations between individual task types.

Table 2. Correlations between paper-based test (pt) and Semideus game data (gt) according to different task types (n =66).

	Estimation (pt) M=82.24%, SD=9.52%	Comparing (pt) M=81.3%, SD=19.17%	Ordering (pt) M=62.1%, SD=32.45%	Density (pt) M=37.08%, SD=25.46%
Estimation (gt) M=86.4%, SD=6.41%	**.60** **	.69 **	.60 **	.54 **
Comparing (gt) M=75.7%, SD=24.18%	.38 *	**.58** **	.56 **	.29 *
Ordering (gt) M=53.25%, SD=27.55%	.60 **	.68 **	**.56** **	.50 **
Density (gt) M=59.38%, SD=33.78%	.45 **	.61 **	.50 **	**.38** *

Correlation significance levels (2-tailed): ** .001 level; * .05 level

Cronbach's alpha coefficient for the four *Semideus* task types was .82, which indicates that the internal consistency of *Semideus* task types were good assessing consistently rational number conceptual knowledge. The alpha coefficient for the same four paper-based task types was .71. We did not evaluate the internal consistency of each task type because theoretically there can be large variation between difficulties of tasks within each task type. For example, it is easier to estimate the magnitude of ½ on the number line than the magnitude of 2/7.
Table 2 also shows that the students performed better in comparison tasks than in ordering tasks. Large standard deviations in comparison and ordering tasks indicate that students' rational number conceptual knowledge varied considerably. From a learning analytics and assessment point of

view, it would be meaningful to consider task accuracy in a more detailed way. For example, the analytics revealed that the most challenging tasks for players seemed to be fraction ordering tasks (M=42.16%, SD=26.22), ordering tasks that included both fraction and decimal numbers (M=44.39%, SD=28.04), and tasks that included equivalent numbers (M=49.99%, SD=39.99). In general, the accuracy information can be used to provide meaningful information for the learners as well as for the teachers, and could be used to facilitate learning. However, such information does not help teachers to identify common rational number misconceptions. The semantic model we have used to describe the tasks of *Semideus* identifies most common misconceptions.

In our study, we tried to identify students who may have misconceptions related to whole number bias [33]. The analysis revealed that 12.5% of the students had a clear whole-number ordering-related bias in fraction numbers. Additionally, the analysis revealed that 13 % of the students had a clear digit count bias in decimal numbers. Both biases were detected if a student's accuracy in bias related tasks was 0% and accuracy in non-bias tasks was at least 50%. In this analysis we used a strict 0% accuracy rule, because the degree of the bias-related tasks was small. These preliminary results show that game performance data can be used to detect misconceptions. This is a very promising result, because previous research has revealed that teachers need support in detecting misconceptions [45] and our study shows that games could be used to provide such support. However, more powerful and reliable ways to detect misconceptions needs to be developed and thus we are currently creating an algorithm that can detect misconceptions and related conceptual change processes in real-time.

According to Anastasi [46], perceived relevance of a test is essential for motivating participants to demonstrate their "full repertoire of skills." If a test is considered to be irrelevant or meaningless, participants are likely to be less willing, which compromises the validity of the test. This is to say, in a game-based assessment context it is important to ensure that play is engaging and players understand the aims of the game-based test. Furthermore, control of the game has to be accurate and easy (good playability), so that the players experience the game as a fair test instrument, the game controls do not negatively influence on the test performance, and the game behavior provides valid information about the knowledge and skills of the player.

In this study we measured user engagement in terms of flow experience and playability. Table 3 shows that players experienced quite high flow experiences when playing *Semideus* (M = 4.24, SD = 0.86).

Table 3. Mean values and standard deviations of flow and playability in *Semideus*.

	M	**SD**
Flow construct	4.24	0.86
• Challenge – skill balance	4.15	1.29
• **Clear goals**	**4.52**	**1.52**
• **Immediate & unambiguous feedback**	**4.80**	**1.17**
• Action awareness merging	4.44	1.08
• Time distortion	3.94	1.63
• Sense of control	4.06	1.29
• Concentration	3.99	1.48
• **Loss of self-consciousness**	**4.71**	**1.55**
• Autotelic experience	3.55	1.42
Playability construct	4.36	1.17

Note. Three highest scoring flow dimensions are bolded.

As we can see, the clear goals, immediate and unambiguous feedback, and loss of self-consciousness flow dimensions scored highest. Players could perceive the goals of the game, and

the game provided feedback that enabled the players to understand how well they were doing with respect to the goals. It was surprising that the loss of self-consciousness dimension scored second highest (M = 4.71, SD = 1.55). This means that players could concentrate on the game content in a way that they did not consider how others might have evaluated their playing performance. This finding is very important in the light of assessment, since players are able to do their best without the familiar pressures normally associated with assessment. Good playability score (M = 4.36, SD = 1.17) indicates that the tilting based user interface was accurate and intuitive to use, even though the game was played without sounds. The autotelic dimension, which refers to enjoyable and intrinsically rewarding experience scored the lowest (M = 3.55, SD = 1.42). This is understandable because we used the game in a summative assessment manner and several engaging game mechanics were excluded from the version used in this study. For example, enemies, mystery boxes, hints, and a learning companion were excluded in order to study *Semideus*'s core gameplay (task types). Nevertheless, the flow and playability results indicate that the quality of the *Semideus* game was good and most of the players enjoyed it. In fact the majority of players reported that they wanted to achieve the maximum three stars from each level (M = 4.97, SD = 1.82). Based on these findings we assume that players were engaged and tried to perform well in the game.

The correlation analysis indicated that there was a significant relationship between flow experience and playability ($r = .56$, $p < .001$), which means that the playability of the game can have a strong influence on overall user experience and the validity of collected data. Moreover, the correlation analysis revealed that there was a significant relationship between overall playing performance and flow experience in *Semideus* ($r = .43$ $p < .001$). This result is consistent with earlier studies that have indicated that level of flow can be used to predict learning (e.g. [47]).

To summarize, the results presented above support Hypothesis 1a and indicate that *Semideus* can be used to assess students' rational number conceptual knowledge. Consistent with Hypothesis 1b, the overall accuracy in the *Semideus* game (M = 70.19%, SD = 16.24%) correlated significantly with procedural fraction knowledge (M = 61.45%, SD=28.71%), $r = .48$, $p < .001$. Further analysis revealed that procedural fraction knowledge correlated most strongly with estimation task accuracy ($r = .52$, $p<.001$) and ordering task accuracy ($r = .50$, $p < .001$). In general, these results are in line with previous research (e.g. [31, 35]), suggesting that instructional approaches that focus on conceptual knowledge facilitate also procedural fraction knowledge.

3.2 Intervention study

In this section the results of the intervention are presented. Two main aims of the intervention were to study how rational number conceptual knowledge influences success in *Wuzzit Trouble* and how playing *Wuzzit Trouble* influences the development of rational number conceptual knowledge.

Points achieved in *Wuzzit Trouble* did not correlate significantly with pre-test playing performance, $r = .44$ $p >.05$. As shown on Figure 4, in the pre-test the accuracy of rational number conceptual knowledge of the control group (M = 84.14, SD = 9.51) was significantly higher than the accuracy of the treatment group (M = 75.15, SD = 10.34), $t(23) = 2.26$, $p < .05$). However, Figure 4 illustrates that the treatment group (M = 82.63, SD = 7.31) almost catched up the control group (M = 85.21, SD = 9.71) in the post-test. In order to study the effectiveness of the treatment (playing *Wuzzit Trouble*), the gain scores were calculated by subtracting the pre-test scores from the post-test scores.

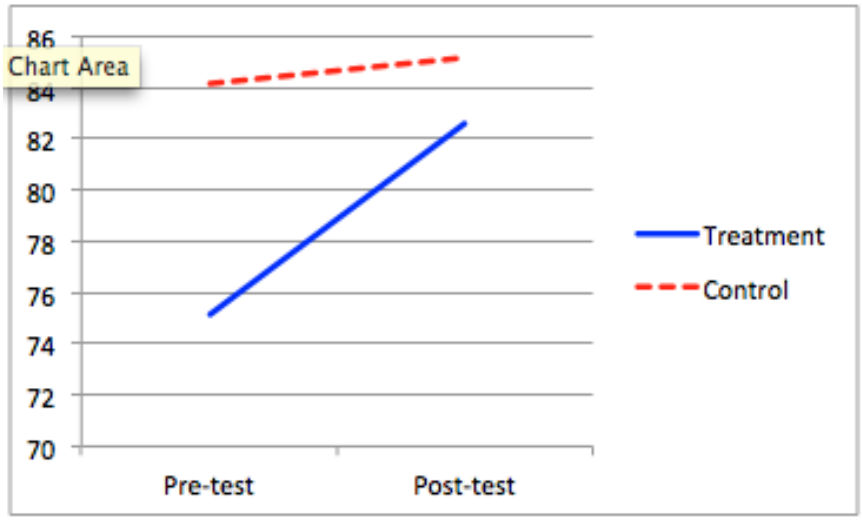

Figure 4. The mean accuracy of pre-test and post-test by conditions.

The t-test analysis indicated that the gain of the treatment group (M = 7.47, SD = 4.31) was significantly higher than the gain of the control group (M = 1.07, SD = 3.93), t(23) = 3.87, p<.001. Based on the t-test results, we calculated the effect size of the intervention. Cohen's d was very high, d = 1.61 indicating that playing *Wuzzit Trouble* influenced significantly the development of students' rational number conceptual knowledge. The more detailed analysis of the game data revealed that largest gains were in density (21%) and ordering (17%) task types while the estimation accuracy gain was only 2%. Cleatly, this was because the estimation accuracy was reasonable high already in the pre-test (M = 88 .6%, SD =6.4%) and there was not much room to improve. The control group got a little bit better in ordering tasks.

The small sample size means that, on their own, these results have to be viewed as at most suggestive. However, it gains more significance in the light of similar findings from another (small) study ([22], in this special issue).

Although the *Semideus* game does not emphasize the speed of solving tasks, we explored task durations between pre- and post-tests. The correlation analyses revealed that task duration did not correlate with task accuracy at all. However, as shown on Figure 5, in the pre-test the control group (M =31.3, SD = 9.23) solved the tasks little bit faster than the treatment group (M = 35.8, SD = 8.54), but the difference was not statistically significant, t(23) = 1.23, p > .05). Nevertheless, figure 5 illustrates that the treatment group (M = 25.58, SD = 7.36) almost equaled the control group (M = 26.65, SD = 8.19) in the post-test. As we can see, both groups solved the task much faster in the post-test. This can be partly explained with getting familiar with the user interface and playing the same game again, but it is reasonable to assume that the treatment group might have benefitted also from playing *Wuzzit Trouble*. (This is totally speculative and more research with a bigger sample is needed to study the learning outcomes of *Wuzzit Trouble* and the meaning of speed in *Semideus*.)

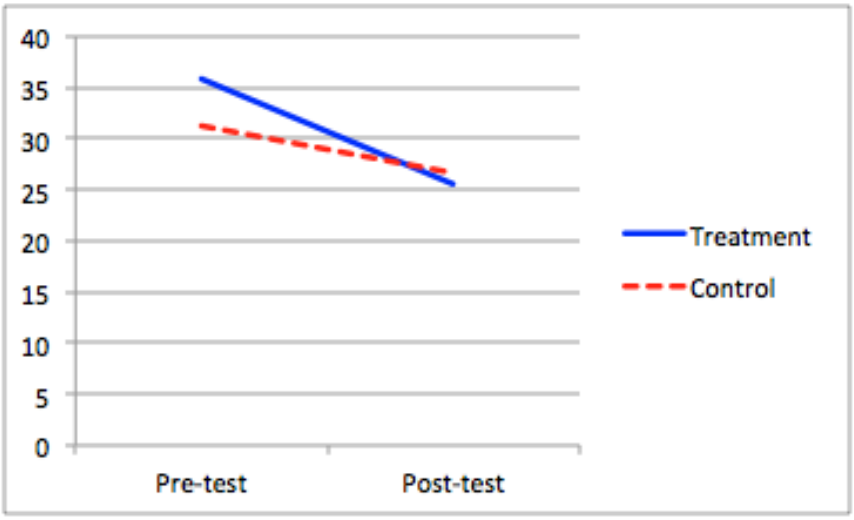

Figure 5. The mean mean task duration of pre-test and post-test by conditions.

As in the validation study, we evaluated user engagement in terms of flow experience and playability. Table 4 shows that players experienced fairly strong flow experiences in *Wuzzit Trouble* (M = 4.01, SD = 1.33). Flow was a little bit weaker than in the *Semideus* game, but in general the experiences were quite similar.

Table 4. Mean values of flow experience and playability in the *Wuzzit Trouble* game.

	M	**SD**
Flow construct	4.01	1.33
• Challenge – skill balance	4.19	1.65
• **Clear goals**	**4.69**	**1.56**
• **Immediate & unambiguous feedback**	**4.67**	**1.59**
• Action awareness merging	4.11	1.77
• Time distortion	3.74	1.89
• Sense of control	3.89	1.72
• Concentration	3.72	1.63
• **Loss of self-consciousness**	**4.31**	**1.86**
• Autotelic experience	3.06	1.58
Playability construct	3.99	1.53

Note. Three highest scoring flow dimensions are bolded.

As we can see, clear goals (M = 4.69, SD = 1.56), immediate & unambiguous feedback (M = 4.67, SD = 1.59), and loss of self-consciousness (M = 4.31, SD = 1.86) flow dimensions scored highest, just as they did in *Semideus*. Similarly the autotelic experience dimension scored lowest (M = 3.06, SD = 1.58). Although players did not consider the game as an intrinsically rewarding experience (they did not realize that it was a representation of some deep mathematics, instead viewing it as "just a game"—which is what the game designers sets out to achieve), most of the players reported that they wanted to achieve all three stars from each level (M = 4.31, SD = 1.74). Based on these findings we assume that players were engaged and tried to perform well in the game.

The correlation analysis indicated that there was a significant relationship between flow experience and playability ($r = .72$, $p < .001$), which means that the playability of the game can influence remarkably on overall user experience. As opposed to the results of the *Semideus* game, the analysis revealed that the overall playing performance (points achieved in *Wuzzit Trouble*) and flow experience did not correlate ($r = .27$ $p > .05$).

4. Discussion and conclusion

The aim of this study was to investigate the usefulness of digital video games in mathematics learning and assessment. To accurately assess learning, it is important to develop non-intrusive and engaging assessment methods that provide valid information about students' mathematical competence. In our validation study, we developed game-based measures for assessing conceptual rational number knowledge and evaluated the validity and reliability of the *Semideus* game as an assessment instrument. As expected, the evaluation results indicated that *Semideus* could be used to assess students' conceptual rational number knowledge at least in summative manner. Moreover, our study showed that the *Semideus* game can be used to identify whole number bias related misconceptions, although more accurate identification solutions needs to be developed.
Our intervention study demonstrated that a game (in this case *Semideus*) can be used as a test instrument in experimental settings and even relative short game based mathematics interventions can be effective. As we expected the students benefited significantly of the three hour *Wuzzit Trouble* intervention. This was the case, even though *Semideus* assesses conceptual rational number knowledge and the gameplay of *Wuzzit Trouble* is founded on whole number arithmetic. This too was expected by researchers, since both games develop number sense and general numeric problem solving ability, important skills that apply to both integer arithmetic and fraction arithmetic, but others familiar only with more traditional math learning games might be surprised. (More on this below.)

4.1 Theoretical and practical implications

The findings of this study have theoretical and practical implications. In general, the results enrich the previous research on children's understanding of rational numbers and game-based assessment. The validation study showed that the gamified number line estimation tasks, the magnitude comparing tasks, and the magnitude ordering tasks could be used to validly assess students' conceptual rational number knowledge. Consistent with previous research [e.g. 31, 35] we found positive interaction between understanding of rational number magnitudes and rational number arithmetic performance. This consistent interaction strengthens the positive results of the *Semideus* validation study. In general, these results indicate that instructional approaches that emphasize understanding of rational number magnitudes (conceptual knowledge) support also arithmetic performance (procedural knowledge) and thus they should be more widely used.
On the other hand, the intervention study showed that the whole number arithmetic training (playing of the *Wuzzit Trouble* game) enhanced students' conceptual rational number knowledge. We want to emphasize that even relatively short periods (of the order of two hours play spread over a few weeks) of engagement with a well-designed math learning game can lead to significant improvements in mathematical competence, with transfer to another area of mathematics (in this case from whole numbers to rational numbers). This finding is consistent with integrated theory of numerical development [23], which emphasizes that, in addition to differences between whole and rational numbers, whole numbers and rational numbers also have important commonalities that are central to general mathematical competence. In spite of the differences between whole and rational numbers, the understanding of whole number magnitudes and fluency in whole number arithmetic, especially fluency in division, create a foundation for understanding rational number magnitudes. This may help explain why playing of the *Wuzzit Trouble* game enhanced the playing performance in the *Semideus* game.
Furthermore, the evaluation of students' flow experiences indicated that games can engage students in learning mathematics. From an assessment point of view, an important finding was that students could concentrate on playing *Semideus* in a way that they did not consider how others might have evaluated their playing/mathematics performance. In other words, students were able to do their best without the familiar pressures normally associated with exams and assessment. Thus, game-based assesment solutions could indirectly support school satisfaction.

4.2 Limitations and future work

Our study has some limitations that call for more reseacrh on the topic. First, the sample size was small, reducing the power of the study. This is the case especially in the *Wuzzit Trouble* study. However, the results are consistent with the study of Pope & Mangram [22] that used written test instruments to study the effectiveness of the *Wuzzit Trouble* game. Though both studies had fairly small sample sizes, the learning outcomes observed were remarkably similar, suggesting that the outcomes observed were genuine — certainly enough to warrant further, more extensive studies, which are already underway.
Second, at the beginning of the study, the rational number knowledge level of the treatment group was much lower than the control group, which may have increased the effect size of the intervention.
Third, due to scheduling restrictions at the participating schools, we could include only one level that was intended to teach the use of the *Semideus* game to participants. Thus, it is possible that some of the players had difficulties controlling the game accurately in the beginning, and that may have affected negatively their overall performance. However, this should not have been a major problem, because participants appreciated the playability of the game.
Fourth, some of the players faced technical problems with the *Semideus* game because their iPads had the wrong iOS version. The players who could not play the game adequately due to technical problems were removed from the analysis.
Finally, the study was spread to period of two months and thus there is a risk that the game playing is not the only factor that has affected the results. The treatment group may have learned rational number knowledge during their free time (rational numbers were not taught in school during the study).
Because of these limitations, more research with bigger sample sizes on the topic is needed.
While our study used *Semideus* as a pre- and post-test and *Wuzzit Trouble* as an intervention, we cannot assume that similar results would have been obtained with the roles reversed, though we suspect that would be the case, and we intend to investigate this in a future study.
Overall, our results about game-based assessment are promising, and we believe that the *Semideus* game can be developed to support also formative assessment approach. In fact, based on the results of this pilot study, we have already made several modifications to *Semideus*. In terms of learning analytics and formative assessment, the most important features that we have implemented are adaptive content, visual aids, and a learning companion. In the future we will study the learning effectiveness of the improved *Semideus* game and validate its assessment power in terms of formative assessment. In fact, we are currently exploring more valid ways to detect misconceptions and ways to visualize players' competencies in real time. Visualizations are provided both for players and teachers. Thus, one of our future aims is to study how teachers can use learning analytics effectively in individualizing teaching in the classrooms.

5. Conclusion

This study showed that well designed digital games can be used for learning and assessment purposes. Considerable mathematical and pedagogical thought went into the design of *Semideus* and *Wuzzit Trouble*. The developers of these games began with a dynamic, visual representation of the underlying mathematics (the number line and integer arithmetic, respectively) and built a game around it. With both of these games, *to solve the game puzzles or challenges is to understand and solve the underlying mathematical problem*. We argue that in order to create successful mathematics games learning science should inform the game design and mathematics should be well integrated to gameplay. Moreover, we want to emphasize that we believe that mathematics games can be more effective if a teacher or an instructor is involved in the game based learning process (assuming that an appropriate pedagogical approach is used).
In fact, our long-term goal is to produce interactive learning experiences and learning analytics tools for teachers based on in-game measures that can predict development of students' domain specific knowledge and reveal students' misconceptions and weaknesses as well as their strengths. Such information will be very useful for teachers in individualizing teaching.
Furthermore, one of the more interesting possibilities that games provide relies on assessing children's conceptual development and mathematical thinking in larger contexts. The big data sets that can be collected with games make it possible to uncover dependencies and patterns behind conceptual change, and compare the performance with other groups, including between countries.

These comparisons could provide totally new insights for curriculum development and assessment. When we can provide valid analysis about learning processes, conceptual changes, and learning assessment, we can provide something new and complementary to current assessment methods such as PISA, TIMSS and PIRLS.

Acknowledgements

This research was funded in part by Academy of Finland (The Future of Learning, Knowledge and Skills programme).

References

[1] Parsons, S., Bynner, J., Does numeracy matter more? London: National Research and Development Centre for Adult Literacy and Numeracy, 2005.
[2] Gee, J. P., What Video Games Have to Teach Us about Learning and Literacy. NY: Macmillan, 2003.
[3] Squire, K., Video Games Literacy: a literacy of expertise, In J. Coiro, M. Knobel, D. Leu & C. Lankshear (Eds) Handbook of Research on New Media Literacies. New York: Macmillan, 2008.
[4] Devlin, K. J., Mathematics Education for a New Era: Video Games as a Medium for Learning. AK Peters Ltd, 2011. http://dx.doi.org/10.1201/b10816
[5] Jere-Folotiya, J., Chansa-Kabali, T., Munachaka, J. C., Sampa, F., Yalukanda, C., Westerholm, J., Lyytinen, H., The effect of using a mobile literacy game to improve literacy levels of grade one students in Zambian schools. Educational Technology Research and Development, 62(4), 417-436, 2014. http://dx.doi.org/10.1007/s11423-014-9342-9
[6] Clark, D. B., Tanner-Smith, E. E., Killingsworth, S. S., Digital Games, Design, and Learning A Systematic Review and Meta-Analysis. Review of educational research, 0034654315582065, 2015.
[7] Wouters, P., van Nimwegen, C., van Oostendorp, H., van der Spek, E. D., A meta-analysis of the cognitive and motivational effects of serious games. Journal of Educational Psychology, 105(2), 249, 2013.
[8] Shin, N., Sutherland, L. M., Norris, C. A., Soloway, E., Effects of game technology on elementary student learning in mathematics. British Journal of Educational Technology, 43(4), 540-560, 2012. http://dx.doi.org/10.1111/j.1467-8535.2011.01197.x
[9] Kiili, K., Perttula, A., Lindstedt, A., Arnab, S., Suominen, M., Flow Experience as a Quality Measure in Evaluating Physically Activating Collaborative Serious Games. International Journal of Serious Games, 1(3), 2014. http://dx.doi.org/10.17083/ijsg.v1i3.23
[10] Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., Boyle, J. M., A systematic literature review of empirical evidence on computer games and serious games. Computers & Education, 59(2), 661-686, 2012. http://dx.doi.org/10.1016/j.compedu.2012.03.004
[11] Whitton, N., Game engagement theory and adult learning. Simulation & Gaming, 42(5) 596–609, 2011. http://dx.doi.org/10.1177/1046878110378587
[12] Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., van Ruijven, T., Wenzler, I., The research and evaluation of serious games: Toward a comprehensive methodology. BJET, 2013.
[13] Chang, M., Evans, M. A., Kim, S., Norton, A., Samur, Y., Differential effects of learning games on mathematics proficiency. Educational Media International, 52(1), 47-57, 2015. http://dx.doi.org/10.1080/09523987.2015.1005427
[14] Adams, D. M., Mayer, R. E., MacNamara, A., Koenig, A., Wainess, R., Narrative games for learning: Testing the discovery and narrative hypotheses. Journal of educational psychology, 104(1), 235, 2012. http://dx.doi.org/10.1037/a0025595
[15] Bellotti, F., Berta, R., De Gloria, A., Designing effective serious games: opportunities and challenges for research. International Journal of Emerging Technologies in Learning (iJET), 5(SI3), 22-35, 2010. http://dx.doi.org/10.3991/ijet.v5s3.1500
[16] Ketamo, H., Devlin, K., Replacing PISA With Global Game Based Assessment. In ECGBL2014-8th European Conference on Games Based Learning: ECGBL2014 (p. 258). Academic Conferences and Publishing International, 2014.

[17] Riconscente, M., Mobile learning game improves 5th graders' fractions knowledge and attitudes. Los Angeles: GameDesk Institute, 2011.
[18] Shute, V. J., Ventura, M., Kim, Y. J. , Assessment and learning of informal physics in Newton's playground. Journal of Educational Research, 106(6), 423–430, 2013. http://dx.doi.org/10.1080/00220671.2013.832970
[19] Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R., Assessment in and of serious games: an overview. Advances in Human-Computer Interaction, 1, 2013. http://dx.doi.org/10.1155/2013/136864
[20] Shute, V. J., Ventura, M., Bauer, M., Zapata-Rivera, D., Melding the power of serious games and embedded assessment to monitor and foster learning. Serious games: Mechanisms and effects, 2, 295-321, 2009.
[21] The US National Research Council, Adding It Up: Helping Children Learn Mathematics. Washington, DC: National Academies Press: National Academy Press, pp. 1–462, 2001.
[22] Pope, H., Mangram, C., Wuzzit Trouble: The Influence of a Digital Math Game on Student Number Sense. Internationa Journal of Serious Games, Vol. 2, Nr. 4, December 2015.
[23] Siegler, R. S., Thompson, C. A., Schneider, M., An integrated theory of whole number and fractions development. Cognitive psychology, 62(4), 273-296, 2011. http://dx.doi.org/10.1016/j.cogpsych.2011.03.001
[24] Torbeyns, J., Schneider, M., Xin, Z., Siegler, R. S., Bridging the gap: Fraction understanding is central to mathematics achievement in students from three different continents. Learning and Instruction, 37, 5-13, 2015. http://dx.doi.org/10.1016/j.learninstruc.2014.03.002
[25] Van Hoof, J. Verschaffel, L., Van Dooren, W. Number sense in the transition from natural to rational numbers. In D.M. Gomez Rojas (Chair), Living apart together? The long and winding road from natural to rational number understanding. Symposium conducted at the biennial meeting of the European Association for Research on Learning and Instruction (EARLI), 2015.
[26] Bailey, D. H., Hoard, M. K., Nugent, L., Geary, D. C., Competence with fractions predicts gains in mathematics achievement. Journal of experimental child psychology, 113(3), 447-455, 2012. http://dx.doi.org/10.1016/j.jecp.2012.06.004
[27] Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., ... Chen, M., Early predictors of high school mathematics achievement. Psychological science, 23(7), 691-697, 2012. http://dx.doi.org/10.1177/0956797612440101
[28] Siegler, R. S., Fazio, L. K., Bailey, D. H., Zhou, X., Fractions: the new frontier for theories of numerical development. Trends in cognitive sciences,17(1), 13-19, 2013. http://dx.doi.org/10.1016/j.tics.2012.11.004
[29] Stafylidou, S., Vosniadou, S., The development of students' understanding of the numerical value of fractions. Learning and instruction,14(5), 503-518, 2004. http://dx.doi.org/10.1016/j.learninstruc.2004.06.015
[30] Vamvakoussi, X., Vosniadou, S., How many decimals are there between two fractions? Aspects of secondary school students' understanding of rational numbers and their notation. Cognition and instruction, 28(2), 181-209, 2010. http://dx.doi.org/10.1080/07370001003676603
[31] Bailey, D. H., Siegler, R. S., Geary, D. C., Early predictors of middle school fraction knowledge. Developmental science, 17(5), 775-785, 2014. http://dx.doi.org/10.1111/desc.12155
[32] DeWolf, M., Vosniadou, S., The representation of fraction magnitudes and the whole number bias reconsidered. Learning and Instruction, 37, 39-49, 2015. http://dx.doi.org/10.1016/j.learninstruc.2014.07.002
[33] Ni, Y., Zhou, Y. D., Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. Educational Psychologist, 40(1), 27-52, 2005. http://dx.doi.org/10.1207/s15326985ep4001_3
[34] Alibali, M. W., Sidney, P. G., Variability in the natural number bias: Who, when, how, and why. Learning and Instruction, 37, 56-61, 2015. http://dx.doi.org/10.1016/j.learninstruc.2015.01.003
[35] Fuchs, L.S. et al., Improving at-risk learners' understanding of fractions. Journal of Educational Psychology, 105, 683, 2013. http://dx.doi.org/10.1037/a0032446
[36] Barth, H.C., Paladino, A.M., The development of numerical estimation: Evidence against a representational shift, Developmental science, 14, 125–135, 2011. http://dx.doi.org/10.1111/j.1467-7687.2010.00962.x

[37] Link, T., Huber, S., Nuerk, H.-C., Moeller, K., Unbounding the mental number line - new evidence on children's spatial representation of numbers, Frontiers in Psychology, 4 (1021), 2014. http://dx.doi.org/10.3389/fpsyg.2013.01021
[38] Durkin, K., Rittle-Johnson, B., Diagnosing misconceptions: Revealing changing decimal fraction knowledge. Learning and Instruction, 37, 21-29, 2015. http://dx.doi.org/10.1016/j.learninstruc.2014.08.003
[39] Pouw, W. T., Van Gog, T., Paas, F., An embedded and embodied cognition review of instructional manipulatives. Educational Psychology Review,26(1), 51-72, 2014. http://dx.doi.org/10.1007/s10648-014-9255-5
[40] Csikszentmihalyi, M., Flow: The Psychology of Optimal Experience. Harper Perennial, New York, 1991.
[41] McMullen, J., Laakkonen, E., Hannula-Sormunen, M., Lehtinen, E., Modeling the developmental trajectories of rational number concept (s).Learning and Instruction, 37, 14-20, 2015. http://dx.doi.org/10.1016/j.learninstruc.2013.12.004
[42] Razali, N. M., Wah, Y. B., Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21-33, 2011.
[43] Doane, D. P., Seward, L. E., Measuring skewness: a forgotten statistic. Journal of Statistics Education, 19(2), 1-18, 2011.
[44] Kerlinger, Kerlinger, F. Foundations of behavioral research (3rd ed.). Harcourt Brace Jovanovich, Orlando, FL, 1986.
[45] Depaepe, F., Torbeyns, J., Vermeersch, N., Janssens, D., Janssen, R., Kelchtermans, G., Van Dooren, W., Teachers' content and pedagogical content knowledge on rational numbers: A comparison of prospective elementary and lower secondary school teachers. Teaching and Teacher Education, 47, 82-92, 2015. http://dx.doi.org/10.1016/j.tate.2014.12.009
[46] Anastasi, A. Psychological testing (6th ed.). Macmillan, New York, 1988.