Evaluating Cognitive and Affective Outcomes of a Digital Game-Based Math Test

K. Kiili, and H. Ketamo

Abstract— Even though digital learning games have become common in education, relatively little is known about the usefulness of game-based assessment. This paper aims to explore if a game-based math test can provide added value to math education with respect to cognitive and affective outcomes. We used in-game measures, embedded in the game called Semideus Exam, focusing on conceptual fraction knowledge. In order to validate the game-based assessment approach, we compared the cognitive outcomes of fifty-one Finnish sixth graders, who completed both paper-based and game-based math tests in a randomized order. In addition, the students' test anxiety and flow experience were measured to evaluate the affective outcomes. The results indicate that the game-based test scores correlated significantly with the paper-based test scores suggesting that the game-based assessment was successfully implemented and the game provided comparable data with the paper-based test approach. More importantly, the results revealed that game-based assessment lowered test anxiety and increased engagement which is likely to decrease assessment bias caused by test anxiety. In addition, the results show that earlier playing experience and gender did not influence the game-based test score suggesting fairness of the game-based assessment approach. Although we identified several benefits of the game-based assessment approach, more evidence is needed on the usefulness and fairness of game-based assessments.

Index Terms— Educational games, Computer Uses in Education, Automatic assessment tools, Education

1 INTRODUCTION

Using games in learning already has a long history. Researchers have increasingly argued that the meaning-making practices taking place when people engage in digital games define a form of literacy that is potentially better suited for addressing the needs of the learners in the 21st century [1], [2]. Previous research has provided some evidence that educational games can support learning [3]. Furthermore, the use of game-based tasks has proven to be an effective approach to learning in the domain of mathematics education [4], [5], [6], [7]. The intrinsic appeal of games can be explained by their ability to satisfy basic psychological needs for competence, autonomy, and relatedness - which when experienced increase students' motivation and engagement [8].

Recently, researchers have also sought to use games as assessment tools [9], [10], [11]. Learning in games has been traditionally evaluated indirectly and separately from the actual gameplay [12]. In fact, according to Kim and Shute [13] game-based assessment is fairly new in education and there is little evidence of how to maximize the effectiveness and validity of game-based assessment without losing engagement. This paper aims at describing children's conceptual knowledge of fraction numbers. Another aim is to explore to what extend game-based tests can provide added value to math education with respect to cognitive and affective outcomes.

- *K. Kiili is with the Tampere University of Technology, P.O. Box 300, FI-28101 Pori, Finland. E-mail: kristian.kiili@tut.fi.*
- *H. Ketamo is with the Satakunta University of Applied Sciences, Tiedepuisto 3 FI-28600 Pori, Finland. E-mail: harri.ketamo@samk.fi.*

*Please note that all acknowledgments should be placed at the end of the paper, before the bibliography (**note that corresponding authorship is not noted in affiliation box, but in acknowledgment section**).*

1.1 Game-Based Assessment

According to Pellegrino, Dibello, and Goldman [14] assessment can be considered a tool designed to observe students' behavior and to produce data that can be used to draw reasonable inferences from about their competencies. Assessment which aims to improve learning is called formative assessment and assessment which is used to assess learning outcomes for such purposes as grading, promotion, and placement is called summative assessment [15]. Test anxiety, referring to individual differences in the extent to which assessments are appraised as threatening [16], is common especially in summative assessment situations. Given the increased testing of school-aged children in several countries, there is a need for developing assessment solutions that lower test anxiety. In general, games can provide an appropriate context for both summative and formative assessment and may decrease test anxiety. Understanding the context of assessment also helps to understand the claims made from the assessment and reach the desired inference [9]. The present study focuses on summative assessment.

It has been argued that conventional educational measures are not suitable for educational games, since conventional measures are usually highly invasive and compromise flow experience [17], [18]. Consequently, Kim and Shute [13] have argued that a well-designed game-based assessment should be enjoyable to the players while providing valid evidence of the players' proficiencies. In order to achieve these goals, the adoption of stealth (embedded) assessment that aims to blur the distinction between assessment and learning has been suggested for educational games [12]. Stealth assessment is an evidence-based approach where the tasks that a student performs are highly interactive, assessment is well integrated to the gameplay and assessment is carried out non-invasively [12], [19]. Stealth assessment has been utilized and studied in several games [10], [20]. Although stealth assessment is expecially suitable for formative assessment, we applied a similar evidence-centered design approach in implementing in-game measures for a game-based test studied in this research. This approach was selected because this study is a part of a process in which we are developing a Semideus School game for formative assessment purposes.

The validity of assessment depends on the conceptual frameworks and processes which have been used to guide the development of assessment [14]. For example, in this study we utilized an assessment triangle [14] which provides a useful framework for designing game-based assessment solutions and for establishing their validity. The vertices of the assessment triangle represent three interconnected key elements underlying any assessment: 1) a model of student cognition and learning in the domain of assessment, 2) a set of assumptions about the observations that will provide evidence of students' competencies relative to the cognitive model, and 3) an interpretation process for making sense of the evidence in regard to the assessment goals.

We want to emphasize that it is very important to ensure that the used game mechanics are founded on the cognitive model and produce valid evidence of targeted competencies. According to Kim and Shute [13] fairness should be addressed with a particular care in game-based assessment because games may function differently across subgroups (e.g., male vs. female, gamers vs. non-gamers). We should be aware that students are not always familiar with different kinds of digital devices, game mechanics and user interface solutions that may cause bias in assessment. Thus, it is also very important as part of a game-based assessment process to investigate how well the players master the user interface of the game and how deeply the game engages the players.

1.2 Cognitive Model about Conceptual Fraction Knowledge

In developing games for learning and assessing fraction knowledge, we have also combined a cognitive model about fractions. Fractions are frequently encountered in daily life. Therefore, a failure to understand the basic concept of fractions can cause difficulties in qualifying for any level of skilled labour [21]. However, there is a great deal of evidence that children find understanding of fractions very difficult. Many children fail to perform adequately even in simple fraction tasks even after a considerable amount of mathematics instruction [22], [23], [24]. A crucial part of fraction knowledge is the successful representation of fraction magnitude (reflecting the relation between the numerator and denominator).

According to conceptual change theories, children form an initial conception of numbers as counting units before they encounter fractions. As a result, later on they draw heavily on this initial understanding to make sense of fractions [24], [25]. The phenomenon called whole number bias originates from people's false belief that all properties of whole numbers can be applied to rational numbers [26]. According to Alibaliv and Sidney [27] whole number bias has been observed in elementary school students, in high school students, in adults, and even in expert mathematicians. Whole number bias has been found to cause difficulties in reasoning the size of fractions [28]. For example, when comparing fraction magnitudes, people sometimes think that the fraction that has larger whole numbers as its parts is larger (e.g. one may think that $1/5$ is larger than $1/3$, because 5 is larger than 3). In fact, fraction comparison tasks are a common way of detecting whole number bias and studying understanding of magnitudes. Research on numerical cognition has shown that the so-called distance effect indicates a successful representation of fraction magnitude [29]. According to distance effect number comparisons with a small distance are more difficult and take longer to be compared than number comparisons with a larger distance [30].

Siegler, Thompson, and Schneider [23] have proposed an integrated theory of numerical development that emphasizes continuity between the acquisition of understanding of whole numbers and fractions. According to their theory significant conceptual change is required to understand that fractions, like whole numbers, represent magnitudes that can be located on number lines. Accordingly, recent findings have suggested that instructional interventions which aim to support conceptual change should target the learners' interpretation of fractions as magnitudes by practicing them on number lines [31], [32], [33]. A common way to train and assess conceptual fraction knowledge is a number line estimation task, which is founded on the concept of a mental number line. The mental number line is often used as a metaphor to describe our spatially oriented mental representation of number magnitude [34].

1.3 The Present Study and Hypotheses

The present study is a part of an ongoing project in which we have developed a game-based rational number research engine called Semideus. The main purpose of the current study is to demonstrate the validity and usefulness of a game created with the Semideus engine in assessing the students' conceptual fraction knowledge. In fact, a crucial issue which has rarely been investigated is how the outcomes of game-based math tests differ from the outcomes of traditional paper-based math tests. In other words, to what extent a game-based approach adds value to mathematics education and whether game-based assessment causes assessment bias. To address this question, we studied Finnish sixth graders' conceptual knowledge of fraction with both a game-based and a paper-based math test consisting of 0–1 number line estimation and 0–1 magnitude comparison tasks. Moreover, the aim of the present work is to illustrate the affective outcomes of using game-based tests in mathematics education. We approached the affective outcomes through two lenses: flow experience and text anxiety.

We made altogether nine predictions. In terms of convergent construct validity, we expected that the game-based test would provide comparable results with the paper-based test with respect to fraction estimation accuracy and fraction comparison performance (Hypothesis 1). In addition, we expected to observe similar, general rational number trends that previous numerical cognition research has identified, suggesting content validity (Hypothesis 2). Because both number line estimation and magnitude comparison are believed to reflect understanding of numerical magnitudes [23] we expected that performances in these tasks would correlate (Hypothesis 2a). We also anticipated to observe a standard distance effect [29], [30] assuming that fractions that are near each other are more difficult and take longer to compare than fractions with larger distance (Hypothesis 2b). Moreover, previous research has shown that fractions consistent with whole number ordering are usually mastered better than fractions inconsistent with whole number ordering [28], [35]. Thus, we expected that students would solve consistent whole number ordering comparison tasks better than inconsistent comparison tasks (Hypothesis 2c). Finally, because understanding of fraction magnitudes should be strongly related to overall mathematical knowledge [23] we hypothesized that knowledge of fraction magnitudes would correlate highly with overall mathematical achievement measured by a previous math grade, thus, suggesting criterion validity (Hypothesis 2d).

In order to examine consequential validity (the positive and negative consequences of assessment for teaching and learning), the rest of the hypotheses were related to the affective outcomes of game-based assessment. Previous research has shown that the use of games in mathematics education can be beneficial for affective outcomes [36], [37]. Based on this, we hypothesized that students would experience significantly lower test anxiety (Hypothesis 3a) and significantly higher flow experience (Hypothesis 3b), when they engage themselves with the game-based math test. In terms of flow theory [38], [39], our game-based test facilitates positive affective outcomes because it provides concrete short-term goals with manageable rules and clear and immediate feedback on the player's performance (e.g., running scores, accuracy information, progress meter, health meter, achieved stars) and rich multimodal information that screens out distractions and facilitates concentration. Moreover, the player's avatar, non-player characters and the general storyline facilitate concentration and immersion [8]. Previous research has shown that flow experience is related to high performance [40], [41]. Therefore, it is probable that flow experience will have a positive relation with the overall game-based test performance (Hypothesis 3c). Moreover, we expected that test anxiety would be negatively associated with test performance (Hypothesis 3d).

According to Kim and Shute [13], the fairness of game-based assessment should be addressed with particular care because games may function differently across subgroups (e.g. male vs. female, gamers vs. non-gamers). An exploratory approach was used to investigate the fairness of the game-based test because we did not find any theoretical basis for setting up a hypothesis.

2 Method

2.1 Participants and Research Design

Three Finnish sixth grade classes participated in the study. Sixty students were involved in the study and 51 of them followed the requested protocol (i.e. conducted the tests and returned all the questionnaires). Thus, the final sample used in the analyses consisted of 51 students with 28 females and 23 males. The mean age (SD) of the students was 11.92 (0.34). All the participants conducted both the paper-based test and the game-based test. A counter-balanced measures design was used to reduce the influence of the test order on the results. The participants were randomly divided into two groups with one group starting with the paper-based test and the other group with the game-based test. Exactly same fraction tasks were used in the both tests.

2.2 Description of the Semideus Research Engine

Semideus is a rational number research engine that can be used to create games that aim at supporting and assessing the development of children's conceptual rational number knowledge. The engine was used to implement a game (a research instrument), Semideus Exam, which was used in this study. The gameplay is founded on tasks that require working with number lines implemented as walkable platforms of a mountain. In the game the player controls a character, called Semideus, who tries to collect gold coins that a goblin has stolen from Zeus. Semideus has discovered the locations of the hidden coins, encrypted in mathematical symbols, and must race the goblin to retrieve the coins from the trails of Mount Olympus.

Fig. 1 shows an example of a number line estimation task and Fig. 2 an example of a magnitude comparison task which were used in the present study. In the number line estimation tasks the player tries to locate a gold coin based on a given number.

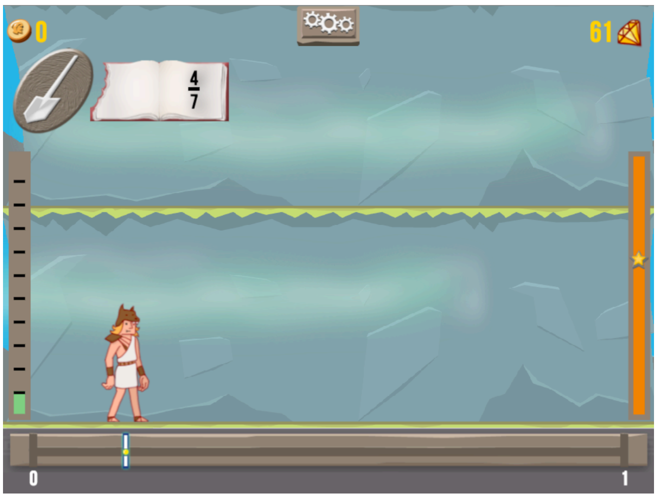

Fig. 1. An example of a number line estimation task in which the player should dig up a coin from spot 4/7.

In magnitude comparison levels, the player has to compare two stones with values on them by arranging the stones in an ascending order with regard to the numerical magnitudes depicted on them. The exact spot on the number line (ranging from smaller numbers on the left to larger numbers on the right) does not matter as long as the order of the stones is correct. The player can also pile up the stones, if he or she thinks that the magnitudes are equivalent.

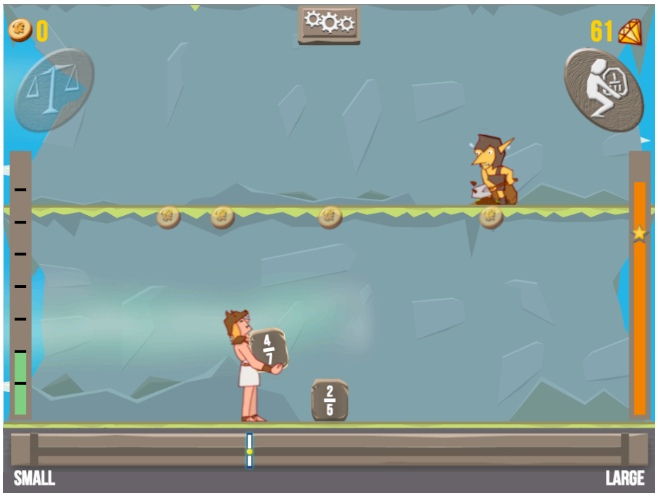

Fig. 2. An example of a magnitude comparison task in which the player should arrange 4/7 and 2/5 in an ascending order.

2.2.1 The Level Structure and Tasks of the Game

The Semideus Exam game was a web-based application played through an iPad browser in a full screen mode. The game world consisted of seven levels. The first two levels were designed for onboarding. The aim of the onboarding levels was to familiarize players with the basic rules and controlling mechanics of the game. The first onboarding level included seven whole number estimation items (number line ranging from 0 to 100) and the second level six whole number comparison items (three digit numbers). In the first three whole number estimation tasks (Level 1) the correct answer/position was indicated by a coin. In this approach a player should immediately understand that the correct solution indicates the position of the coin on the number line. Moreover, these items were used to evaluate how well players mastered character movements carried out by tilting the tablet.

The actual test phase consisted of five levels (Levels 3-7). The tasks of the test levels were designed according to the cogntitive model of conceptual fraction knowledge. The test phase started with ten fraction estimation items (Level 3: number line ranging from 0 to 1). The fractions were selected in such a way that they covered the whole range of the number line. The second and the major part of the test phase consisted of fraction comparison items (Levels 4-7). The order of the fraction items within levels was randomized. The test included altogether 28 fraction comparison tasks from which 14 were consistent with whole number ordering and 14 were inconsistent with whole number ordering (Table 1).

TABLE 1
FRACTION COMPARISON PAIRS

	Consistent			Inconsistent		
Small distance	1/2	vs.	4/7	2/5	vs.	1/2
	2/3	vs.	4/5	2/3	vs.	4/9
	5/9	vs.	3/8	2/9	vs.	1/3
	3/7	vs.	1/4	3/7	vs.	2/3
	4/7	vs.	2/5	5/8	vs.	3/4
	9/10	vs.	7/9	7/11	vs.	10/19
	3/11	vs.	5/12	6/19	vs.	5/12
Large distance	1/2	vs.	8/9	1/9	vs.	1/2
	2/5	vs.	5/7	3/9	vs.	3/4
	1/3	vs.	3/4	3/4	vs.	4/9
	6/11	vs.	11/13	3/9	vs.	2/3
	13/18	vs.	4/11	7/8	vs.	8/15
	8/17	vs.	1/16	12/22	vs.	11/12
	4/11	vs.	13/18	10/11	vs.	11/23

The numerical magnitude of each of the fractions was less than one. The comparison pairs were distance controlled. The distance between small distance gatecory comparison pairs was below 0.3 and the distance between large distance gatecory comparison pairs was over 0.3. The average distance of small distance pairs was .14 and the average of large distance pairs was .37. The average distance of consistent comparisons was .25 and the inconsistent comparisons .26. Moreover, the fraction comparison tasks were designed to have pairs that were on the different sides of ½ (6 pairs/whole number ordering condition), the same side of ½ (6 pairs/ whole number ordering condition), and include ½ (2 pairs/ whole number ordering condition). These factors were equated across conditions and counterbalanced so that half of the time the larger fraction stone appeared on top of the smaller stone and the other half time the smaller fraction stone appeared on top of the larger stone.

2.2.2 Configuration of the Game

The game was configured in such a way that each participant could play all the seven levels once starting from level one. Although a player could run out of virtual energy in a level (100 energy units in the beginning), he or she was still able to complete the level. However, on the mountain top (end of the level) the player did not earn the bonus which was given on the basis of remaining energy. The players were allowed only one answer to each task within the levels.

Feedback was shown after each answer. When the location was estimated accurately, the right position of the coin was indicated by a green marker on the number line and the estimation accuracy percentage was shown. Additionally, the player got 100-500 coins depending on the degree of correctness (over 98% correct = 500 coins; 95% - 98% = 300 coins; 92% - 94% = 100 coins). For inaccurate estimations the player lost 15 units of energy and the right place of the coin was shown with a green marker. After feedback, the player progressed to the next platform (i.e., faced the next task). In the magnitude comparison tasks the number of coins awarded to the player depended on the time taken to answer (under 9 seconds = 500 coins; 9-12 seconds = 400 coins; 12-16 seconds = 300 coins; 16-20 seconds = 200 coins; over 20 seconds = 100 coins; wrong answer = 0 coins and loss of 20 energy units).

After completing a level (reaching the top of the mountain) the player got additional feedback: 1-3 stars and earned coins were shown (i.e., one star for completing the level, one star for collecting enough coins, and one star from accuracy reflecting that enough energy was left). Additionally, a bonus was given on the basis of remaining energy (energy % * 500 coins).

The movement of the Semideus character was controlled by tilting the device to the left or right. Movement speed depended on the angle of tilting with a steeper angle resulting in higher speed.

2.3 Measures

A demographic survey was used to collect background information of the participants. The digital survey included questions about the participants' gender, previous game playing experience/frequency and last math grade in the school (possibly ranging from 4 to 10).

User experience was measured in terms of flow experience [38] and playability. Flow experience was measured with a 9-item digital questionnaire developed by the authors. The items included were derived from the flow scale which has been used in recent serious games studies [39], [42]. The included flow dimensions 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 3-item digital questionnaire developed by the authors. The dimensions included were ease of use, intuitiveness of the user interface and controlling accuracy of the user interface. A 6-point Likert-type response format was used in both the flow and playability questionnaires.

Test anxiety was measured with a modified 12-item version of the Children's Test Anxiety Scale that is appropriate for use with children aged 8–12 [43]. The statements were changed in such a way that they refer to the just finished test instead of general test experiences. The self-report scale consisted of three subscales: thoughts (e.g. “When I was taking the test, I was wondering if I will pass.”), off-task behaviours (e.g., “When I was taking the test, I was looking around the room.”) and autonomic reactions (e.g. “When I was taking the test my hand was shaking.”). The response format consisted of a four-point scale (almost never – almost always). Furthermore, a slightly modified version of the scale was used in the game-based test. For example, the statement “When I was taking the test, I was playing with my pencil or pen” was modified to “When I was taking the test, I was playing with my test equipment.”.

The Semideus Exam game continuously logged players' detailed playing behavior on a secured server according to a semantic model that is used to describe all the tasks of the game (based on the cognitive model of conceptual fraction knowledge). Based on the semantics, each task is tagged with keywords that describe the task in terms of rational number competencies. We have developed a data-analyzing tool that can be used to fetch data based on the tags. The same interface is used to create realtime learning analytics and automatic assessment reports. In this study, the teacher got an automatic report of the game-based test which included each player's estimation accuracy, overall comparison performance, performance in whole number ordering, consistent and inconsistent comparisons, possible flag about clear whole number bias, and possible flag about user interface mastering problems.

2.4 Procedure

The students were tested in three groups (based on classes) during a regular school day. First, the researchers introduced the game and explained the progress of the study to the students by reading aloud all the instructions. The teacher was instructed to explain to the students that their math test score will be based on both the paper-based and the game-based test performance. Second, the students got their personal passwords for the game. It was also required in all questionnaires and in the paper-based test for identification purposes. When the students had got their passwords and the progression of the test was clear to everybody, the first test began. Half of the students in each class did first the game-based test and the other half started with the paper-based test. The students got 30 minutes to complete the test. They were not allowed to discuss during the tests. After completing a test, the students were instructed to fill in the flow experience and the test anxiety questionnaires. Furthermore, after the game-based test, the players were instructed to fill in the playability questionnaire as well.

2.5 Analysis

Several variables related to test performance were computed. Number line estimation task accuracy was computed as 100*abs (correct value – estimated value)/the numerical range of the number line. The percentage of correctly solved tasks was computed in the magnitude comparison tasks. The overall game-based/paper-based test performance was used in some fraction knowledge analyses. It is an average of number line estimation accuracy and magnitude comparison performance.

Three constructs related to affective outcomes were computed, i.e. a flow construct (mean of nine flow dimensions), a playability construct (mean of two dimensions), and a test anxiety construct (mean of 12 items). Cronbach's alpha coefficient was used to evaluate the internal consistency of these affective measures.

The analysis of the collected data revealed that the data did not follow normal distribution. Thus, non-parametric tests were used in comparative analyses. Furthermore, correlation analyses were used to study the relation between different variables.

3 Results and Discussion

3.1 Onboarding phase

In order to study the adoption of the user interface we explored the students' performance in the onboarding phase (whole number estimation and whole number comparison tasks). A coin was shown in the first three whole number estimation tutorial tasks. It indicated the correct position of the whole number on the number line approximately. The students' accuracy was good in these tutorial tasks (tasks 1-3: M = 95.1%, SD = 10.57%). When the coin indicating the correct solution was no longer available, whole number estimation accuracy decreased (tasks 4-7: M = 93.03%, SD = 8.62%), but it was still good. There was one student who did not understand the tilting user interface during the onboarding levels and he or she did not move the game character at all which resulted in low accuracy. However, the results indicate that the students adopted tilting control during the onboarding phase pretty well.

The analysis of comparison performance revealed that some students faced difficulties in the whole number comparison tasks. Yet, the students' performance was satisfactory in six whole number comparison tasks (M = 92.48%, SD = 20.9%), as indicated by correctly solved comparisons. Accuracy increased during the onboarding phase while the mean accuracy of the last two comparison tasks was higher (M = 95.65%, SD = 20.6%) than the overall accuracy. Thus, we believe that a longer onboarding phase could have decreased mistakes caused by an unfamiliar user interface leading to smaller measurement errors.

The students' answers were analyzed more deeply in order to find the reasons for mistakes made in the whole number comparison tasks. The analysis indicated that 10 out of 51 students made comparison mistakes. Eight of these players made only one mistake. The following mistakes were identified: player did not move stones at all in the first task (3), player arranged stones in wrong order (3), and player answered without moving stones after a successfully performed comparison (2). These results indicate that most of the students could handle the stone carrying user interface after the onboarding phase. As a result, we argue that the user interface of the game is adequate for studying and assessing the students' conceptual fraction knowledge and the risk for a measurement error is quite low. It is noteworthy that accidental answers are always possible and hard to prevent without compromising the enjoyment aspect. In fact, the students tend to make ordering mistakes now and then even in paper-based context (ordering items to descending order instead of ascending order). However, in order to lower the measurement error risk a longer or an adaptive onboarding phase should be used in the future studies and in authentic assessment contexts.

3.2 Game-Based and Paper-Based Test Performance

First, we investigated, whether the paper-based test and the game-based test were comparable in terms of the construct validity of assessment. We expected (Hypothesis 1) that both tests produce similar results on the students' conceptual fraction knowledge. As expected, a correlation analysis indicated that there was a statistically significant relation between the overall paper-based test performance (M = 87,53%, SD = 10.73%, Mdn = 90.43%) and the overall game-based test performance (M = 82.28%, SD = 9.86%, Mdn = 84.35%), $r = .73$, $p < .001$, suggesting construct validity. However, a related-samples Wilcoxon signed-rank test showed that the students performed significantly better in the paper-based test than in the game-based test, $z = 3.949$, $p < .001$. Moreover, we considered construct validity in terms of individual task types. This analysis revealed that correlation between the estimation tasks was higher (paper: M = 92.12%, SD = 6.48%; game: M = 90.23%, SD = 6.95%; $r = .74$, $p < .001$) than between the comparison tasks (paper: M = 80.95%, SD = 16.76%; game; M = 74.33%, SD = 15.36%; $r = .63$, $p < .001$). Although the students performed better in the paper-based test, these findings support our Hypothesis 1. However, in order to decrease possible bias caused by accidental comparison anwers we will modify the interaction model of the answering button for future studies.

Second, we investigated the existence of effects and characteristics that are common in conceptual fractions knowledge. As expected in Hypothesis 2a, estimation accuracy was strongly related to the comparison performance (game: $r = .49$, $p < .001$; paper: $r = .63$, $p < .001$) indicating that both tasks reflect magnitude understanding. In line with our Hypothesis 2b, we observed a significant distance effect in the fraction comparison tasks. A related-samples Wilcoxon signed-rank test indicated that the students' game performance in the comparison pairs with large distance was significantly better (M = 79.5%, SD = 17.42%, Mdn = 80.95%) than in the comparison pairs with small distance (M = 68.74%, SD = 16%, Mdn = 71.43%), $z = 4.723$, $p < .001$. Moreover, the students compared the fraction pairs with large distance (M = 6622ms, SD = 2433ms, Mdn = 6001ms) significantly faster than the comparison pairs with small distance (M = 6977ms, SD = 2584ms, Mdn = 6145ms), $z = 2.381$, $p = .017$.

Unlike expected in Hypothesis 2c, a related-samples Wilcoxon signed-rank test indicated that the students did not perform significantly better in the whole number ordering consistent comparison pairs (M = 79.97%, SD = 12.85%, Mdn = 81.81%) than in the whole number inconsistent comparison pairs (M = 68.68%, SD = 29.55%, Mdn = 78.57%), $z = 1.566$, $p = .117$. This is in line with some of the recent studies which have reported inconsistent results regarding the whole number bias on fraction magnitude comparison [e.g. 25]). DeWolf and Vosniadou [25] explained the better performance in inconsistent tasks by assuming that some people may follow the reasoning that 'the smaller the whole number components the larger the value of the fraction'. Furthermore, consistent with our Hypothesis 2d the students' overall performance in both the paper-based and the game-based fraction tests were strongly related to their previous math grades (paper: r = .53, p < .001 and game: r =. 65, p < .001). To summarize, these findings indicate that the game-based test produced comparable cognitive results with the paper-based test and the Semideus Exam could be used for assessment purposes.

3.3 Flow Experience and Test Anxiety

One premise of game-based assessment relies on positive affective outcomes facilitating consequential validity. We studied the affective outcomes with respect to flow experience and test anxiety. The analysis indicated satisfactory internal consistency of the flow construct (game: α = .73; paper: α = .81) and of the test anxiety construct (game: α = .75; paper: α = .82). The more detailed analysis of the test anxiety dimensions showed that the internal consistency of thoughts (game: α = .76; paper: α = .80) and autonomic reactions (game: α = .86; paper: α = .85) dimensions were good, but the internal consistency of off-task behavior was poor (game: α = .49; paper: α = .38) in the present Finnish sample. Table 2 shows the means and standard deviations of the flow and test anxiety dimensions.

TABLE 2
DESCRIPTIVE STATISTICS OF FLOW AND TEST ANXIETY

	Game		Paper	
	M	SD	M	SD
Flow construct	4.46	.74	3.98	.86
Challenge-Skill Balance	3.69	1.52	4.00	1.23
Clear Goals	4.94	1.17	3.98	1.29
Unambiguous Feedback	4.51	1.21	3.78	1.32
Action-Awareness Merging	4.61	1.06	3.98	1.32
Transformation of time	3.96	1.65	3.49	1.63
Sense of control	4.63	1.13	3.88	1.21
Concentration	4.63	1.22	4.26	1.48
Loss of Self-Consciousness	5.22	1.22	4.86	1.50
Autotelic Experience	3.94	1.50	3.63	1.33
Anxiety construct	1.40	.33	1.59	.41
Thoughts	1.76	.64	2.06	.70
Autonomic Reactions	1.16	.35	1.25	.44
Off-Task Behavior	1.29	.44	1.45	.40

As expected in Hypothesis 3a, a related-samples Wilcoxon signed-rank test indicated that the students experienced significantly lower anxiety in the game-based test (M = 1.40, SD = .32, Mdn = 1.33) than in the paper-based test (M = 1.58, SD = .057, Mdn = 1.5), z = -2.848, p = .004 (see Table 2 for descriptives). Accordingly, in line with Hypothesis 3b, a related samples Wilcoxon signed-rank test confirmed that the students also experienced significantly higher flow experience in the game-based test (M = 4.46, SD = .74, Mdn = 4.56) than in the paper-based test (M = 3.98, SD = .86, Mdn = 4.22), z = 4.121, p < .001 (see Table 2 for descriptives). Moreover, the correlation analysis confirmed our Hypothesis 3c which assumed that there is a significant positive relation between the game-based test performance and flow experience, r = .52, p < .001. In addition, the correlation analysis confirmed our Hypothesis 3d, which assumed that there is a significant negative relation between game-based test performance and test anxiety, r = -.31, p < .026. These positive results on affective outcomes support the use of game-based assessment as game-based assessment might reduce anxiety-related bias in assessment. However, the fairness of the game-based test has to be considered in more detail within the subgroups.

3.4 Fairness of the Game-Based Test

In order to further ensure the validity and fairness of game-based assessment it is crucial to consider the subgroups of the students. We studied the fairness of the game-based test in relation to the students' playing frequency (gamers vs. non-gamers) and gender. 62% of the students reported that they play digital games almost daily, 30% once a week, and 8% rarely. Correlation analysis indicated that the playing frequency did not relate to success in the game-based test, r = .012, p = .935. We categorized the students who play games almost daily as gamers and the rest as non-gamers for further analyses. Independent samples Kruskal-Wallis test indicated that there were no statistically significant differences in the experienced flow level, $H(2)$ = 3.803, p = .149, experienced anxiety level, $H(2)$ = 2.836, p = .242 and perceived playability of the game, $H(2)$ = 2.618, p = .270, between gamers and non-gamers. The high playability score (M = 4.65, SD = .97; the internal consistency: α = .79) indicated that the tilting based user interface was accurate and intuitive to use, even though the game was played without sounds. In general, these findings mean that the gamers did not have an unfair advantage over non-gamers in succeeding in the game-based test.

Although the gamers did not have an advantage over the non-gamers with respect to accuracy, they adopted the user interface faster. Independent samples Kruskal-Wallis test indicated that the gamers estimated the whole number magnitudes (M = 14400ms; SD = 4409ms; Mdn = 13299ms) significantly faster than non-gamers (M = 16174ms; SD = 3868ms; Mdn = 15985ms), $H(2)$ = 6.536, p = .038. Accordingly, the gamers compared the whole number magnitudes (M = 6665ms; SD = 2657ms; Mdn = 6013ms) significantly faster than the non-gamers (M = 9789ms; SD = 4907ms; Mdn = 8445ms), $H(2)$ = 8.464, p = .015. In the actual test phase the gamers compared fractions significantly faster (M = 6007; SD = 2141; Mdn = 5744) than the non-gamers (M = 7703; SD = 2142; Mdn = 7905), $H(2)$ = 9.628, p = .008. However, there was no statistically significant difference in fraction estimation duration between the gamers (M = 9318ms; SD = 1987ms; Mdn = 9196ms) and the non-gamers (M = 9599ms; SD = 2385ms; Mdn = 10113ms), $H(2)$ = 1.951, p = .377. We assume that the estimation user interface was easier to adopt and the non-gamers learned to use it faster than the more complex comparison user interface which also required lifting and dropping the stones. These results indicate that the gamers may have an advantage over non-gamers if time is an important factor in the assessment.

A Mann-Whitney U test was run to determine whether there were differences in the experienced flow level, experienced anxiety level and perceived playability of the game between females and males. There were no statistically significant differences in the median scores of experienced flow ($U = 261$, $z = -1.157$, $p = .247$), experienced anxiety ($U = 336$, $z = .267$, $p = .790$), and perceived playability of the game ($U = 295$, $z = -.516$, $p = .606$) between females and males.

Furthermore, most of the students would prefer to participate in the game-based math test ($M = 4.67$, $SD = 1.47$). A Mann-Whitney U test was run to determine whether there were differences in the preferred test type between females and males. There were no statistically significant differences in the median test preference score between females and males, $U = 344$, $z = .433$, $p = .665$. In other words, 82.4% of the students would like to complete a game-based math test instead of a traditional paper-based math test. Test preference was not related to the math grade ($r = .02$, $p = .867$), playing frequency ($r = .21$, $p = .151$), game-based test performance ($r = -.11$, $p = .465$), or the paper-based test performance ($r = -.12$, $p = .410$). However, the students who appreciated the playability of the game also tended to prefer the game-based test, $r = .31$, $p =.029$. Moreover, consistent with the motivational model of video game engagement [8], playability (mastery of controls) was positively related to flow experience (enjoyment), $r = .49$, $p < .001$. As the model states, mastery of controls is not implicitly satisfying, but it unlocks the potential of the game to meet the player's psychological needs.

All in all, despite the fact that the gamers may have little advantage over the non-gamers we found some evidence that game-based assessment can be a fair form of assessment and most of the students may be willing to adopt game-based tests in school. However, we have to remember that the results can not be generalized to other games and more research is needed on different kinds of games.

3.5 Learning Analytics

A game-based assessment can also provide added value to the teacher. In the present study the teachers got an automatic report of the game-based test. The test report included descriptive statistics of their students' performance at individual and class level. The report distinguished estimation accuracy, overall comparison performance, performance in whole number ordering consistent and inconsistent comparisons as well as the overall test score (percentage). In addition to these basic metrics, the learning analytics module identified the students that had a clear whole number bias misconception and had problems in mastering the user interface leading to a biased test score.

In the present study, the whole number bias was detected if the student's performance was below 20% in the whole number ordering inconsistent comparisons and performance was over 85% in the whole number ordering consistent comparisons (due to the distance effect only the tasks in which the distance between fractions was over 0.3 were taken into consideration). According to this logic, seven participants had a clear whole number bias misconception. The automatic identification of misconceptions is very important, because the teachers tend to have problems to identify misconceptions based on traditional paper-based tasks validly. In general, the teachers of this study appreciated the test report a lot and they thought that it facilitated their work. Automatic report saved their time and helped them to focus on teaching the issues that were most challenging to their students.

4 Conclusions, Limitations, and Future Directions

In the present study we investigated the usefulness of game-based math assessment. To this end, we developed a Semideus Exam game prototype which was used to assess the students' conceptual fraction knowledge. We studied the cognitive outcomes of game-based assessment by comparing game-based test performance to paper-based test performance. Moreover, we investigated the fairness of the game-based assessment approach and the possible benefits of game-based assessment with respect to the affective outcomes.

The results of the study showed that the Semideus Exam game could be used to assess the students' conceptual fraction knowledge validly. However, we want to emphasize that the students should have enough time to familiarize themselves with the game controls and the game meachanics before valid assessment can be carried out with a game. In terms of affective outcomes, game-based assessment can lower test anxiety and increase engagement or at least flow experience. This is an important finding because test anxiety has a negative relation and flow experience a positive relation with test performance. Based on these findings we argue that game-based assessment could increase school satisfaction and it is also likely to decrease assessment bias caused by test anxiety. Moreover, the results revealed that most of the students were willing to do game-based math tests in school regardless of their playing experience (gamer vs. non-gamer) or gender. In fact, earlier playing frequency did not relate to success in the game-based test, thus suggesting the fairness of the game-based assessment approach. To summarize, the results provide some evidence that the game-based assessment approach might work well in math education with respect to both cognitive and affective outcomes.

This study has some limitations that call for more research on the topic. First, the sample size was quite small, thus reducing the significance of the study. Second, this study investigated game-based assessment with only one game and in one assessment context which also reduces the significance of the study. The benefits and risks of game-based assessment should be studied with a wide variety of games and in different assessment contexts.

Particularly, the importance of previous game playing experience needs to be studied more exhaustively with respect to different kinds of games. For example, the meaning of earlier playing experience or gender might be different in educational games which are based on more complex entertainment game mechanics. In this case, it was our aim to use simple game mechanics which everyone can easily adopt. Third, although the three teachers involved in the present study did appreciate the automatic report of the game-based test, the study did not focus on the usefulness of game-based assessment in terms of the teachers' and educators' demands and opinions. The educators' demands and opinions of game-based assessment are crucial for the diffusion of game-based assessment into daily practices. Therefore, we will focus on this topic in our future studies. To this end, we will develop a new Semideus version with more detailed learning analytics which also support a formative assessment approach.

To conclude, games can provide a different and detailed view of learning assessment. The most interesting possibilities that the games provide rely on assessing children's conceptual development in larger contexts. The big data sets that can be collected with online games like Semideus also make it possible to uncover dependencies and patterns behind conceptual change and to compare performance within and between countries. These comparisons could provide totally new insights for curriculum development and assessment. When the games can provide valid analysis about the learning process and learning assessment, we could provide something new and complementing to current large scale assessments such as PISA, TIMSS, and PIRLS. In fact, one of our future aims is to assess the development of conceptual rational number knowledge in a large scale with games.

ACKNOWLEDGMENT

The present study was funded by grant 289140 awarded to the first author by the Academy of Finland.

REFERENCES

[1] J.P. Gee, *What Video Games Have to Teach Us about Learning and Literacy*, New York, Macmillan, 2003.

[2] K.J. Devlin, *Mathematics Education for a New Era: Video Games as a Medium for Learning*, Florida, CRC Press, 2011.

[3] P. Wouters, C. van Nimwegen, H. van Oostendorp, & E.D van der Spek, "A meta-analysis of the cognitive and motivational effects of serious games", *Journal of Educational Psychology*, vol. 105, no. 2, pp. 249, 2013.

[4] J. ter Vrugte, T. de Jong, S. Vandercruysse, P. Wouters, H. van Oostendorp, & J. Elen, "Computer game-based mathematics education: Embedded faded worked examples facilitate knowledge acquisition", *Learning and Instruction* (Available online 23 November 2016), In Press.

[5] M. Bakker, M. van den Heuvel-Panhuizen, & A. Robitzsch, "Effects of playing mathematics computer games on primary school students' multiplicative reasoning ability", *Contemporary Educational Psychology*, vol. 40, pp. 55-71, 2015.

[6] K. Moeller, U. Fischer, H.C. Nuerk, & U. Cress, "Computers in mathematics education–Training the mental number line", *Computers in Human Behavior*, 48, pp. 597-607, 2015.

[7] M.M. Riconscente, "Results from a controlled study of the iPad fractions game Motion Math", *Games and Culture*, vol. 8, no. 4, pp. 186-214, 2013.

[8] A.K. Przybylski, C.S. Rigby, & R.M. Ryan, "A motivational model of video game engagement", *Review of general psychology*, vol. 14, no. 2, pp. 154-166, 2010.

[9] K.E. DiCerbo, R.J. Mislevy, and J.T. Behrens, Inference in game-based assessment," *Using Games and Simulations for Teaching and Assessment: Key Issues*, H.F. O'Neil, E.L. Baker, and R.S. Perez, eds., New York, NY: Routledge, Taylor & Francis Group, pp. 253-279, 2016.

[10] V.J. Shute, L. Wang, S. Greiff, W. Zhao, and G. Moore, "Measuring problem solving skills via stealth assessment in an engaging video game," *Computers in Human Behavior*, vol. 63, pp. 106-117, 2016.

[11] Á. Serrano-Laguna, J. Torrente, P. Moreno-Ger, and B. Fernández-Manjón, "Application of learning analytics in educational videogames," *Entertainment Computing*, vol. 5, no. 4, pp. 313-322, 2014.

[12] V.J. Shute, "Stealth assessment in computer-based games to support learning," *Computer games and instruction*, vol. 55, no. 2, pp. 503-524, 2011.

[13] Y.J. Kim and V.J. Shute, "The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment," *Computers & Education*, vol. 87, pp. 340-356, 2015

[14] J.W. Pellegrino, L.V. DiBello, and S.R. Goldman, "A framework for conceptualizing and evaluating the validity of instructionally relevant assessments," *Educational Psychologist*, vol. 51, no. 1, pp. 59-81, 2016.

[15] V.J. Shute, M. Ventura, and Y.J. Kim, "Assessment and learning of qualitative physics in newton's playground," *The Journal of Educational Research*, vol. 106, no. 6, pp. 423-430, 2013.

[16] C.D. Spielberger, and P.R. Vagg, "Test anxiety: A transactional process model," *Test anxiety: Theory, Assessment and Treatment*, C. D. Speilberger and P. R. Vagg, eds., Bristol: Taylor & Frances, pp. 3-14, 1995.

[17] R. Van Eck, "Digital game-based learning: It's not just the digital natives who are restless," *EDUCAUSE review*, vol. 41, no. 2, pp. 16-18, 2006.

[18] F. Bellotti, B. Kapralos, K. Lee, P. Moreno-Ger, and R. Berta, "Assessment in and of serious games: an overview. *Advances in Human-Computer Interaction*, *2013*, 1.

[19] V.J. Shute and Y.J. Kim, "Formative and stealth assessment," *Handbook of research on educational communications and technology"* New York: Springer, pp. 311-321, 2014.

[20] I. Mayer, D. van Dierendonck, T. van Ruijven, and I. Wenzler, "Stealth assessment of teams in a digital game environment," *Proc. International Conference on Games and Learning Alliance*, pp. 224-235, 2013.

[21] A. Ischebeck, M. Schocke, and M. Delazer, "The processing and representation of fractions within the brain: An fMRI investigation," *NeuroImage*, vol. 47, no. 1, pp. 403-413, 2009.

[22] R.S. Siegler, L.K. Fazio, D.H. Bailey, and X. Zhou, "Fractions: the new frontier for theories of numerical development," *Trends in cognitive sciences*, vol. 17, no. 1, pp. 13-19, 2013.

[23] R.S. Siegler, C.A. Thompson, and M. Schneider, M. "An integrated theory of whole number and fractions develop-

ment," *Cognitive psychology*, vol. 62, no. 4, pp. 273-296, 2011.

[24] S. Stafylidou, and S. Vosniadou, "The development of students' understanding of the numerical value of fractions," *Learning and instruction*, vol. 14, no. 5, pp. 503-518, 2004.

[25] M. DeWolf and S. Vosniadou, "The representation of fraction magnitudes and the whole number bias reconsidered," *Learning and Instruction*, vol. *37*, pp. 39-49, 2015.

[26] Y. Ni and Y.D. Zhou, "Teaching and learning fraction and rational numbers: The origins and implications of whole number bias," *Educational Psychologist*, vol. 40, no. 1, pp. 27-52, 2005.

[27] M.W. Alibali, and P.G. Sidney, "Variability in the natural number bias: Who, when, how, and why," *Learning and Instruction,* vol. 37, pp. 56-61, 2015.

[28] J. Van Hoof, T. Lijnen, L. Verschaffel, and W. Van Dooren, "Are secondary school students still hampered by the natural number bias? A reaction time study on fraction comparison tasks," *Research in Mathematics Education*, vol. 15, no. 2, pp. 154-164, 2013.

[29] M. Schneider, and R.S. Siegler, "Representations of the magnitudes of fractions," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 36, no. 5, pp. 1227-1238, 2010.

[30] R.S. Moyer, and T.K. Landauer, "Time required for judgements of numerical inequality," *Nature*, vol. 215, pp. 1519–1520, 1967.

[31] J. Torbeyns, M. Schneider, Z. Xin, and R.S. Siegler, "Bridging the gap: Fraction understanding is central to mathematics achievement in students from three different continents," *Learning and Instruction*, vol. 37, pp. 5-13, 2015.

[32] D.H. Bailey, R.S. Siegler, and D.C. Geary, "Early predictors of middle school fraction knowledge," *Developmental Science*, vol. 17, no. 5, pp. 775-785, 2014.

[33] L.S. Fuchs, R.F. Schumacher, J. Long, J. Namkung, C.L. Hamlett., P.T. Cirino, N.C. Jordan, R. Siegler, R. Gersten, and P. Changas, "Improving at-risk learners' under- standing of fractions," *Journal of Educational Psychology,* vol. 105, pp. 683–700, 2013.

[34] G. Wood, and M.H. Fischer, (2008). "Numbers, space, and action–from finger counting to the mental number line and beyond," *Cortex*, vol. *44*, no. 4, pp. 353-358, 2008.

[35] J. Van Hoof, J. Vandewalle, L. Verschaffel, and W. Van Dooren, "In search for the natural number bias in secondary school students' interpretation of the effect of arithmetical operations," *Learning and Instruction*, vol. 37, pp. 30-38, 2015.

[36] E.N. Castellar, J. Van Looy, A. Szmalec, and L. De Marez, "Improving arithmetic skills through gameplay: assessment of the effectiveness of an educational game in terms of cognitive and affective learning outcomes," *Information Sciences*, vol. *264*, pp. 19-31, 2014.

[37] F. Ke, "A case study of computer gaming for math: Engaged learning from gameplay?," *Computers & Education*, vol. 51, no. 4, pp. 1609-1620, 2008.

[38] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*, New York, Harper Perennial, 1991.

[39] K. Kiili, T. Lainema, S.D. Freitas, and S. Arnab, "Flow framework for analyzing the quality of educational games," *Entertainment Computing*, vol. 5, no. 4, pp. 367-377, 2014.

[40] S. Engeser, and F. Rheinberg, "Flow, performance and moderators of challenge-skill balance," *Motivation and Emotion*, vol. 32, no. 3, pp. 158-172, 2008.

[41] S. Engeser, eds., *Advances in flow research,* New York, NY: Springer, 2012.

[42] K. Kiili, K. Devlin, A. Perttula, P. Tuomi, and A. Lindstedt, "Using video games to combine learning and assessment in mathematics education," *International Journal of Serious Games*, vol. 2, no. 4., 2015.

[43] D.W. Putwain, and R.A. Daniels, "Is the relationship between competence beliefs and test anxiety influenced by goal orientation?," *Learning and Individual Differences*, vol. 20, no. 1, pp. 8-13, 2010.

First K. Kiili, Ph.D. (Doctor of Philosophy, Multimedia, 2006), M.Ed. (2001). Kiili has worked in Tampere University of Technology since 2001. Currently he is Academy Research Fellow in Tampere University of Technology. Has been a visiting scholar in Stanford University in years 2009, 2013, and 2014-2015. Editorial board member of International Journal of Serious Games and European Conference on Game-Based Learning. The nomination as a researcher of the year of University Consortium of Pori, 2011. Outstanding paper award in ED-MEDIA conference 2006. He has got several research grants and has been principal investigator in several game-based learning projects. Kiili has published over 80 international peer-reviewed papers. His current research interests are related to learning sciences, game-based learning and assessment, and mathematics learning games.

Second H. Ketamo, Ph.D. (information technology, 2002), is an independent researcher and start-up entrepreneur. Previously he has been e.g. Academy of Finland -granted post-doc researcher, founder of gameMiner (game AI/data mining), director of education at Satakunta University of Applied Sciences and founder of SkillPixels ltd. (game-based learning). Currently he is founder of HeadAI ltd. (AI/robotics) and senior fellow at University of Turku. Ketamo has published 86 international peer-reviewed papers and had more than 200 presentations in international forums. His current research interests are related to learning sciences, artificial intelligence and semantic computing.