Computers in Human Behavior 65 (2016) 391–401

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

The higher the score, the higher the learning outcome? Heterogeneous impacts of leaderboards and choice within educational videogames

Steve Nebel*, Maik Beege, Sascha Schneider, Günter Daniel Rey

E-Learning and New Media, Faculty of Humanities, Technische Universität Chemnitz, Germany

ARTICLE INFO

Article history:
Received 29 April 2016
Received in revised form
1 June 2016
Accepted 27 August 2016
Available online 9 September 2016

Keywords:
Competition
Choice
Educational videogames
Feedback intervention
Leaderboards

ABSTRACT

Interactive multimedia environments such as educational videogames offer great potential for learning in groups with multiple players. Multiplayer games might lead to competition among the learners which is frequently used to motivate them to play again. Additionally, competitive outcomes as discrepancies to a desired standard might differ between players and this type of feedback might influence learning. Therefore, the experiment seeks to investigate learning effects of different amounts of standard discrepancy and the choice to repeat levels. Standard discrepancy was operationalized by either showing a high learner score (low standard discrepancy) or a medium learner score (large standard discrepancy) at a leaderboard. Choice to repeat a level (possibility to repeat vs. no possibility to repeat) was manipulated by presenting or hiding a repeat button. An experiment was conducted with 85 students who played a jump-and-run game in order to learn facts about three allegorical paintings. Results revealed an effect of standard discrepancy on retention performance with higher scores for the high standard discrepancy condition. Choice did not influence learning outcomes, but improved motivational and emotional measures. Findings underpin the new role of leaderboards as feedback mechanisms.

Leaderboards as displays of ranks for comparison (Seaborn & Fels, 2015) are popular among different applications in digital media. They are frequently used within internet-based games as they can be persistent, accumulated long-term and represent skill levels better than single play sessions (Wang & Sun, 2011). Since the early days of arcade cabinets, leaderboards are used with-incommercial games to increase replayability. For example, ARPG's (Action Role Playing Games) like Diablo III (2012) or Path of Exile (2013) utilize season based leaderboards. As a consequence of their popularity, leaderboards are interesting for gamification. This concept is defined as “the intentional use of game elements for a gameful experience of non-game tasks and contexts” (Seaborn & Fels, 2015, p. 17). For this, leaderboards are often included in quizzes like a usability-quiz (Hemke, Meyer, Hühne, Schneider, & Wohlgemuth, 2014) or a quiz for hearing-impaired learners (Glova, Asuncion, Martin, Manzan, & Pagtaconan, 2014). They can even be found in applications for project management (e.g., Kudos Badges, 2015) or within tools to foster user participation during software development (Halan, Rossen, Cendan, & Lok, 2010). Considering these examples, it is not surprising that leaderboards appear advantageous to creators of educational learning materials. However, sufficient research analyzing individual learner outcomes when interacting with leaderboards is missing. Even if leaderboards are included in empirical comparisons they are often part of a larger gamification strategy (Bajko, Hodson, Seaborn, Livingstone, & Fels, 2015) or included in combination with achievements and other competitive mechanics (Landers & Landers, 2015) which limits the interpretation tremendously. Empirical studies addressing gamification elements often lack of sufficient statistical analysis to generate comparable effects (Seaborn & Fels, 2015). Even if valuable data is collected, the researchers regularly target other factors than learning (e.g. fun, Butler, 2013). Therefore, creators of educational content still need sufficient empirical studies to base their decision on whether to include leaderboards or not. Additionally, studies do not provide sufficient information on how leaderboards influence the individual learner. With this experiment, we seek to provide further empirical evidence to guide the

* Corresponding author. E-Learning and New Media, Faculty of Humanities, Technische Universität Chemnitz, Straße der Nationen 12, 09111 Chemnitz, Germany.

E-mail address: steve.nebel@phil.tu-chemnitz.de (S. Nebel).

http://dx.doi.org/10.1016/j.chb.2016.08.042
0747-5632/

implementation of leaderboards and to close the scientific gap of individual impacts.

1. Literature review

1.1. Leaderboards in (educational) videogames

Leaderboards can contain and categorize various elements, such as the number of correct answers, achieved goals or time spent. They can also be used to evaluate and categorize solutions for given tasks (e.g., Foldit, Beta). The required data is usually easy to acquire as it is already part of the game. The resulting list is comparably easy to integrate into an educational concept as no interference with gameplay occurs. These technical advantages might partially explain the popularity of leaderboards. However, leaderboards influence several other aspects of educational gaming. For example, the increased difficulty by leaderboards also increases perceived value of the achieved results, thus increasing memorability (Wang & Sun, 2011). When presented at the end of the game, leaderboards are typically used to increase the motivation to play again and subsequently increase time-on-task (Landers & Landers, 2015). Leaderboards provide entertainment, a sense of accomplishment and memories linking play events to specific rewards (Wang & Sun, 2011). Leaderboards serve as source of motivation (Schubert, Paulsen, & Hager, 2014; Willems et al., 2014) as learners see that their work is recognized (Domínguez et al., 2013). Beyond these motivational impacts, leaderboards can influence goals and the perception of progress within educational videogames by mapping progress and inciting actions (Seaborn, Pennefather, & Fels, 2013). For example, if students lack of engagement in the early stages of a class or game they fall back in a leaderboard and start taking actions (Barata, Gama, Jorge, & Gonçalves, 2013). Therefore, leaderboards provide stimuli or goals (e.g., Glova et al., 2014; Hemke et al., 2014) which might help to complete learning tasks (Domínguez et al., 2013). Leaderboards might also be useful when players are not able to beat the complete game. Players can try to beat their own records (Wang & Sun, 2011) and, subsequently, perceive achievement even during the early stages of the learning process. In sum, leaderboards can provide long-term goals (e.g., be the best) and short-term goals (e.g., improve a self-determined number of places), which both support motivation and orientation during playtime. In addition, leaderboards, as the easiest form of social interaction (Domínguez et al., 2013; Wang & Sun, 2011), are perceived as more influencing than achievements or similar progression indicators. Wang and Sun (2011) claim that the reward of a satisfying position within the leaderboard might encourage social interaction as exchanging information might improve performance and, subsequently, closeness to a desired status. Leaderboards increase collaboration (Schubert et al., 2014) and competition (Sarangi & Shah, 2015) as well. This might be amplified even further, as providing information serves as a possibility to show off learned skills. Regarding these factors, it should be noted that the acceptance of (public) competitive factors might be culturally dependent (Barata et al., 2013; Cheong, Filippou, & Cheong, 2014; Schubert et al., 2014) and the impacts of leaderboards might vary within different cultural backgrounds. Additionally, the limited information value of some leaderboards might lead to learners feeling uncomfortable with their position (Domínguez et al., 2013). For example, learners do not know if others actually learned more or if they just played better. In this vein, it is shown that competition within leaderboards could lead to demotivation if the distance to other students gets too high. Therefore, leaderboards require enough users in order to ensure the existence of comparable competitors (Willems et al., 2014).

1.2. Heterogonous effects of competition

It is hypothesized that leaderboards affect motivation, post-test performance and behaviors through competition (Cagiltay, Ozcelik, & Ozcelik, 2015; Simões, Redondo, & Vilas, 2013). Competitive gameplay might increase interest (Plass et al., 2013), enjoyment (Vorderer, Hartmann, & Klimmt, 2003), or attention, excitement and involvement (Vandercruysse, Vandewaetere, Cornillie, & Clarebout, 2013). However, recent research on the effect of competition within various group constellations has shown heterogeneous effects (Nebel, Schneider & Rey, 2016). Among other things, it can be derived from this study that the impact of competition might vary from player to player relative to the personal gameplay experience and personal traits. Although competition as a form of social comparison should lead to an unidirectional drive upward (Festinger, 1954), this is not distributed homogenously among the players, and subsequently, within the leaderboard. Some learners might perceive stronger competition as others relative to their proximity to standards (Garcia & Tor, 2007, 2009; Garcia, Tor, & Gonzalez, 2006). Therefore, the effects of competition might vary among the players as well. For example, players close to a standard are more likely to replay parts of the game (Butler, 2013). In contrast, learners that are allowed to control the task themselves tend to practice a different task after good trials (Wu & Magill, 2011). Therefore, players in good positions might not choose to play again but rather play another level to learn something different. In addition, personal traits influence how enjoyable leaderboards actually are. For example, introverts might appreciate an offline-leaderboard more than extroverts (Codish & Ravid, 2014). Nonetheless, generally leaderboards are rather motivating than demotivating (Schubert et al., 2014).

1.3. Leaderboards as feedback mechanism

Leaderboards do not only induce competitive effects, they also provide information on how a player performed. The simplest element is the rank itself (e.g., 34rd). It can be regarded as a *praise for task performance*, although this kind of feedback contains very limited learning related information (Hattie & Timperley, 2007). Achievements can be classified as glory mechanisms (Wang & Sun, 2011) as well, although it is important to differ between achievements and leaderboards. The first are permanently granted and indicate an already reached standard and signal no need for further behavior. The latter also indicates a certain skill level but, additionally, a difference towards a potentially more desirable standard. However, the information leaderboards provide does not represent continuous feedback as it appears only punctual. Therefore, leaderboards can be compared to feedback interventions (FI), which are defined as "actions taken by external agents to provide information regarding some aspects of one's performance" (Kluger & DeNisi, 1996, p. 255), whereas the external agent is the game mechanic that forces the player to recognize the leaderboard. These FI impact pleasantness and arousal. Depending on its type (positive or negative) FI can result in positive or negative mood (Kluger & DeNisi, 1996). In line with some aspects of competition, the *Feedback Intervention Theory* (FIT) states that behavior is regulated by a comparison of feedback to standards (Kluger & DeNisi, 1996). Learners with a negative feedback (feedback that the performance is below a desired standard) either increase their effort, abandon or modify the standard, or reject the feedback message (Kluger & DeNisi, 1996). In line with the goal setting theory (Locke & Latham, 1990, 2002), Feedback Intervention Theory argues that arousal is elevated because the feedback-standard gap increases (Kluger & DeNisi, 1996). As a consequence, learners who perform already within the boundaries of a standard might not increase their effort. This might be in contrast to the postulated effects of competition which indicate stronger effects as players perform closer to certain standards.

Several other factors might lead to heterogeneous impacts of feedback provided by leaderboards. For example, “commitment to goals” is a mediator of the effectiveness of positive and negative feedback (Hattie & Timperley, 2007). Upon receiving negative feedback, players might become unsatisfied with their performance, set higher performance goals and subsequently perform better. In contrast, positive feedback could foster interest and increase the chance to persist in the activity. Additionally, there might be an interaction with self-efficiency as highly self-efficacious learners might seek negative feedback at first, to excel at the task (Hattie & Timperley, 2007). Finally, Hattie and Timperley (2007) state that feedback works on four levels. Feedback might be focused on the task, the process, self-regulation or the self as a person. Leaderboards as a feedback mechanism work on the task-level (*How good am I doing, compared to my final goal?*) and on the self-level (*Am I performing well?*). The task level is explicit and direct, while the self-level depends on individual competitive factors (*Is 34rd a good performance for me?*) and may vary from player to player. Therefore, heterogeneity within the impacts of leaderboards on learners might be further amplified.

1.4. The present experiment

As the literature review highlighted, closeness to standards within a leaderboard might induce different effects on learners that have not been sufficiently analyzed. The underling theories cannot describe learner behavior within educational videogames completely and need further diversification and enhancement supported by empirical data. Therefore, we seek to systematically manipulate closeness to a desired standard and measure its impact on cognitive, motivational and learning outcomes. A previous study on the impacts of user position within a leaderboard analyzed high vs. low position within the leaderboard (player scores better/worse than any of the other players; Butler, 2013). However, both variations represent proximity to a standard (winning/loosing). In contrast, *standard discrepancy* is manipulated by placing the player close to a standard (rather high score, thus close to winning) and far away from any standard (middle position, neither winning nor loosing). As leaderboards are used to provoke repeated learning cycles and self-controlled practice is suspected to increase learning outcomes in general (Wu & Magill, 2011), we integrated an *opportunity to repeat* levels as the second factor *choice*. With this, we seek to further enrich the empirical basis for theories on leaderboards.

To investigate the impact of the group manipulation on competition and effort, three hypothesis are postulated. As the *n-effect* predicts higher competitive effects within closeness to specific standards (Garcia & Tor, 2007, 2009; Garcia et al., 2006), the first hypothesis is formulated:

Hypothesis 1. *Learners within the low standard discrepancy groups (LSD) perceive higher competition than learners within the high standard discrepancy groups (HSD).*

Additionally, FIT predicts lowered effort for players who perform close to their desired standard (Kluger & DeNisi, 1996). For this, the second hypothesis is postulated:

Hypothesis 2a. *Learners within the low standard discrepancy groups invest lower competitive effort than learners within the high standard discrepancy groups.*

As it is expected that the possibility to repeat parts of the game gives the player the impression that they can undo their mistakes and improve anytime, subsequently requiring less competitive effort while playing on their first run, another hypothesis is formulated:

Hypothesis 2b. *Learners with possibilities to repeat the levels (PR) invest less competitive effort than learners without this possibility (WR).*

As previous studies highlight the importance of motivation in the field of gamification (Seaborn & Fels, 2015) and relative autonomous motivation is related to academic performance (Kusurkar, Ten Cate, Vos, Westers, & Croiset, 2013), we were interested in the impacts of the manipulation. It is expected that increased competitiveness through proximity to a standard increases motivation. Furthermore, it is hypothesized that the self-determined approach of the replay mechanism might influence relative autonomous motivation positively. Additionally, the leaderboard as praise for task performance and the replay mechanic as a method to undo mistakes and errors should influence emotions positively. More specifically, it is assumed that positive emotions should increase as the proximity to the desired standard decreases. In contrast, negative emotions should decrease as players could repeat elements of the game. As a consequence, the following hypothesis is postulated:

Hypothesis 3a. *Learners within the low standard discrepancy groups have higher relative autonomous motivation and perceive more positive emotions than learners within the high standard discrepancy groups.*

Hypothesis 3b. *Learners with possibilities to repeat the levels have higher relative autonomous motivation and perceive less negative emotions than learners without this possibility.*

Both the leaderboard as feedback mechanism (Kluger & DeNisi, 1996) and the leaderboard as competitive factor should influence learning (Nebel et al., 2016). As we cannot safely hypothesize if the FI or the effects of competition affect the learning outcome more, the direction of the learning assumptions cannot be determined. Therefore, an undirected hypothesis is postulated:

Hypothesis 4a. *Learners within the low standard discrepancy groups differ in their learning outcomes compared to learners within the high standard discrepancy groups.*

Additionally, as self-controlled practice might increase learning outcomes (Wu & Magill, 2011), the following hypothesis is postulated:

Hypothesis 4b. *Learners with possibilities to repeat the levels learn more than learners without this possibility.*

As goal orientations have numerous impacts on academic performance and affective states (Pekrun, Elliot, & Maier, 2009), the experiment is expanded with an analysis of goal orientations. Because of the exploratory approach, no specific hypothesis are presented.

2. Method

2.1. Materials and design

As the feedback mechanism with leaderboards is rather simple and contains relatively little information, the gameplay mechanic should be simple as well. Therefore, a simple jump-and-run mechanic as the basis for the educational videogame was implemented. The game developing environment Unreal Engine 4 (2015) and their *Paper 2D* sprite-based plug-in was used to create three levels where the player could jump across different paintings and learn about their meanings and creation. To ensure low prior experience regarding the learning content, the rather unfamiliar learning topic of allegories and comparably unknown artists were used. More specifically, the paintings *Allegory on the Tulip Mania* from Jan Brueghel the Younger (1640), *Allegory of Four Seasons* from Johann Georg Platzer (1750) and *Allegory of the Consequences of the Peace of Utrecht* from Paolo De Matteis (1714) were included.

Upon designing the leaderboard, it had to be decided which information the leaderboard should contain, when to include the leaderboard into gameplay and how to manipulate the closeness-factor. Leaderboards that could be activated every time the player desired, could be categorized as permanent feedback-seeking behavior. As the theoretical frame for this experiment was based on interventions, we decided to include the leaderboard only after each level. Thus, its characteristic as punctual feedback intervention was highlighted. Furthermore, as velocity is regarded as an FI that directs attention to the motivational level which improves performance (Kluger & DeNisi, 1996), this value was utilized as the organizational core-element for the leaderboard. To further foster the perception of an actual achievement while performing close to the desired standard of winning, different colored trophies (e.g., gold for the first place, silver for the second) were included to represent different rankings (Domínguez et al., 2013) and a short statement was added indicating comparable competitors (Fig. 1). To manipulate the position within the leaderboard, it had to be filled with realistic values depending on the performance of the individual test subject. In order to achieve this, the place within the leaderboard was calculated first. Depending on the experimental condition, a random place within the 2.5%–5% interval (e.g., 4th of 87 players) for the close to standard condition, or within the 45%–55% interval (e.g., 43rd of 87 players) for the high standard discrepancy group was calculated. Therefore, players were not placed on the same rank every level (which might have felt unnatural) but within a predetermined range. To fill the remaining places of the leaderboard with useful and realistic data, the fastest possible time anyone could beat the levels had to be determined. Afterwards, a multiplier for values lower (better ranks) and higher (lower ranks) than the player, based on the fastest playtime and the real player performance was calculated.

$$M_{low} = \frac{Player\ Time - Fastest\ Time}{3} \quad M_{high} = Player\ Time - Fastest\ Time$$

Additionally, standard normal distributed random numbers for each missing leaderboard rank was generated with the polar method (Glasserman, 2003; Marsaglia & Bray, 1964). Finally, a linear combination of the absolute random numbers and the corresponding multiplier had to be added (or subtracted, depending on the rank) to the player time. After sorting the resulting dataset and recalculating values that laid below the fastest possible time (which were very rare as we used the divisor three on this side of the leaderboard), we received leaderboards with a realistic distribution, but random content based on the individual player performance. To avoid a possible confounding with players' performance, we only showed the closest scores. As a consequence, players within the high standard discrepancy groups did not know what performance is needed to reach the first place, and thus, had to orient themselves on the presented mediocre ranks. Additionally, this leaderboard screen was used to present buttons to advance to the next level. Furthermore, a button to repeat the level and the necessary explanation for the PR group was included here. Examples of the final screens are shown in Fig. 1. Finally, the materials have been pre-tested several times, to ensure the game and the leaderboard work as intended.

2.2. Tasks

The first task for the participants was to master a basic and sequential gameplay tutorial (Fig. 2). This level taught the players how to navigate and how to interact with their surroundings. Additionally, they were told about the goal of the jump-and-run (reaching the best possible rank within the leaderboard) and how to finish the game. After making sure that no more questions were unanswered, the players faced the experimental tasks.

The participants could explore one painting per level and were told to find five green *information bubbles* (Fig. 3). These bubbles disappeared upon contact with the player figure and a text block containing learning information was presented. After seven seconds (a timespan evaluated during the pre-tests), the text block

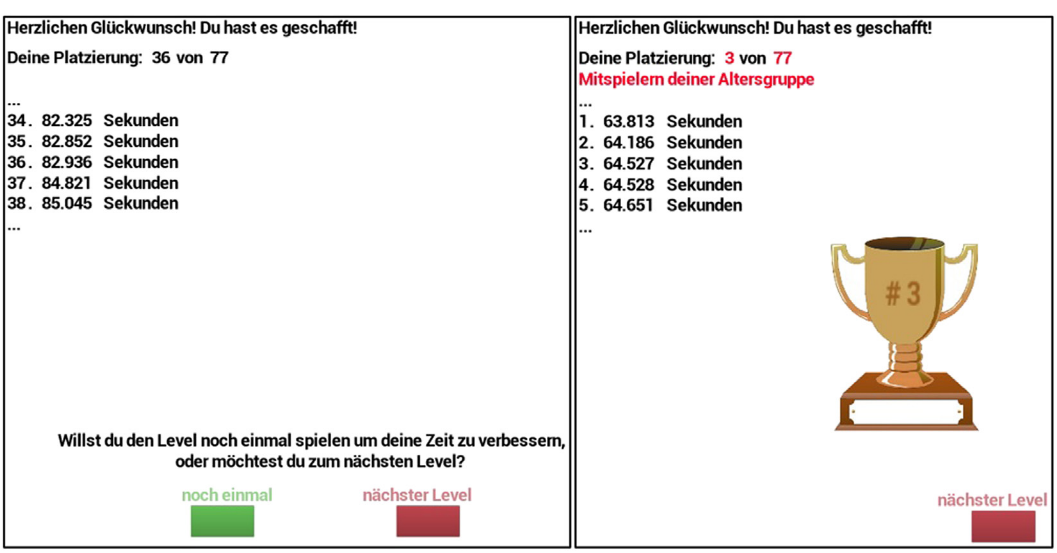

Fig. 1. *Leaderboard Screens*: HSD*PR condition on the left, translation: "Congratulations, you made it! Your Place: 36 of 77 Do you want to replay the level to improve your time, or do you want do proceed to the next level?", LSD*WR on the right, translation: "Congratulations, you made it! Your Place: 3 of 77 for players within your age range".

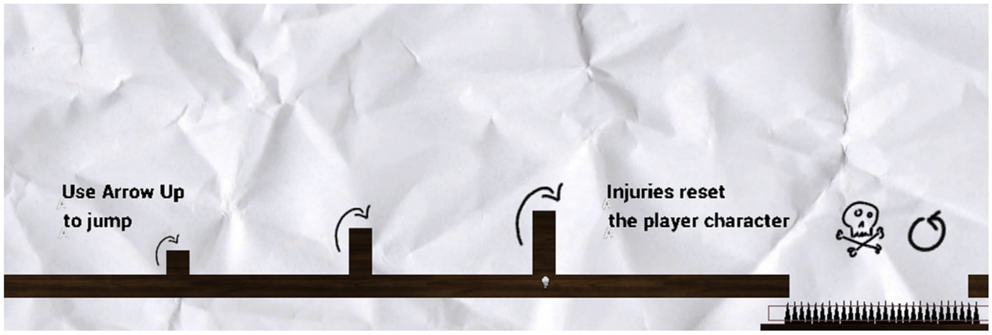

Fig. 2. *Segment of the tutorial (translated version)*: to provide a consistent art-theme, a background with a paper texture and instructions looking like sketches was used.

Fig. 3. *Level 1*: a small segment of level one and the *Allegory on the Tulip Mania*, the counter on the top indicates one of five questions have been answered correctly, the blank space indicates a missing question in this area, on the left, a green information bubble is visible. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

disappeared and the bubble was visible again. The players could collect the bubbles as often as they wished, in case they had not read and learned everything within the information blocks yet. After being collected for the first time, each information bubble unlocked a specific red *question bubble* somewhere else on the painting. Upon contact with the player character, this bubble disappeared and a question concerning the content of the connected information bubble was presented. If the player chose the right of six presented answers, both connected bubbles disappeared permanently and a counter indicated that the number of correct answers had increased. Additionally, a segment of the painting that had been cut out, was restored. This simple restauration of the paintings served as a backstory and basic motivation for the players. If the player, however, gave the wrong answer, the question bubble was deactivated and the player had to collect the information bubble again. Thus, we wanted to reduce the potential of arbitrary guessing and forced the player to review the related information. If the player successfully answered all five questions, he could advance to the next level.

2.3. Participants

Overall, 85 students (78.8% female; age: $M = 22.22$, $SD = 3.73$) from the TU Chemnitz participated in this experiment. 23 participants were assigned to the HSD + WR condition, 19 participants to the HSD + PR condition, 21 students to the LSD + WR condition and 22 students to the LSD + PR condition. Students are enrolled in media and communication studies (71.8%), instructional and media psychology (24.7%) and informatics and communication studies (3.5%). Each participant received a 1.5 h course credit. Overall, students had medium prior knowledge: $M = 2.88$, $SD = 0.92$ on a scale ranging from one to five. Prior knowledge is included as a covariate in the statistical analyses as this might influence the effect of competition (Ter Vrugte et al., 2015).

2.4. Measures

To capture the various impacts of the experimental manipulation and to gain insight into the relationships with goal orientation, several measures were used. First of all, prior experience was measured on a self-created two-item scale (*"How would you describe your prior knowledge in terms of video games, e.g. PC, console, mobile phone and devices?"* and *"How would you describe your prior knowledge in terms of "Jump and Runs", e.g. Super Mario."*) which reached in a five point Likert scale from *"very bad"* to *"very good"* and showed a sufficient reliability of $\alpha = 0.74$. Afterwards, goal orientation was addressed. More specifically, the *3 × 2 Achievement Goal Questionnaire* (Elliot, Murayama, & Pekrun, 2011) was used. Besides *approach* and *avoidance* orientations, the questions capture the *task*, *one-self* and *others* goal orientations. Thus, they could provide more information than the common *task/mastery* approach (Elliot & Murayama, 2008; Elliot, 1999). This was especially useful because we were interested in the effects of social interactions. The questionnaire contained six subscales, one for each goal construct. Furthermore, each segment contained three statements, presenting a possible goal for the player. This resulted in 18 sentences altogether. Statements like *"To outperform other students on the exams in this class"* had to be rated on seven point Likert scales, ranging from *"not true of me"* to *"extremely true of me"*. To ensure consistency with the experiment, wording had to be slightly modified. For example, *"this class"* was replaced with *"this experiment"*. The scales show sufficient reliability ranging from $\alpha = 0.80$ to 0.92, similar to other implementations (Elliot et al., 2011). To address mental load and mental effort, a questionnaire provided by Krell (2015) was used. This questionnaire contained 12 questions, six for each component. Statements like *"The tasks were challenging"* had to be rated on seven point Likert scales ranging from *"not at all"* to *"totally"*. The reliability turned out to be sufficient for mental effort ($\alpha = 0.84$) and mental load ($\alpha = 0.84$). This was followed by two extended items from the *Game Experience Questionnaire* (IJsselsteijn, in preparation) addressing perceived competition ("*I felt challenged – to be better than others*") and competitive effort ("*I had to put a lot of effort into it – to be better than others*"). Participants had to rate these statements on five point Likert scales ranging from *"completely true"* to *"not true at all"*. Afterwards, the *Achievement Emotions Questionnaire* (AEQ) containing various emotional constructs (Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011; Pekrun, Götz, & Perry, 2005) was used. The questionnaire contained items that needs to be answered before, during and after the sessions. As the experiment was completely new and participants did not know what to expect, questions that needed to be presented before the learning phase could not be used. Additionally, we did not want to disturb the gaming phase, therefore only questions that could be answered after the learning phase were included. Furthermore, the AEQ differs between learning-related, class-related and test-related emotions. As we were interested in how the educational videogame and the manipulations impacts emotions in the context of the overall classroom experience, we focused on the class-related scales. Thus, 15 questions remained, addressing enjoyment, pride, anger, shame and hopelessness. Statements like *"I am happy that I understood the material"* had to be rated on five point Likert scales ranging from *"strongly disagree"* to *"strongly agree"*. Apart from hopelessness, the scales showed reliabilities ranging from $\alpha = 0.74$ to $\alpha = 0.87$, similar to the literature (Pekrun et al., 2011). However, hopelessness should be interpreted with caution because of the low reliability ($\alpha = 0.61$). This was followed by the *Situational Motivation Scale* (SIMS) addressing motivation within four subscales (Guay, Vallerand, & Blanchard, 2000). This questionnaire included 16 items, equally distributed among the subscales of intrinsic motivation, identified regulation, external regulation and amotivation. The participants had to rate why they were currently engaged in this activity with statements like *"Because I think that this activity is pleasant"* on seven point Likert scales ranging from *"corresponds not at all"* to *"corresponds exactly"*. The resulting scales provided sufficient reliability ranging from $\alpha = 0.77$ to $\alpha = 0.89$, consistent with the literature (Guay et al., 2000). Only the identified regulation subscale showed borderline reliability ($\alpha = 0.65$) and should be interpreted with caution. Finally, learning outcomes were addressed. Fifteen items were implemented, asking the same questions the players had to answer in order to beat the game successfully. The participants had to choose from four possible answers, whereas only one solution was correct. Thus, a *retention* score was aggregated.

In addition to the measurements, two additional values were calculated. First, *relative autonomous motivation*, as this showed to be a valuable indicator in previous studies (Kusurkar et al., 2013; Vansteenkiste, Lens, De Witte, De Witte, & Deci, 2004; Vansteenkiste, Zhou, Lens, & Soenens, 2005), representing a weighted score for the relative overall motivation. In line with previous approaches, we assigned weights to intrinsic motivation (+2), identified regulation (+1), external regulation (−1) and amotivation (−2) taken from the SIMS. After summing up these four scales, we received the resulting relative autonomous motivation score. Secondly, *instructional efficiency* was calculated, using the formula of Van Gog and Paas (2008):

$$Efficiency = \frac{zP_{test} - zE_{test}}{\sqrt{2}}$$

Retention scores served as test performance (P_{test}). In line with previous approaches (Nebel et al., 2016), test effort was subsided with learning related measures to address *procedural efficiency*. More specifically, mental effort was used to calculate *efficiency effort* and mental load was used to determine *efficiency load*.

2.5. Procedure

Prior to the experimental sessions, a computer laboratory was prepared. For this, the electronical questionnaires created with *Lime Survey* (2015) were loaded and the educational videogame was installed on a Windows 7 PC with a 24″ monitor. The participants were tested individually in order to ensure optimal experimental manipulation. For example, other participants in the same room could induce effects of social comparison that could have interfered with the manipulation. The students were randomly assigned to one of the four experimental conditions by drawing lots. At first, the participants had to answer the demographic questionnaire, the goal orientation questionnaire and the prior experience test on screen. Afterwards, they had to play the game, starting with a tutorial which was implemented to teach basic navigation functions. The three test levels lasted about 22 min ($SD = 7.6$). After the game was finished, a second questionnaire had to be fulfilled on screen, containing competition, mental load, mental effort, motivation and emotion scales followed by the retention knowledge test. If all tests were completed, the participants were thanked and left the room. Altogether, the experiment lasted between 32 and 61 min.

3. Results

3.1. Perceived competition

Hypothesis 1. *Learners within the low standard discrepancy groups*

(*LSD*) *perceive higher competition than learners within the high standard discrepancy groups* (*HSD*).

In order to verify hypothesis 1, a univariate analysis of covariance (ANCOVA) was conducted for *perceived competition* as dependent variable, *standard discrepancy* as between-subject factor, and *gaming experience* as covariate. All pre-defined test assumptions were not significantly impaired, Levene's test F (1, 83) = 0.25; p = 0.62. No significant differences are shown for standard discrepancy, F (1,82) = 4.23, p = 0.076, η_p^2 = 0.04. However, descriptive results show a tendency towards higher gaming experience corrected (M = 2.88) scores of perceived competition for the low discrepancy group (M_{LSD} = 3.44, SE = 0.18) in contrast to high discrepancy group (M_{HSD} = 2.98, SE = 0.18). Therefore, hypothesis 1 cannot be supported, although a tendency appeared.

3.2. Competitive effort

Hypothesis 2a. *Learners within the low standard discrepancy groups invest lower competitive effort than learners within the high standard discrepancy groups.*

Hypothesis 2b. *Learners with possibilities to repeat the levels* (*PR*) *invest less competitive effort than learners without this possibility* (*WR*).

In order to verify hypothesis 2, two ANCOVAs were conducted for *competitive effort* as dependent variable, *standard discrepancy* and *choice* as between-subject factors, and *gaming experience* as covariate. All pre-defined test assumptions were not significantly violated, Levene's test F(3, 81) = 1.36; p = 0.26. Significant differences are shown for standard discrepancy, F(1, 80) = 6.93, p = 0.01, η_p^2 = 0.08. Participants with high standard discrepancy (M_{HSD} = 3.09, SE = 0.17) reported higher competitive effort than students with low standard discrepancy (M_{LSD} = 2.47, SE = 0.16). In addition, a significant main effect was found for choice, F(1,80) = 7.72, p = 0.007, η_p^2 = 0.09. Students with a possibility to repeat (M_{PR} = 2.45, SE = 1.17) perceived their amount of competitive effort lower than students without an option to repeat (M_{WR} = 3.11, SE = 0.16). No significant interaction was found, F(1,80) = 3.9, p = 0.052, η_p^2 = 0.05, however, a tendency was observed. Descriptive results show that participants in the high standard discrepancy without an option to repeat (M_{HSD*WR} = 3.65, SE = 0.23) perceived less competitive effort than participants in the low standard discrepancy group with an option to repeat (M_{LSD*PR} = 2.37, SE = 0.23). Altogether, hypothesis 2a and 2b can be supported and in addition, preliminary evidence for a positive interaction was found.

3.3. Motivation & emotion

Hypothesis 3a. *Learners within the low standard discrepancy groups have higher relative autonomous motivation and perceive more positive emotions than learners within the high standard discrepancy groups.*

Hypothesis 3b. *Learners with possibilities to repeat the levels have higher relative autonomous motivation and perceive less negative emotions than learners without this possibility.*

In order to verify the motivational component of hypothesis 3, two ANCOVAs were conducted for *relative autonomous motivation* as dependent variable, *standard discrepancy* and *choice* as between-subject factors, and *gaming experience* as covariate. All pre-defined test assumptions were not significantly violated, Levene's test F(3, 81) = 0.19; p = 0.90. No significant differences are shown for standard discrepancy, F(1,80) = 0.1, p = 0.75, η_p^2 = 0.001. In contrast, a significant main effect was found for choice, F(1,80) = 6.97, p = 0.01, η_p^2 = 0.08. Students with a possibility to repeat (M_{PR} = 6.01, SE = 0.90) perceived a higher relative autonomous motivation than students without an option to repeat (M_{WR} = 2.70, SE = 0.87). No significant interaction was found, F(1,80) = 0.21, p = 0.65, η_p^2 = 0.003.

To shed more light on each concept within the motivation scale, follow-up ANCOVAs were conducted for either *intrinsic motivation, identified regulation, external regulation, amotivation* as dependent variables, *choice* as between-subject factor, and *gaming experience* as covariate. Concerning intrinsic motivation, all pre-defined test assumptions were not significantly violated, Levene's test F(1, 83) = 0.004; p = 0.95. No significant main effect was found for choice, F(1,82) = 3.53, p = 0.06, η_p^2 = 0.04. However, a tendency towards high scores within the group perceiving an additional option to repeat (M_{PR} = 4.82, SE = 0.19) than their counterparts can be revealed (M_{WR} = 4.32, SE = 0.18). Regarding identified regulation, no significant violation of pre-defined test assumptions are shown, Levene's test F(1, 83) = 0.001; p = 0.97, and a significant effect is demonstrated, F (1,82) = 4.50, p = 0.04, η_p^2 = 0.05. Students with a possibility to repeat (M_{PR} = 4.96, SE = 0.16) perceived a higher amount of identified regulation than students without an option to repeat (M_{WR} = 4.48, SE = 0.16). For external regulation, all pre-defined test assumptions were not significantly violated, Levene's test F(1, 83) = 0.32; p = 0.57. Again, a significance tendency towards lower scores within the group with a choice (M_{PR} = 4.01, SE = 0.23) than without a choice (M_{WR} = 4.62, SE = 0.22) can be seen, F(1,82) = 3.45, p = 0.07, η_p^2 = 0.04. Concerning amotivation, pre-defined test assumptions were not significantly violated, Levene's test F(1, 83) = 1.23; p = 0.27. A significant difference can be shown, F(1,82) = 7.22, p = 0.01, η_p^2 = 0.08. Participants with a choice to repeat (M_{PR} = 2.25, SE = 0.17) reported lower scores of amotivation than students without a choice to repeat (M_{WR} = 2.90, SE = 0.16).

In order to handle the assumption regarding emotions, a multivariate analysis of covariance (MANCOVA) was conducted for *standard discrepancy* and *choice* as independent variables, *enjoyment, pride, anger, shame, hopelessness* as dependent variables, and *gaming experience* as covariate. All pre-defined test assumptions were not significantly violated, Box's M (45, 15624.0) = 77.803, p = 0.012 (As suggested in the literature, we used α = 0.001 for the Box's-M-test; Field, 2009). Significant main effects were found for choice, Wilks's statistics $\wedge$ = 0.86 F (5,76) = 2.50, p = 0.04, η_p^2 = 0.14, but neither for standard discrepancy, Wilk's $\wedge$ = 0.92, F (5,76) = 1.37, p = 0.24, η_p^2 = 0.08, nor for the interaction, Wilk's $\wedge$ = 0.96 F (5,76) = 0.57, p = 0.73, η_p^2 = 0.04. Follow-up ANCOVAs with *choice* as independent variable reveal no significant effects for pride, F(1,82) = 0.16, p = 0.69, η_p^2 = 0.002, and shame, F(1,82) = 1.36, p = 0.25, η_p^2 = 0.02. In contrast, significant effects can be shown for enjoyment, F(1,82) = 6.53, p = 0.01, η_p^2 = 0.07, hopelessness, F(1,82) = 6.31, p = 0.01, η_p^2 = 0.07, whereas no significant violation of predefined test assumptions occurred, Levene's test for enjoyment, F(1, 83) = 0.01; p = 0.92, and for hopelessness, F(1, 83) = 1.95; p = 0.17. Students reported higher scores of enjoyment (M_{PR} = 3.87, SE = 0.13) and lower scores of hopelessness (M_{PR} = 1.38, SE = 0.13) with a choice to repeat. Within the WR condition, the scores of enjoyment (M_{WR} = 3.40, SE = 0.13) and hopelessness (M_{WR} = 1.83, SE = 0.12) were significantly decreased. According to the fact that pre-defined test assumptions were significantly violated for anger, Levene's test F(1, 83) = 8.21; p = 0.005, a Mann-Whitney-U-test was conducted, which revealed a significant difference, U = 702.5, z = −2.06 p = 0.04, r = −0.22. Since a median comparison shows no differences ($Mdn_{PR} = 1 = Mdn_{WR}$), descriptive results concerning the arithmetic mean reveal that students with a choice reported lower levels of anger ($M_{PR} = 1.24$, $SD = 0.56$) than students without a choice ($M_{WR} = 1.53$, $SD = 0.80$). Therefore, hypothesis 3a cannot be confirmed. Furthermore, statistical analysis revealed support for hypothesis 3b regarding relative autonomous motivation, enjoyment, hopelessness and anger but not for pride and shame.

3.4. Learning

Hypothesis 4a. *Learners within the low standard discrepancy groups differ in their learning outcomes compared to learners within the high standard discrepancy groups.*

Hypothesis 4b. *Learners with possibilities to repeat the levels learn more than learners without this possibility.*

In order to come to decision whether one of these hypotheses are correct, an ANCOVA was conducted for either *retention* as dependent variable, *choice* and *standard discrepancy* as between-subject factors, and *gaming experience* as covariate. Since predefined test assumptions are violated, Levene's test $F(3, 81) = 5.35$; $p = 0.002$, a *U*-test revealed a significant difference for standard discrepancy, $U = 633.5$, $z = -2.52$ $p = 0.012$, $r = -0.27$, but not for choice, $U = 759.5$, $z = -1.33$ $p = 0.18$, $r = -0.14$. Median comparison shows higher scores of retention for high standard discrepancy ($Mdn_{HSD} = 14$) than low standard discrepancy ($Mdn_{LSD} = 13$). Therefore, hypothesis 5a can be confirmed, whereas hypothesis 5b was not supported.

To gain deeper insight into the learning process, efficiency was analyzed. More specifically, a MANCOVA was conducted with the independent variables *standard discrepancy* and *choice*, the dependent variables *efficiency effort*, *efficiency load*, and the covariate *gaming experience*. All-predefined test assumptions were not significantly violated, Box's M (9, 68,994.5) $= 4.44$, $p = 0.90$. Significant main effects were found for choice, Wilks's statistics $\wedge = 0.90$ $F(2,79) = 4.43$, $p = 0.015$, $\eta^2_p = 0.10$, but not for standard discrepancy, Wilks's $\wedge = 0.96$ $F(2,79) = 1.73$, $p = 0.18$, $\eta^2_p = 0.04$, and not for the interaction, Wilks's $\wedge = 0.99$ $F(2,79) = 0.14$, $p = 0.87$, $\eta^2_p = 0.003$. A follow-up ANCOVA for the factor choice revealed a significant difference for efficiency load, $F(1,80) = 8.97$, $p = 0.004$, $\eta^2_p = 0.10$, and for efficiency effort, $F(1,80) = 4.06$, $p = 0.047$, $\eta^2_p = 0.05$, whereas all pre-defined test assumptions were not significantly violated, Levene's test for efficiency load, $F(3, 81) = 0.45$; $p = 0.72$ and Levene's test for efficiency effort, $F(3, 81) = 0.32$; $p = 0.81$. Students in the group receiving choices ($M_{PR} = 0.36$, $SE = 0.16$) were more efficient concerning mental load than students without choices ($M_{WR} = -0.33$, $SE = 0.16$). In addition, participants with choices ($M_{PR} = 0.21$, $SE = 0.15$) were more efficient regarding mental effort than students receiving no choice ($M_{WR} = -0.20$, $SE = 0.14$).

3.5. Goal orientations

Additionally, overall pearson-correlation coefficients were conducted to explore the coherence between goal orientation (GO) and cognitive and competitive factors. Results are outlined in Table 1.

Results show that mental effort is rather positively related to task orientation and self/other avoidance. Mental load is also positively related to task orientations and self-approach. Perceived competition is positively related to task orientations and other goals but not related to self-approach and avoidance. Competitive effort is positively related to all goal orientations except self-avoidance. Furthermore competitive effort is positively related to mental load as well as to mental effort while perceived competition is just related to mental effort.

4. Discussion

The experiment delivers various interesting and previously undiscovered outcomes. First, a significant difference regarding perceived competition could not be found. Nonetheless, a tendency in line with the hypothesized direction indicating higher perceived competition within the LSD conditions could be observed. Regarding the competitive effort measure, players within the LSD and the PR groups reported significantly less competitive effort, than their experimental counterparts. Additionally, a tendency towards a positive interaction of the experimental factors appeared. Concerning motivational measures, a significant difference for choice but not for standard discrepancy within the relative autonomous motivation scale could be observed. The option to repeat increased the motivation of students. The analyses of all motivation scales revealed a tendency for higher intrinsic and extrinsic motivation within the PR condition. Additionally, significant differences for identified regulation and amotivation were observed. More specifically, participants within the PR groups showed higher intrinsic motivation and identified regulation, and lower external regulation and amotivation than members of the WR groups. A similar pattern emerged among the emotional measures. A significant effect for choice but not for standard discrepancy was observed. Further tests revealed higher enjoyment and lower anger and hopelessness assessments within the PR conditions. Additionally, no differences regarding pride and shame were observed.

A central outcome of our experiment was found within the learning measures. The calculations showed an effect for the standard discrepancy condition but not for the second factor choice. More specially, a significantly higher learning result within the HSD condition was observed. In contrast, efficiency measures revealed a significant difference for choice, but nor for standard discrepancy. The following calculations showed increased efficiency load and efficiency effort scores within the PR condition. Finally, the exploratory analysis of goal orientation revealed some interesting connections as well, although it should be noted that correlations do not prove cause-and-effect patterns. For example, other-goal orientations are more related to competitive factors such as perceived competition and competitive effort than to overall mental load or mental effort. Furthermore, a significant relation between task goal orientation and all cognitive and competitive factors appeared. In contrast, the self-goal orientation revealed a more heterogeneous picture. While self-approach goal orientation correlated with mental load and competitive effort, self-avoidance goal orientation showed a significant correlation with mental effort. Finally, perceived competition was not related to self-goal orientation at all.

The findings suggest that manipulating the standard discrepancy does not impact the perceived competition significantly. This might be explained by the relatively scarce presentation of the leaderboard. Participants were not confronted with their relative position all the time, but only after the levels. Therefore, the impact of their relative position might be dampened by the remaining gameplay. This was not manipulated differently between the groups. In contrast, significant effects emerged during the investigation of competitive effort measures. In line with FIT (Kluger & DeNisi, 1996), higher competitive effort for players within the HSD groups could be observed. When players perceived a discrepancy between their actual position and the desired goal, they increased their effort. Interestingly, players close to a standard

Table 1
Correlations among goal orientations (GO), cognitive and competitive factors.

Variable	1	2	3	4	5	6	7	8	9	10
1. GO task-approach	–									
2. GO task-avoidance	0.66***	–								
3. GO self-approach	0.46***	0.43***	–							
4. GO self-avoidance	0.43***	0.50***	0.84***	–						
5. GO other-approach	0.40***	0.43***	0.63***	0.65***	–					
6. GO other-avoidance	0.46***	0.48***	0.66***	0.69***	0.87***	–				
7. Mental effort	0.60***	0.52***	0.21	0.27*	0.21	0.24*	–			
8. Mental load	0.36**	0.25*	0.22*	0.18	0.13	0.19	0.32**			
9. Perceived competition	0.26*	0.27*	0.17	0.13	0.41***	0.28**	0.32**	-0.10	–	
10. Competitive effort	0.34**	0.40***	0.22*	0.12	0.26*	0.24*	0.35**	0.39**	0.34**	–

Note. $N = 85$ for all measures. * indicates $p < 0.05$. ** indicates $p < 0.01$. *** indicates $p < 0.001$.

reported lower effort, although the increased perceived competition (on a descriptive level) would indicate reverse explanations. These results indicate that punctual presented leaderboards might be rather a feedback tool than an instrument for permanent competitive gameplay. Additionally, the possibility to repeat as a mechanic to undo errors and to improve anytime, reduced competitive effort. Player could simply *try out* how good they are on their first run and then decide whether they want to invest more effort to improve their standing or simply continue the game. This also explains the observed tendency towards an interaction between standard discrepancy and the possibility to repeat. As players within the HSD group wanted to improve and noticed that they cannot repeat levels, they might have invested even more effort to improve on their first run in the following level. In contrast, as players who could try out how good they are additionally experienced that they are within standard range, they might have lowered their competitive effort even more.

The lack of effects on motivational measures within the manipulation of standard discrepancy can be explained similar to the impacts on perceived competition. As the leaderboard did not influence players during playtime, motivational measures did not reach significance. In contrast, the option to repeat influenced how players perceived their gameplay and thus, their relatively autonomous motivation. When a choice is presented, the rather self-determined constructs (Ryan & Deci, 2000) of intrinsic motivation and identified regulation increased, whereas the reported values for rather non-self-determined concepts of external regulation and amotivation were lowered. Giving players a choice to organize their playing process will shift their motivation towards a self-determined orientation. A similar pattern could be observed for measures regarding emotion. Again, leaderboards did not influence the overall gaming experience in order to induce effects, but choice significantly altered how the gameplay was perceived. The option to undo errors and improve at one's convenience reduced anger, hopelessness and increased overall enjoyment. Furthermore, players who perceived positive emotions like enjoyment could choose to keep playing to maintain this feeling, while players who perceive negative emotions like anger could alter their playstyle on the repetition to reduce this feeling.

Regarding learning outcomes, different results appeared. No significant effect for the factor choice was found, indicating that learners reached a skill level on the first playthrough that was not easy to increase by simple repetition. This might be very different for other educational videogames and might have occurred, because only five questions per level were included. Therefore, only a relatively low number of 15 facts was implemented which might be comparably easy to memorize on the first playthrough. However, standard discrepancy influences all learning outcomes. In line with results from competitive effort, players who perceived that they have a larger standard discrepancy learned more. This might be because of their increased effort, but learning outcomes neither correlate with competitive effort nor with mental effort. It seems more likely that players who were told that their performance is not sufficient (HSD), kept information longer, and therefore performed better on the test after the play sessions. In contrast, players who were told that their performance is already sufficient (LSD) were more likely to abandon the previously learned information. This might have decreased their post-test outcomes.

In contrast to the retention outcomes, calculations of efficiency revealed only significant impacts for the factor choice. Again, the punctual presentation of the leaderboard might not have influenced the overall gameplay in order to affect learning efficiency. In contrast, the option to repeat components of the game and to organize oneself, induced measureable effects. As players could decide when to focus on certain parts of the learning material and repeat elements if necessary, they learned more effectively. However, it should be noted that the used videogame did not provide much content besides learning. If another game contains more entertainment-based content, the implementation of choice might also increase the use of non-educational game elements, and subsequently, harm efficiency.

Exploratory correlations show that both task-approach and task-avoidance goals are related to all cognitive and competitive measures. It can be deduced that players who choose the given task as their main focus of goal orientation invested more mental effort and subsequently showed increased cognitive load. Additionally, it can be assumed that these players accepted the competitive element of the game more. This seems comprehensible, as this was specifically demanded by the given task. However, such a coherent picture does not appear when self- and other-goal orientations are examined. Self-approach orientations are related to mental load and competitive effort. This indicates that the improvement of the performance regarding to oneself induces some sort of competition and therefore, adds mental load and competitive effort. In contrast, self-avoidance goals show only one connection with mental effort. This implies that a more defensively oriented self-avoidance goal structure increases mental effort without inducing further competition. This seems reasonable, as learners with approach orientations seem to be more likely to participate in a competitive scenario than avoidance oriented players. Finally, both goal orientations focusing on others show positive correlations with competitive effort and perceived competition. Minding that the competitive scenario was built around comparisons with other players, this connection appears logical. Players who focus on these goal orientations seem to concentrate more on competitive facets, based on the result of increased reported competition. Additionally, they might use the competitive mechanic more frequently resulting in increased competitive effort. In contrast to the others-approach, only others-avoidance goal orientation show a significant correlation with mental effort. This indicates a slightly stronger effect for others-avoidance orientations over others-approach orientations. This effect might stem from the experimental nature of the test. Students might have tried to succeed while being observed within an experimental study. If the educational videogame is part of a voluntarily play session, this result might look different.

4.1. Implications

The experiment offers several theoretical and practical implications. First of all, results show many congruities with the FIT. It was demonstrated that manipulating the leaderboard does not only induce competitive effects, but also serves as an element of feedback that needs to be considered. For example, as the position within the leaderboard effects the overall gameplay only marginally, strong effects on perceived competition could not be observed. In contrast, the feedback of being close or far from desired standards did effect competitive effort and learning. This leads to the practical implication that leaderboards should be interwoven with gameplay or presented more often, to increase its competitive impact. Additionally, their use should be reconsidered regarding the implied feedback, especially if players might easily reach scores close to their desired standards. Likewise, it can be derived that the assumed beneficial effects of leaderboards might not be distributed equally among all players. Connecting these implications with our research questions, we successfully showed heterogeneous effects of leaderboards among learners. Furthermore, we demonstrated that learners' behavior and performance while interacting with leaderboards might be explained by different theories such as the FIT or theories of social comparison. However, none of them is capable of entirely predicting the observed outcomes.

The second implication can be drawn from the manipulation of choice. Despite that this factor did only impact learning efficiency and not learning outcomes, it was demonstrated that presenting a choice did lead to a more self-determined motivational orientation and to beneficial impacts within emotions. Thus, a simple way to empower students while giving them more control over their learning process was demonstrated. This influenced the perception of the learning material positively. Altogether, we could not show an entirely positive picture, but a more specific view on how choice might affect the player within educational videogames. Another implication can be drawn from exploratory analysis of goal orientation. As task-orientations show the strongest correlations with cognitive and competitive measures, it can be derived that these orientations appear as the most potent orientations regarding beneficial interactions with educational videogames. As designers seek to optimize the given task, players who orient themselves at these tasks might benefit the most from such considerations. Additionally, as only minor differences but no dominant pattern between avoidance and approach goals were found, it can be concluded that competitive gameplay induced by leaderboards might not be affected by one of these orientations fundamentally more. This might be especially reassuring, minding the culturally different perceptions of competition. Finally, a successful solution how to teach simple information with an entertaining jump-and-run game was demonstrated. We received an overall positive feedback that can be underpinned by some of the regular measures as well. For example, items within the AEQ (e.g., „*After this educational videogame, I start looking forward to the next educational videogame*") or the SIMS (e.g., "*Because this activity is fun*") were rated above average.

4.2. Limitations

Research analyzing complex media like educational videogames is limited by several factors. First of all, conclusions can only be safely applied the type of game used within the experiment. Although, we have tried to make manipulations as generalizable as possible, different implementations of leaderboards or choice might result in different outcomes. For example, leaderboards could make the game more social, but this is not harnessed within the design as no user interaction were provided. Therefore, some beneficial effects of leaderboards might not appear. Additionally, the leaderboard may have targeted the wrong value. It contained who plays the fastest, which is not necessarily the player who learns the most. Maybe results would have been different, if the leaderboard contained values like the amount of information collected per correct answer. This might have motivated the players not to learn faster, but to learn more efficiently. Additionally, players might have spent more time on the task if they are more interested in the topic than in the gamification component. Thus, they might have learned more although the strictly time-based leaderboard indicated a lower performance. This discrepancy might have affected the impact of the leaderboard. Learning measures limit the experiment as well. Further measurements of competition and learning, especially learning over time and the addressed value of the learning content is necessary to underpin the provided explanations. Furthermore, homogeneity of variance was not given in learning data. Therefore, *U*-tests were computed to investigate whether experimental manipulation impacts learning. Since nonparametric methods have a lower power than parametric equivalents (e.g., Renkewitz & Sedlmeier, 2013), the lack of influence of choice on learning outcomes could be explained by the limitations of nonparametric methods. Finally, two of our scales showed low reliability. Therefore, further research focusing on these two scales might show more facets that might have been overlooked within this study.

4.3. Future directions

This research needs to be updated frequently, as new, more complex and detailed forms of representations of user skills and expertise emerge. For example, the BERGE framework (Seaborn et al., 2013) provides multiple layers of information and tries to induce prosocial behavior through group based leaderboards. Furthermore, the relation between competitive effort and learning should be investigated further. Although standard discrepancy effected both measures, it remains unclear how these concepts are related. Especially, as they show no direct correlation. Finally, theoretical frameworks regarding educational videogames, do not include personal traits or represent unique playthroughs sufficiently. This research has successfully demonstrated that impacts of educational videogames can not only differ because of personal differences, but also because of a difference between the individual playthroughs. This issue gains importance as educational videogames can be very complex compared to traditional learning media and provide very different experiences. Conclusions for educational videogames that root within frameworks that focus on traditional learning material (e.g., Cognitive Load Theory, Sweller, van Merrienboer, & Paas, 1998), need to include this issue more. Finally, it remains a big challenge for designers of educational videogames to ensure that players perceive the learning experience they want them to.

References

Kudos Badges.(2015) [Developed by ISW].

Bajko, R., Hodson, J., Seaborn, K., Livingstone, P., & Fels, D.. (2015). Guilds, die rolls, and leaderboards: Gamification of two undergraduate multimedia and social media courses. Paper presented at the proceedings of the EDSIG conference.

Barata, G., Gama, S., Jorge, J., & Gonçalves, D. (2013). *So fun it hurts—gamifying an*

engineering course foundations of augmented cognition (pp. 639–648). Springer. http://dx.doi.org/10.1007/978-3-642-39454-6_68.
Brueghel, J., & the Younger. (1640). *Allegory on the Tulip mania* [Painting]. Retrieved from: https://upload.wikimedia.org/wikipedia/commons/e/e7/Allegorie_der_Tulipomanie.jpg.
Butler, C. (2013). *The effect of leaderboard ranking on players' perception of gaming fun online communities and social computing* (pp. 129–136). Springer. http://dx.doi.org/10.1007/978-3-642-39371-6_15.
Cagiltay, N. E., Ozcelik, E., & Ozcelik, N. S. (2015). The effect of competition on learning in games. *Computers & Education, 87*, 35–41. http://dx.doi.org/10.1016/j.compedu.2015.04.001.
Cheong, C., Filippou, J., & Cheong, F. (2014). Towards the gamification of learning: Investigating student perceptions of game elements. *Journal of Information Systems Education, 25*(3), 233.
Codish, D., & Ravid, G.. (2014). Personality based gamification-Educational gamification for extroverts and introverts. Paper presented at the proc. 9th Chais conf. for the study of innovation and learning technologies: Learning in the technological era.
De Matteis, P. (1714). *Allegory of the consequences of the Peace of Utrecht* [Painting]. Retrieved from: https://upload.wikimedia.org/wikipedia/commons/7/76/Paolo_De_Matteis_-_Allegory_of_the_Consequences_of_the_Peace_of_Utrecht_-_Google_Art_Project.jpg.
Diablo III. (2012). *[Developed by blizzard Entertainment]: Activision blizzard*.
Domínguez, A., Saenz-de-Navarrete, J., De-Marcos, L., Fernández-Sanz, L., Pagés, C., & Martínez-Herráiz, J.-J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & Education, 63*, 380–392. http://dx.doi.org/10.1016/j.compedu.2012.12.020.
Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34*(3), 169–189. http://dx.doi.org/10.1207/s15326985ep3403_3.
Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 100*, 613–628. http://dx.doi.org/10.1037/0022-0663.100.3.613.
Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3 × 2 achievement goal model. *Journal of Educational Psychology, 103*(3), 632. http://dx.doi.org/10.1037/a0023952.
Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*(2), 117–140. http://dx.doi.org/10.1177/001872675400700202.
Field, A. (2009). *Discovering statistics using SPSS*. Sage Publications.
Foldit. (Beta). [Developed by University of Washington Departments of Computer Science & Engineering and Biochemistry].
Garcia, S. M., & Tor, A. (2007). Rankings, standards, and competition: Task vs. scale comparisons. *Organizational Behavior and Human Decision Processes, 102*(1), 95–108. http://dx.doi.org/10.1016/j.obhdp.2006.10.004.
Garcia, S. M., & Tor, A. (2009). The n-effect more competitors, less competition. *Psychological Science, 20*(7), 871–877. http://dx.doi.org/10.1111/j.1467-9280.2009.02385.x.
Garcia, S. M., Tor, A., & Gonzalez, R. (2006). Ranks and rivals: A theory of competition. *Personality and Social Psychology Bulletin, 32*(7), 970–982. http://dx.doi.org/10.1177/0146167206287640.
Glasserman, P. (2003). *Monte Carlo methods in financial engineering*. New York: Springer-Verlag.
Glova, A. J. M., Asuncion, N. S. M., Martin, J. M. L., Manzan, L. V., & Pagtaconan, W. C. R.. (2014). SignApp: A mobile learning tool for hearing-impaired learners. 12th national eLearning conference.
Guay, F., Vallerand, R. J., & Blanchard, C. (2000). On the assessment of situational intrinsic and extrinsic motivation: The Situational Motivation Scale (SIMS). *Motivation and Emotion, 24*(3), 175–213.
Halan, S., Rossen, B., Cendan, J., & Lok, B.. (2010). High score!-motivation strategies for user participation in virtual human development. Paper presented at the intelligent virtual agents.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. http://dx.doi.org/10.3102/003465430298487.
Hemke, F., Meyer, H. A., Hühne, K., Schneider, M., & Wohlgemuth, V. (2014). Usability-Quiz: Die Begriffswelt der Usability durch ein Lernspiel vermitteln. In *Proceedings of the 28th EnviroInfo 2014 conference*.
IJsselsteijn, W. A., de Kort, Y. A. W. & Poels, K. (in preparation). The Game Experience Questionnaire: Development of a self-report measure to assess the psychological impact of digital games. Manuscript in preparation.
Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254. http://dx.doi.org/10.1037/0033-2909.119.2.254.
Krell, M. (2015). Evaluating an instrument to measure mental load and mental effort using Item Response Theory. *Science Education Review Letters, 2015*.
Kusurkar, R., Ten Cate, T. J., Vos, C., Westers, P., & Croiset, G. (2013). How motivation affects academic performance: A structural equation modelling analysis. *Advances in Health Sciences Education, 18*(1), 57–69. http://dx.doi.org/10.1007/s10459-012-9354-3.
Landers, R. N., & Landers, A. K. (2015). An empirical test of the theory of gamified learning the effect of leaderboards on time-on-task and academic performance. *Simulation & Gaming*. http://dx.doi.org/10.1177/1046878114563662.
Lime Survey. (2015). *[Developped by the LimeSurvey project team] LimeSurvey GmbH*.
Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*. Prentice-Hall, Inc.
Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*, 705–717. http://dx.doi.org/10.1037/0003-066X.57.9.705.
Nebel, S., Schneider, S., & Rey, G. D. (2016). From duels to classroom competition: Social competition and learning in educational videogames within different group sizes. *Computers in Human Behavior, Part A*. ISSN: 0747-5632, *55*, 384–398. http://dx.doi.org/10.1016/j.chb.2015.09.035.
Marsaglia, G., & Bray, T. A. (1964). A convenient method for generating normal variables. *SIAM Review, 6*(3), 260–264.
Path of exile.(2013) [Developped by grinding gear games].
Pekrun, R., Elliot, A. J., & Maier, M. A. (2009). Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of Educational Psychology, 101*(1), 115.
Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36–48. http://dx.doi.org/10.1016/j.cedpsych.2010.10.002.
Pekrun, R., Götz, T., & Perry, R. P. (2005). *Achievement emotions questionnaire (AEQ). user's manual*. To be published Manuscript. Munich: University of Munich https://de.scribd.com/doc/217451779/2005-AEQ-Manual.
Plass, J. L., O'Keefe, P. A., Homer, B. D., Case, J., Hayward, E. O., Stein, M., et al. (2013). The impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation. *Journal of Educational Psychology, 105*, 1050–1066. http://dx.doi.org/10.1037/a0032688.
Platzer, J. G. (1750). *Allegory of the four seasons* [Painting]. Retrieved from: https://upload.wikimedia.org/wikipedia/commons/6/69/Johann_Georg_Platzer_001.jpg.
Renkewitz, F., & Sedlmeier, P. (2013). *Forschungsmethoden und Statistik in der Psychologie* (2nd ed.). München: Pearson Studium.
Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68. http://dx.doi.org/10.1037/0003-066X.55.1.68.
Sarangi, S., & Shah, S. (2015). Individuals, teams and organizations score with gamification: Tool can help to motivate employees and boost performance. *Human Resource Management International Digest, 23*(4), 24–27. http://dx.doi.org/10.1108/HRMID-05-2015-0074.
Schubert, P., Paulsen, L., & Hager, J. (2014). Auswirkungen von Gamification in Enterprise Collaboration Systems. In A. Butz, M. Koch, & J. Schlichter (Eds.), *Mensch und Computer 2014 Tagungsband* (pp. 3–14). München: Oldenbourg Wissenschaftsverlag.
Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies, 74*, 14–31. http://dx.doi.org/10.1016/j.ijhcs.2014.09.006.
Seaborn, K., Pennefather, P., & Fels, D. I.. (2013). Reimagining leaderboards: Towards gamifying competency models through social game mechanics. Paper presented at the proceedings of the first international conference on gameful design, research, and applications.
Simões, J., Redondo, R. D., & Vilas, A. F. (2013). A social gamification framework for a K-6 learning platform. *Computers in Human Behavior, 29*(2), 345–353. http://dx.doi.org/10.1016/j.chb.2012.06.007.
Sweller, J., van Merrienboer, J. G., & Paas, F. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296. http://dx.doi.org/10.1023/A:1022193728205.
Ter Vrugte, J., De Jong, T., Vandercruysse, S., Wouters, P., Van Oostendorp, H., & Elen, J. (2015). How competition and heterogeneous collaboration interact in prevocational game-based mathematics education. *Computers & Education, 89*, 42–52. http://dx.doi.org/10.1016/j.compedu.2015.08.010.
Unreal Engine 4.(2015) [Developed by epic games].
Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*(1), 16–26. http://dx.doi.org/10.1080/00461520701756248.
Vandercruysse, S., Vandewaetere, M., Cornillie, F., & Clarebout, G. (2013). Competition and students' perceptions in a game-based language learning environment. *Educational Technology Research and Development, 61*, 927–950. http://dx.doi.org/10.1007/s11423-013-9314-5.
Vansteenkiste, M., Lens, W., De Witte, S., De Witte, H., & Deci, E. L. (2004). The 'why' and 'why not' of job search behaviour: Their relation to searching, unemployment experience, and well-being. *European Journal of Social Psychology, 34*(3), 345–363. http://dx.doi.org/10.1002/ejsp.202.
Vansteenkiste, M., Zhou, M., Lens, W., & Soenens, B. (2005). Experiences of autonomy and control among Chinese learners: Vitalizing or immobilizing? *Journal of Educational Psychology, 97*(3), 468.
Vorderer, P., Hartmann, T., & Klimmt, C.. (2003). Explaining the enjoyment of playing video games: The role of competition. Paper presented at the proceedings of the second international conference on entertainment computing.
Wang, H., & Sun, C.-T. (2011). Game reward systems: Gaming experiences and social meanings. In *Proceedings of the 5th Digital Games Research Association* (pp. 14–17).
Willems, C., Fricke, N., Meier, S., Meissner, R., Rollmann, K.-A., Voelcker, S., et al. (2014). Motivating the masses-gamified massive open online courses on openHPI. In *Proceedings of EDULEARN*.
Wu, W. F., & Magill, R. A. (2011). Allowing learners to choose: Self-controlled practice schedules for learning multiple movement patterns. *Research quarterly for exercise and sport, 82*(3), 449–457.