Gamifying a Simulation: Do a Game Goal, Choice, Points, and Praise Enhance Learning?

Notice

This is the author's version of a work that was accepted for publication in Journal of Educational Computing Research. Changes resulting from the publishing process, such as structural formatting, and other quality control mechanisms may not be reflected in this document. A definitive version was accepted for publication in Journal of Educational Computing Research (2018), DOI: 10.1177/0735633118797330. The paper was accepted: 8-AUG-2018.

Authors

Cyril Brom (corresponding author)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, 118 00, Prague, Czech Republic
brom@ksvi.mff.cuni.cz

Tereza Stárková
Faculty of Arts, Faculty of Mathematics and Physics, Charles University
tereza.starek@gmail.com

Edita Bromová
Faculty of Mathematics and Physics, Charles University
edita@email.com

Filip Děchtěrenko
Institute of Psychology, The Czech Academy of Sciences
filip.dechterenko@gmail.com

Acknowledgement

This study was primarily funded by Czech Grant Science Foundation (GA ČR). The beginning of this study was supported by Project nr. P407/12/P152 and the study's second half by Project nr. 15-14715S. Work of F. D. was supported by RVO 68081740.

We thank research assistants who helped to conduct the experiments, most notably: V. Dobrovolný, M. Stejskalová, K. Brožová. We also thank Laboratory of Behavioral and Linguistic Studies in Prague, where part of the study was conducted.

Abstract

Despite the increased interest in gamification approaches, there is a lack of comparative studies that shed light on the applicability of these approaches in educational contexts. In this explorative study, with an experimental design, university learners ($N = 98$) studied a complex process (i.e., how to brew beer) in a two-hour-long computerized simulation. In the experimental condition, the simulation featured the following game design elements: game goals, increased freedom of choice, points, virtual currency, and praise (i.e., a gamified simulation). These elements were absent in the simulation versions used in the two control conditions. No differences in learning outcomes and intrinsic motivation variables between the gamified simulation and its non-gamified versions were observed. The gamified simulation was perceived to be significantly easier than the non-gamified versions ($\eta_p^2 = 0.10$; $d = 0.74, 0.42$). Of the game elements used in this study, the participants perceived most positively a clear, game-like goal. The findings are consistent with self-determination theory, cognitive-affective theory of learning from media, and cognitive load theory. The findings also support the emerging notion that caution should be applied when using gamification approaches in educational contexts.

Keywords: gamification, serious games, simulations, multimedia learning, motivation, learning outcomes

Article

Introduction

Gamification is the use of game design elements in non-game contexts (Deterding et al., 2011). In educational settings, this term means the use of these elements in non-game educational materials and methods, which is referred to herein as *edu-gamification*. Common elements include points, achievement badges and leaderboards; clear, game-like goals and increased challenges; narratives; increased freedom of choice; and immediate feedback (see Darejeh and Salim, 2016; Dicheva et al., 2015). Unlike the digital game-based learning approach, which focuses on the use of "entire" games, the gamification approach puts emphasis only on the individual elements of games.

Edu-gamification approaches are posited to have motivational benefits, which are believed to help address education-relevant issues such as increasing participation in learning activities, familiarizing students with a new environment (e.g., a campus or library) and improving learning outcomes (e.g., Dicheva et al., 2015; Hamari et al., 2014; Hakulinen et al., 2013). However, comparative studies investigating the actual effects of a gamified educational experience compared to a control condition are still emerging, and the alleged claims are thus not well supported by solid empirical evidence (see Ortiz-Rojas, Chiluiza, & Valcke, 2017, for an early review). A vast majority of studies simply report user evaluations and/or log analyses (reviewed in Dicheva et al., 2015; Hamari et al., 2014). Much needed information is lacking about which added game design elements are beneficial or detrimental in educational settings.

The majority of edu-gamification solutions pertain to mobile applications or web-based learning platforms, for instance, those used in university courses (Darejeh and Salim, 2016; Dicheva et al., 2015; see also Ortiz-Rojas et al., 2017). It is less common to gamify stand-alone educational software, such as a computerized simulation of a complex process to be learned. In addition, the majority of solutions focus on computer science and ICT topics (i.e., "Information and Communication Technologies") (Dicheva et al., 2015).

This work presents an explorative experimental study of whether a common edu-gamification approach can improve learning using a stand-alone computer simulation of a non-ICT topic. The study thus investigated edu-gamification in a context in which comparative studies are still lacking. The game design elements in question included a game-like goal, points (which are converted to virtual "money" needed for achieving the goal), increased freedom of choice and praise that provides immediate feedback. In our study, university participants learned how to brew beer from one of the simulation version (that takes approximately 2 hours) and their intrinsic motivation, learning outcomes and subjective evaluation were assessed and compared (i.e., between-subject design). The study has the potential to indicate whether there are advantages to using these game elements in learning from computerized applications among students, similar to those in our research sample.

Background

Edu-gamification Studies with Comparative Design

Gamification is a topic that has evolved rapidly over the previous decade (cf. Deterding et al., 2011), and it has been applied in various contexts (see Darajeh and Salim, 2016; Dicheva et al., 2015; Hamari et al., 2014; Morschheuser et al., 2016). Gamified

experiences are thought to boost participants' engagement and motivation (see, e.g., Landers and Calan, 2011; Dicheva et al., 2015; McGonigal, 2011; Morschheuser et al., 2017; see also Darajeh and Salim, 2016) and thereby positively influence users' behavior. Gamifying educational experiences, i.e., edu-gamification, can be viewed as a gamification sub-field.

There is very limited knowledge about which game elements, if any, facilitate learning or improve other education-relevant outcomes, for what types of learners and in what contexts. Generally, edu-gamification case studies and studies with pre-post design (i.e., without control groups) have yielded partly positive findings (see Dicheva et al., 2015; Hamari et al., 2014), but studies with control groups have yielded fewer positive findings than studies without control groups. In particular, the recent review of gamified courses by Ortiz-Rojas and colleagues (2017), which combined single-group studies with comparative ones, described nine comparative studies with null/mixed results, three with positive results, and two with negative results[1].

For example, badges and/or points with leaderboards sometimes increase participation in various activities (Anderson et al., 2014; Barata et al., 2013; Denny, 2013; Halan et al., 2010; Hamari, 2017) or improve students' time management (Hakulinen et al., 2013). However, when Hanus and Fox (2015) used badges, optional "coins" and a leaderboard in one version of a university course on communication skills, they found that the gamified version did not improve students' grades and self-reported class effort. This version was also detrimental to their intrinsic motivation, satisfaction and, to some extent, sense of empowerment. In a set of three studies involving university students engaged in various variations of an ICT course, de-Marcos and colleagues (Domínguez et al., 2013; de-Marcos et al., 2014; de-Marcos et al., 2016) consistently found that competition-driven gamified course

[1] These numbers are not given in the review, but they were revealed to us by the authors of the review via an email dated 10 October 2017.

versions (using tasks setting short-term goals, trophies, badges and a leaderboard) helped students complete their homework assignments better. However, the gamified course versions hindered conceptual learning. In addition, roughly ¾ of participants in the gamified course versions rarely or never used gamified add-ons (Domínguez et al., 2013; de-Marcos et al., 2014).

Katz and colleagues (2014), using gamified software, engaged primary-school children in working-memory training by means of one of seven versions of a cognitive training application with various combinations of game-based elements. Training performance was best when achieved points were *not* displayed to the learners, which also resulted in a marginally lower perceived effort. The application version with all game-based elements was next to worst in terms of training performance. Near-transfer performance, enjoyment, excitation and perceived difficulty were unaffected by the manipulations. In contrast, using a similar set of subjects, Sandberg and colleagues (2014) showed that when a gamified vocabulary learning application (a narrative context, a challenge specified in the terms of the narrative, achievement medals and smileys, choice) was combined with an adaptive difficulty add-on, the enhanced application's version improved home learning. In a study involving university students, Gauthier (2015) reported a slight non-significant improvement in learning outcomes and no increase in enjoyment[2]; the learners were studying vascular anatomy using a web-based learning environment and the gamified version featured points, energy limit, leaderboard, achievements.

[2] Enjoyment data were not reported in the paper, but they were sent to us by the author via an email dated 29 October 2017.

Digital Game-based Learning Studies

One can assume that the domain of digital game-based learning (DGBL) might reveal which game elements facilitate learning and which do not. However, this is, by and large, not the case.

Value-added DGBL studies investigate whether adding a particular feature or set of features to an educational game enhances learning (compared to the same game without the feature). Because these studies manipulate just one or a few features, they are well suited to investigating the effects of individual game elements. Value-added DGBL studies are on the rise. However, with the exception of enhanced scaffolding designs, they have so far been few: the recent meta-analysis by Clark and colleagues (2016) identified eleven of them. Their results are inconclusive (Clark et al., 2016; see also Mayer, 2014, Ch. 5; Wouters & Oostendorp, 2017). For instance, the effects of presence vs. absence of competitive elements in educational games could inform gamification research looking into the effects of competition-related elements such as points, badges and leaderboards. Nevertheless, conflicting results were reported regarding the presence of competitive elements in educational games (e.g., cf. Plass et al., 2013 with Ke, 2008).

Seductive Details

Seductive, or extraneous, details (e.g., Mayer, 2009) are interesting visual or auditory additions to learning materials that provide tangentially relevant information not necessary for comprehending the core instructional message. Certain game-based elements can also be viewed as seductive details in a broader sense. These include a game narrative or user interface elements added to depict the number of achieved points.

Seductive details do not improve learning in general (Garner et al., 1992; Mayer, 2009; Rey, 2012). Although they may be beneficial under some conditions (Park et al., 2015), one should always be aware that the disadvantages of adding an extraneous detail to an educational tool may outweigh its potential benefits. For instance, the presence of a narrative (in a game-based context) was shown to be beneficial for children aged approximately 10 years (Cordova and Lepper, 1996; Sandberg et al., 2014) but not necessarily for college learners (Adams et al., 2012; Johnson-Glenberg & Megowan-Romanowicz, 2017).

Theoretical Standpoints

The abovementioned outcomes are not surprising when cognitive-motivational theoretical perspectives are considered. Three theories and points are important in the present context. First, the general idea that gamification approaches enhance learning by being "motivating and engaging" can be theoretically underpinned by the cognitive-affective theory of learning from media (CATLM; Moreno, 2005) and the notion of intrinsic motivation. Intrinsic motivation refers to doing an activity for its own sake, because it is inherently interesting and enjoyable (e.g., Ryan & Deci, 2000). Second, according to cognitive load theory (Sweller et al., 2011), most gamification methods are distracting, thereby countering possible effects of increased motivation and engagement. Some methods, however, can be beneficial by way of an easing of cognitive processing. Third, contrary to common expectations, some gaming elements when used outside of games may in fact not be "motivating and engaging" in the first place. This idea derives from self-determination theory (SDT) (Deci and Ryan, 1985). We now consider these three perspectives in turn.

Cognitive-affective theory of learning from media and affective-motivational factors. The first key assumption of the CATLM (Moreno, 2005) is that learners must be

mentally active when they process incoming information, i.e., they must be cognitively engaged. High cognitive engagement generally leads to better learning outcomes than does low cognitive engagement. The second key assumption is that certain affective-motivational factors can increase or decrease cognitive engagement. These factors include temporary affective-motivational states that develop within short time periods (i.e., minutes) during a learning experience or afterward. Pekrun organized these states, particularly those related to achievement contexts, along dimensions of valence (positive vs. negative), activation (activating vs. deactivating), and object focus (i.e., related to the undertaken activity or to the outcome) (Pekrun, 2006; Pekrun & Linnenbrink-Garcia, 2012). Here, we focus on activity-related states. Deactivating activity-related states such as boredom tend to produce low cognitive engagement, which hampers learning. In contrast, activating activity-related states, especially if they are positive, tend to boost cognitive engagement and thereby facilitate learning (the effects of negative activating states are more complex; e.g., Pekrun, 2006, p. 326). These positive-activating states, such as enjoyment, situational interest, or flow, are intrinsic motivation factors (cf. Ryan & Deci, 2000; Schiefele, 1999). One goal of gamification approaches is to target these factors and thereby boost cognitive engagement and learning outcomes (Figure 1).

Cognitive load theory. Cognitive resources the learner uses to process incoming information are limited. Elements of the learning environment designed to increase intrinsic motivation (for example, user interface representations of points or badges) must also be processed by learners. According to cognitive load theory (Sweller et al., 2011), this processing consumes a portion of the (limited) cognitive resources that could be otherwise devoted to processing meaningful learning content. These elements may thus function as seductive details. In addition, they may promote thinking not relevant to learning (e.g., Pekrun and Linnenbrink-Garcia, 2012, p. 264) such as thoughts pertaining to increased

competitive pressure. All in all, intrinsic motivation triggered by these elements may not always enhance learning outcomes, because these elements may induce an unnecessary cognitive load. Although learners may be highly cognitively engaged, their attention may be deflected away from the learning task.

However, at the same time, some game elements (e.g., subgoals, points) may better allocate limited cognitive resources by better structuring the learning task. This would have a positive effect on learning (Figure 1).

--- Insert Figure 1 around here ---

Self-determination theory. The third point to consider is that some gaming elements, when used outside of games, may not be motivational in the first place. This idea derives from SDT (Deci and Ryan, 1985; see also van Roy and Zaman, 2017), in which intrinsic motivation is the principal construct. SDT maintains that intrinsic motivation is fostered when the learning environment helps satisfy the learner's needs for autonomy, competence, and relatedness. The first two of these needs, i.e., autonomy and competence, are relevant here. When a game element undermines one of these needs, intrinsic motivation can be reduced.

Badges, leaderboards, and especially points are commonly assumed to motivate and/or provide useful feedback and to structure the task at hand (see, e.g., Dicheva et al., 2015; Hamari, 2017). However, they can actually be perceived by learners as expected tangible rewards. Expected tangible rewards may reduce the perception of autonomy and, according to SDT, undermine intrinsic motivation (Deci et al., 1999; but see also Cameron et

al., 2001). When learners feel they have earned too few points or badges, their perception of competence may decrease.

On a more positive side, choice can enable higher control in terms of SDT and thus increase the learner's sense of autonomy, which can enhance intrinsic motivation (Patall et al., 2008). A clear goal can help structure the learning task and thereby boost, or at least not reduce, learners' sense of competence and possibly increase their sense of autonomy. Moreover, based on cognitive load theory, structuring (also called segmenting) of learning can reduce cognitive load (Mayer, 2009, Ch. 9; see also above). Praise, as a verbal reward (Cameron et al., 2001, p. 3; Deci et al., 1999), can boost a learner's sense of competence (Vansteenkiste et al., 2009; p. 672). Badges or points, when they provide feedback regarding competence, can support the need for competence (van Roy and Zaman, 2017).

Summary. Individual game elements can have positive and negative effects in educational contexts (see Table 1). SDT makes it clear that the presence of a game element does not necessarily increase intrinsic motivation. If it does, then it can help allocate cognitive resources more effectively (according to CATLM). However, this may still not improve learning because the element may impose a high unnecessary cognitive load (according to cognitive load theory; see Figure 1). For instance, it is known that greater freedom of choice generally increases intrinsic motivation, but it does not always improve subsequent learning (Patall, Cooper, & Robinson, 2008). Likewise, praise has been empirically shown to be one of the least effective types of feedback (Hattie & Timperley, 2007). Whether the positives of added game elements outweigh the negatives cannot be predicted based on a theoretical basis. This question has to be examined empirically. Currently, there are substantial gaps in knowledge regarding which game elements (or combinations thereof) are beneficial (or detrimental or neutral) to learning; and for which types of learners and in which learning situations.

--- Insert Table 1 around here ---

This Study

This study contributes to the existing gamification literature in that it investigated the effects of common gamification elements in a context (i.e., a stand-alone computer simulation on a non-ICT topic) that has been the focus of few comparative studies. University learners study how to brew beer, either from a gamified, roughly two-hour-long, computerized simulation or from one of two control versions of the same simulation (random assignment). The simulation is known from previous research to be motivating and to promote learning (Brom, Bromová, Děchtěrenko, Buchtová, & Pergel, 2014). However, there are activities, such as participating in a simulation game, that are even more motivating for the target audience (Brom, Buchtová, Šisler, Děchtěrenko, Palme, & Glenk, 2014). Therefore, there is a room for improvement concerning motivation, and gamification can, conceivably, help with that. The simulation is of moderate difficulty for university learners (neither too difficult, nor too easy), as determined during pilot research.

Game elements

The game elements under investigation were as follows:

- a clear game goal,

- increased freedom in choice of tasks to work on,
- points given for doing well, which become virtual "money" needed for achieving a game goal,
- and verbal rewards in the form of praise with some learning-related information, which strengthens immediate feedback.

These game elements were selected for reasons of ecological validity; they are quite common in games and, according to our experience, are often used in gamification endeavors in practice (especially in the case of interventions for a single user). In addition, they were suggested by some learners participating in our previous study (Brom et al., 2014) as possible improvements to the (non-gamified) simulation, and they were verified as being appreciated by learners in pilot studies.

Why did we not employ fewer or more game elements? On the one hand, it would be methodologically "purer" to investigate the effects of a single element, but this choice would have limited sense from practical point of view, because the elements we selected were interdependent (e.g., the positive effects of a clear game goal might be undermined without increased freedom of choice). On the other hand, additional elements are frequently used in practice and are thus worth investigating. However, combining too many elements in a single manipulation is methodologically problematic in that it would complicate the interpretation of findings. We thus had to compromise; instead of combining all elements that merit investigation, we selected a "minimal set" of elements, such that removing a single element from the set would artificially reduce the believability of the intervention.

Research Questions

This was an exploratory study with three main research questions:

Q1: What is the net effect of the selected game elements on intrinsic motivation factors?

Q2: What is the net effect of the selected game elements on learning outcomes?

Q3: Do intrinsic motivation factors mediate the effect of gamification on learning outcomes?

We refrained from making directional predictions regarding the impact of the gamification approach because, as argued above, positive and negative effects could be expected (see Table 1 and Figure 1).

This study also began with a supplemental research question:

Q4: What is the net effect of the selected game elements on three supplementary variables: perceived difficulty, perceived learning, and generalized negative affect?

These three variables were not central to the study, but we measured them in nearly all our studies for future meta-analytical purposes; thus, we report the data here for the sake of completeness. Some research linked the perceived difficulty construct to certain aspects of cognitive load (e.g., DeLeeuw & Mayer, 2008), but this link has been questioned (see de Jong, 2010). In general, cognitive load and cognitive engagement, in particular, are notoriously difficult to measure, especially in long-duration intervention studies (e.g., Brünken, Plass, & Leutner, 2003; Brünken, Seufert, Paas, & 2010; de Jong, 2010). Even new instruments (e.g., Leppink et al., 2014) are not without issues (Stárková, Lukavský, Javora, &

Brom, submitted). Therefore, in this study, we relied on measuring learning outcomes and intrinsic motivation.

The relationships between variables investigated in this study are summarized in Figure 2. Intrinsic motivation was measured via five proxy variables, which are also depicted in this figure.

--- Insert Figure 2 around here ---

Comparison Conditions

We used two control groups, which we now discuss in detail. We previously used the simulation referred to in this study to investigate the so-called personalization principle (Brom et al., 2014). The personalization principle states that learners tend to learn better when instructional texts are written in a conversational rather than a formal style (Mayer, 2009). This principle is well supported empirically in short lessons (those spanning less than approximately thirty minutes) with text in English (see Ginns et al., 2013). We have studied it in the Czech context using a long lesson, i.e., with the beer brewing simulation. We generally found no differences between the use of the two language styles for instructional texts (Brom et al., 2014). Nevertheless, we used both conversational and formal simulation versions for comparison in the present study. The reason is that the language style in the conversational version was supported by a background narrative (because of the longer exposure), and the addition of a story element can itself be considered a gamification step (Dicheva et al., 2015; Hamari et al., 2014).

Elements of all three conditions are summarized in Table 2. From the gamification perspective, the simulation versions can be arranged on a spectrum from no gamification to a full-fledged game as follows: non-gamified formal version < non-gamified conversational version < gamified version.[3]

--- Insert Table 2 around here ---

Methods

Participants and Design

One hundred and six Czech university students were recruited. These students participated for course credit. Of these students, 98 were included in the analysis (M_{age} = 23.05 ± 2.53 [± SD]; 55 % males). These participants had diverse study backgrounds (psychology, computer science, art, new media studies, and philology). We included only low-prior-knowledge learners, i.e., those who scored fewer than 15 points on a test of self-assessed prior knowledge (detailed in Measures). An additional eight participants were excluded: two for high prior knowledge of beer brewing, two for being very tired at the

[3] To contrast the experimental condition with the two control conditions, we call both control conditions "non-gamified" despite the fact that the conversational non-gamified condition features a simple narrative. Alternatively, we could say that we had two experimental conditions (i.e., Experimental and Control 2 from Table 2) and one control condition (i.e., Control 1 from Table 2), but we prefer the former terminology.

experiment's beginning, one for not being a Czech or Slovak native speaker[4], two for not comprehending the instructions and one for feeling poorly during the experiment.

There were one experimental group (gamified, G, $n = 31$) and two control groups (non-gamified conversational, NC, $n = 34$; and non-gamified formal, NF, $n = 33$). The participants were randomly assigned to the conditions, and their genders and areas of study were balanced.

Materials – Simulation, Framing

The intervention is an interactive simulation originally developed by us for a previous study of ours (Brom et al., 2014). It was developed using the Netlogo toolkit (Wilensky, 1999). For the present experiments, three versions were used that corresponded to the study's conditions.[5] All the versions consist of four parts:

1. The *tutorial*, which demonstrates how to control the simulation (10-20 minutes).
2. The *linear part*, which demonstrates how to brew beer when every step is performed correctly (30-50 minutes).
3. The *error part*, which demonstrates the consequences of making errors or of not following the standard procedure as previously described (35-60 minutes).
4. The *task-solving part*, in which the learner brews several beers of a specific type in the simulation (30-40 minutes).

[4] The Slovak language is very similar to the Czech language. Many Slovak students study in the Czech Republic, and it is generally little trouble for Slovak university students to understand or speak Czech fluently.
[5] During the development, we followed multimedia learning principles (Mayer, 2009) whenever possible. Key steps in our designing of the gamified version can be mapped onto the steps of the gamification design method by Morschheuser and colleagues (2017) (although their method is tailored to business domains).

The simulation is presented in the Czech language and is self-paced. The graphical interface (Figure 3, 4) includes the following elements in all three versions: textual instructions, an animation panel showing the contents of the fermentation vessels, an explanation panel describing the meanings of graphical elements, graphs and histograms showing the amounts of ingredients in the product, an adjustable thermometer, buttons for controlling the processes, an "Assessment" button for providing immediate feedback, and a slider for controlling the simulation speed. Instructional texts are depicted on individual screens, and the learner can return to previous screens. The tutorial presents 10 screens; the linear part, 24 screens; and the error part, 33 screens. There are two types of instructions: process instructions that describe the beer brewing process, and tutorial instructions that tell the learner what to do next. During the task-solving part, process instructions from the linear part are available to the learner (but tutorial instructions are not displayed). The numbers of words presented in the simulation versions are follows: in the NF version, 6,138 words; NC version, 6,750; and G version, 6,865. The simulation can be controlled by several means, depending on the production phase (primarily via control buttons and adjustment of temperature). Several key ingredients are shown in the product, e.g., enzymes, starch and yeast (depending on the brewing phase). The amounts of ingredients can be monitored using graphs and histograms. When the simulation is running, this information is constantly updated, and the content of the vessel is animated.

Because the learner is awarded points in the G version, the simulation interface (in the G version only) also features a panel depicting the number of points. Eventually, the learner can sell beer produced in the G version, and production of this beer costs them virtual money. Therefore, the G version's simulation interface features additional information regarding the prices of ingredients and energy consumption (these data can be regarded as extraneous details, but they are necessary for the gamification purposes; see Figure 4).

--- Insert Figure 3 around here ---

--- Insert Figure 4 around here ---

In the NC and G versions, the on-screen instructions are written in a conversational style. The NF version is written in a formal style (see Table 3 for an example and Brom et al., 2014, for details on how the conversational style was produced). The instructional content is the same in all the versions.

To justify the use of the conversational style of instructions in the NC and G versions, these two versions also include the following background story (taken from Brom et al., 2014), as explained by the administrator at the start of the lesson:

Imagine you are from a family that owns a family brewery from Baroque times. After the Second World War, your grandpa was trained to become a brew master. In the fifties, the communists confiscated your family brewery, but it was returned to your family after the Velvet Revolution in the nineties. Afterward, your grandpa ran the brewery for approximately 20 years, but he is now 85 years old and is looking for his successor. You are one of the people he has chosen to perhaps take on this role. This does not mean the brewery is yours, but it could be. However, your grandpa is a cautious man. He commissioned the development of a simulation modeling your family brewery. Now, he will let his chosen few work with it as best they can. Only then will he allow the very best candidate to be trained at the real brewery and possibly succeed him. Your grandpa will speak to you via text instructions for the duration of the simulation. Everything written in the instructions is what your grandpa would say.

The instructions are then written as if the grandpa character is speaking to the learner.

The G version presents the following addendum to the background story:

Based on how well you accomplish your tasks [given to you during the tutorial, the linear part or the error part], the grandpa character will assign you points. These points will be converted to money in the final, task-solving part. In this last part, your overarching goal will be to brew beers to sell them and earn money to buy a new fermentation vessel.

During the tutorial, the linear part and the error part, learners are given certain small tasks (e.g., to add the correct amount of malt to the fermentation vessel). There are 19 tasks in total (in each version). In the NF and NC versions, learners automatically advance after they solve a task. Correct information is stated in the next instructional text when relevant (e.g., in Instruction #6 from Table 3). In the G version, learners are also praised by the grandpa character and awarded points if the task has been completed correctly (see Table 3, Instruction #6). The praise typically contains learning-relevant information and thus can be viewed as a form of additional immediate feedback.

When a beer of a specific type is brewed in the task-solving part, learners simply bottle the product at the end by clicking on the "bottle" button. Afterward, learners receive a final assessment concerning the beer's quality. In the NC and NF versions, they can then restart the simulation. In the G version, they can click on the "sell beer" button and sell the product (sales depend on the beer's quality).

--- Insert Table 3 around here ---

Procedure

All sessions were organized in the morning. The participants were tested in groups of 1-5 people, each sitting at a separate computer in a laboratory (with a screen at least a 17" wide). Each computer had two blank A4 sheets of paper and a pen in front of it. All the participants in a given group received the same simulation version.

The experimental schedule is depicted in Figure 5. After the introduction, the participants were given the background questionnaire and the questionnaire on perceived prior knowledge. Next, the background narrative for the NC and G groups and the addendum regarding the goal for the G group (see previous section) were stated. In the control conditions, the treatment was strictly referred to as a "simulation", whereas it was called a "game" in the G condition.

Thereafter, the simulation started. After the tutorial part, the learners received the first in situ questionnaire and then immediately continued with the linear part of the simulation. After each of the subsequent simulation phases ended, the participants filled in the next in situ questionnaire (i.e., the 2nd, 3rd, or 4th) and were offered a short break. During the lesson, the participants could take notes on the sheets of paper they received, if they so wished. They knew that their notes would be taken away for the final knowledge tests, but they could still use them during the task-solving part.

--- Insert Figure 5 around here ---

In the task-solving part, the tasks were assigned one after another by the administrator in the control groups. The last task was assigned during the 29th minute at the latest, after the

first task had started. The tasks were administered to the control groups in the same order. They increased in difficulty and were as follows:

1. Please brew 13-degree beer in the simulation environment.

2. Please brew 10-degree beer that contains 5-6 % sugar (i.e., more than it normally should).

3. Please brew 11-degree beer that is spoiled (i.e., contains acetone).

4. Please brew a drinkable 10-degree beer in less than 50 days.

No participant completed all four tasks in less than 30 minutes.

We strived for a similar hands-on experience in the gamified condition. To this end, we set the game goal (i.e., to earn a specific sum of money for a new fermentation vessel) so that it could be achieved by completing approximately 2 tasks, which was the average from the previous experiment (Brom et al., 2014) with the same task-assignment protocol. The learners could choose which beer to brew in the G condition. However, a wholesale price was set for every beer type, and one beer always had a better retail price due to "current market preferences", making its production more attractive for the learner. The ranking of beer types by attractiveness was the same as the order in which the participants in the control groups were assigned the tasks. In this way, the G group participants were prompted to brew beers in the same order as the control group participants did (many in the G group did indeed follow this order). The G condition participants were not allowed to start a new task after 29 minutes had passed, even if they did not achieve the game goal.

We note that the game goal forces the learners to calculate their profit from the brewed beers, which is irrelevant for the learning goal and could cause unnecessary distraction. To avoid this distraction, we gave the G group participants a list with the average

production costs of all typical beer types that can be brewed in the simulation (i.e., seven) when the task-solving part started. This way, the participants needed only to subtract the production cost from the wholesale price to calculate their profit.

After the last break, we announced that the learning phase had ended. The participants were given the post hoc questionnaire and were administered the retention and transfer tests. When taking the retention test, the participants could return to previous questions, unlike in the case of the transfer test.

At the end, the participants were quickly interviewed and thanked. Approximately one month after the experiment, the participants arrived for the delayed testing session (usually held in the morning). Alternative versions of the tests were used (differing from the immediate testing session). The order in which these tests were administered was counterbalanced across the participants. Afterward, they completed a battery of psychological tests. The participants were informed in advance that the experiment would consist of two parts, but they were not informed of the content of the second part.

Measures

The paper materials consisted of a background questionnaire, four in situ questionnaires, a post hoc questionnaire, a test of graphing in science and a retention and transfer test. All the questionnaires and tests were of the pen-and-paper type. The questions are presented in the Supplementary Materials.

Control variables and demographic data. The background questionnaire asked the participants about their age, gender, area of study, native language and possible vision difficulties. To check whether the groups were balanced with respect to several variables known to correlate with learning outcomes pertaining to acquisition of mental models (see

Brom et al., 2014; Brom & Děchtěrenko, 2015), the background questionnaire also included one question on self-assessed knowledge of mathematics (6-point Likert item) and ICT skills (6-point Likert item), one question on the frequency of playing videogames (4-point ordinal item), one question on the frequency of playing live action experiential/simulation games (5-point ordinal item) and one question on self-assessed ability of acquiring mental models (7-point Likert item). For the same reasons, two questions on prior tiredness were included (7-point Likert item). An average of these two questions is treated as an "energy" variable. Likewise, the questionnaire asked one question regarding prior attitude (7-point Likert item).

After the tutorial part, we measured initial anxiety (as part of the first in situ questionnaire). To assess this variable, we used three 7-point Likert-type questions from the Questionnaire on Current Motivation (Rheinberg et al., 2001; e.g., "When I think about the task, I feel somewhat concerned"; $\alpha = .81$). We did not use the full questionnaire's version due to time constraints. We also note that this variable was intentionally measured after the tutorial part, i.e., after the participants had become familiar with controlling the simulation.

Because the simulation relayed information partly via graphs, it was important to determine whether the groups were balanced with respect to graphing skills. To this end, we administered a shortened version of the test of graphing in science (McKenzie and Padilla, 1986) with 9 questions, each of which could be awarded 0 or 1 point ($\alpha = .76$). The test was given in the delayed testing sessions, i.e., a month after the intervention.

Self-assessed prior knowledge. To avoid cuing the participants on what they should remember, we measured perceived prior domain knowledge rather than administering a full pre-test. From a pilot experiment, we also knew that administering full tests to naïve participants (which lasts approximately 30 minutes) is frustrating for them. This could influence their attitude toward the experiment and thereby the entire learning process. The

participants therefore self-assessed their prior domain knowledge as follows: 1) They were asked to check if any of six conditions apply (e.g., "My relatives, or I personally, brew beer" and "I know what *Saccharomyces cerevisiae* is"). 2) The participants were asked to write whether they had ever tried to learn about beer brewing (an open-ended question). 3) They were asked to assess their knowledge of why and when alcohol is created during the beer brewing process (4-point ordinal item). 4) They were asked to what extent they can explain why a morning headache can be worse when drinking non-alcoholic beer as opposed to alcoholic beer the evening before (6-point Likert item). 5) They were asked how often they discuss the topic of beer brewing with their friends or family (6-point Likert item). 6) They were asked to describe their knowledge of beer brewing, wine-making and whiskey production (6-point Likert item). The score range on the self-assessment was 0-32 ($\alpha = .68$). The scoring is detailed in Brom et al. (2014).

Intrinsic motivation variables. Intrinsic motivation as a current state is typically measured using interest/enjoyment questionnaires (e.g., McAuley, Dunan, & Tammen, 1989; cf. Ryan & Deci, 2000). Intrinsic motivation has been empirically linked to positive-activating affective-motivational states (e.g., Heidig, Müller, & Reichelt, 2015; Peng, Lin, Pfeiffer, & Winn, 2012; Plass et al., 2014; Sabourin & Lester, 2014). These states include situational interest, flow, learning involvement, generalized positive affect, and enjoyment. Although these states are often highly correlated, there is preliminary evidence that they may be differentially related to increased cognitive engagement and learning outcomes (Brom et al., 2017). Therefore, in this study, we measured the abovementioned five states rather than the umbrella construct of intrinsic motivation.

Situational interest is usually viewed as a temporary state of concentration and enjoyment caused by the features of a specific situation (Hidi and Renninger, 2006; Schiefele, 1999, p. 263). Thus, immediately after the tutorial, we measured whether this state

was triggered by the learning task (i.e., along with initial anxiety, in the first in situ questionnaire). We call this variable *initial interest*. We measured it using five 7-point Likert items from the Questionnaire on Current Motivation (Rheinberg et al., 2001; e.g., "Today's topic seems very interesting to me", "I am eager to see how I will perform on today's task"; α = .82).

Generalized positive affect encompasses various positively valenced activating feelings, for instance, feelings of excitation, activity, attentiveness, or enthusiasm (Watson & Tellegen, 1985). We measured this variable using a validated instrument, the PANAS (Positive and Negative Affect Schedule; Watson et al., 1988). This questionnaire consists of two 10-item mood scales: one for positive and the other for negative affect (5-point Likert items). The PANAS was administered twice during the intervention (as part of the in situ questionnaires; see Figure 5) with instructions "to assess one's feelings during the latest part of the simulation [the list of 20 feelings]" (α = .87, .88).

Flow is typically defined as pleasant absorption during an activity that one takes part in (Csikszentmihalyi, 1975), including increased attention to and concentration on the object of the activity. We measured flow twice during the treatment by administering a validated instrument, the Flow Short Scale (Rheinberg et al., 2003) (ten 7-point Likert items; e.g., "I do not notice time passing", "I feel I have everything under control", "I am completely lost in thought"; α = .93, .90).

Learning involvement is also related to task concentration but also involves positive feelings derived from learning and perceived competence. This variable was assessed three times during the treatment using a questionnaire of our design with eight 7-point Likert items (e.g., "So far, I have enjoyed brewing beer", "I always knew what to do next"; α = .86, .88, .81). The items were inspired by various questionnaires for assessing

motivation/interest/involvement-related constructs (e.g., Schraw et al., 1995; Isen and Reeve, 2005).

Enjoyment can be viewed as an emotional state that occurs when an activity is positively valued and is sufficiently controllable by the learner (Pekrun, 2006; p. 323). This variable was assessed in the post hoc questionnaire using two 6-point Likert items ("I enjoyed doing this activity", "I would describe this activity as very interesting").

Knowledge tests. We used the same retention and transfer tests (cf. Mayer, 2009) as in our previous study (Brom et al., 2014). These tests were carefully piloted before that study started (Brom et al., 2014). Each test consisted of two complementary versions (one for administering after the learning experience ended, one for delayed assessment).

The *retention tests* contained 10 short-answer and multiple-choice questions (e.g., "Write down names of the four main phases of beer brewing in the correct order, as you learned today.") and one open-ended question ("Please explain what happens during the fermentation phase and what main products are created during this phase. Imagine you are writing a short encyclopedia entry for beginners."). The test was scored using a precise pre-made key. The score range was 0-31 (Immediate: $\alpha = .66$, Delayed: $\alpha = .73$).

The *transfer tests* contained 6 or 8 open-ended questions, paired across the versions (one question in the shorter version was "paired" to three questions in the longer version). An example is as follows: "We got rid of bacteria during the boiling phase. However, after the conditioning, the product still contains acetone (which is a product of bacteria). When and how could acetone have gotten into the beer? Write down **every possibility** you can imagine" (emphasis in the original). Each question was typed on a separate A4 sheet of paper. The score ranges were 0-17 (18). Cronbach's α were .80 and .63 for the immediate tests and .82 and .75 for the delayed tests. Our transfer tests have an established rating system based on

analysis of key "idea units" in the answers. This system generates substantial agreement between independent raters. Nevertheless, for a check in this study, one rater scored all the tests, and a second rater scored the tests of 22 randomly selected participants (~20 % of the total recruited sample). The agreement for each question was in the range of r = .87-1.00, which we consider good. The scores from the first rater were used in the analysis.

Manipulation check variables. In the gamified condition only, the participants also answered, during the 2nd and 3rd administrations of the in situ questionnaire, two manipulation check questions pertaining to the presence of added game elements: "For you personally, would it be better if the grandpa praised you: *much more often – much less often?*" (5-point Likert item); and "For you personally, how important was it that the grandpa awarded you points?". The latter question had two sub-questions, regarding importance and valence (7-point Likert items). These questions were replaced by the following two questions in the 4th administration: "For you personally, how important was it that money was part of the game?" and "For you personally, how important was it that you had to achieve a game goal?". Both of these questions had two sub-questions, regarding importance and valence (7-point Likert items).

Supplementary variables. In the 2nd – 4th in situ questionnaires, the participants also rated currently perceived difficulty using one 7-point Likert item. In the post hoc questionnaire, one question addressed overall perceived difficulty. Scores from these four questions were averaged for subsequent analysis (α = .70). The feedback questionnaire also yielded data on perceived learning (two 6-point Likert items). Finally, although negative

affect is not central to this study, we report negative affect data from the two administrations of the PANAS for the sake of completeness (α = .84, .88).[6]

Data Analysis

The analysis was conducted using the statistical program R (R Core Team, 2016). Effect sizes were expressed either using η_p^2 (analysis of (co)variance) and classified into small ($\eta_p^2 \sim 0.01$), medium ($\eta_p^2 \sim 0.06$) and large ($\eta_p^2 \sim 0.14$) effect sizes or using Cohen's *d* (t-test) and classified into small ($d \sim 0.2$), medium ($d \sim 0.5$) and large ($d \sim 0.8$) effect sizes (Cohen, 1988). The flow data were converted to T-norms provided with the standardized Flow Short Scale (Rheinberg, 2004) (final scale of 21-74). The average value from the converted flow questionnaires was used. Similarly, we used averages of the generalized positive and negative affect measures and of the three learning involvement and four perceived difficulty measures.

Results

Are the Groups Balanced?

We first tested whether the groups were balanced with respect to age, energy, prior attitude towards the experiment, frequency of playing videogames and experiential games, self-assessed knowledge of mathematics and ICT, self-assessed ability of acquiring mental models, self-assessed prior knowledge, initial anxiety, graphing skills and time spent on intervention (Table 4). The groups were balanced with respect to all these variables except

[6] Although this experiment was not part of a larger study, we also used it to pilot several questions irrelevant to the present purpose.

for age. This variable correlates with several dependent variables (Table 5), and we thus included it as a covariate in the main analysis.

--- Insert Table 4 around here ---

--- Insert Table 5 around here ---

Is Gamification Noticed by Participants?

Second, we tested whether the participants in the G condition reported that they noticed the gamification elements. These tests serve as manipulation checks. The data (Table 6) indicate that, whereas a few participants were probably oblivious to the gamification, the game design elements had some relevance for the majority of them. The game goal received the best assessment, both in terms of importance and valence. The virtual money and points were also perceived positively. The participants reported that they were satisfied with the amount of praise given.

--- Insert Table 6 around here ---

Are There any Effects of Gamification?

Descriptive data are included in Table 7. Because the five intrinsic motivation variables (i.e., initial interest, learning involvement, positive affect, flow, enjoyment) and

four test score variables (i.e., retention immediate and delayed, and transfer immediate and delayed) were inter-correlated (Table 5), we analyzed the effects of gamification on these variables using MANCOVAs with age as a covariate. There were no differences between the groups (Pillai test; intrinsic motivation variables: $F(10, 172) = 0.98$; $p = .460$; learning outcomes: $F(4, 360) = 0.97$; $p = .425$[7]). Thus, the answer to the first and the second research questions is that the gamification had no detectable effects on intrinsic motivation or learning outcomes.

We analyzed between-group differences for the remaining dependent variables using ANCOVAs with age as a covariate. A significant difference was detected concerning perceived difficulty ($F(2, 88) = 4.76$, $p = .011$; $\eta_p^2 = .10$; 95 % CI [.02, .20]), but no significant differences were detected concerning negative affect ($F(2, 92) = 0.39$, $p = .676$; $\eta_p^2 = .01$; 95 % CI [.00, .04]) and perceived learning ($F(2, 92) = 0.52$, $p = .597$; $\eta_p^2 = .01$; 95 % CI [.00, .05]). Perceived difficulty varied: the post hoc tests showed that the experience was perceived as easier by those in the gamified group than by those in the non-gamified formal group (Tukey's test; $t(60) = 2.9$; $p = .013$; $d = 0.74$; 95 % CI [0.21, 1.27]). The difference also tended in the same direction concerning the non-gamified conversational group and spanned a moderate range, though the difference was not significant ($t(61) = 1.64$; $p = 0.234$; $d = 0.42$; 95 % CI [-0.10, 0.93]).[8] Thus, the answer to the fourth research question

[7]Assumptions for multivariate normality, homogeneity of variance, and homogeneity of variance-covariance matrices were met for a MANCOVA involving test scores. Assumptions for the intrinsic motivation MANCOVA were also met, except for the normality assumption. However, MANCOVAs are robust with respect to deviances from a normal distribution (Olson, 1974). We also tested between-group differences for individual variables using (a) an ANCOVA with age as a covariate (on the whole sample), (b) an ANCOVA with age as a covariate (with outliers removed), (c) a non-parametric Kruskal-Wallis test, and (d) a factor from exploratory factor analysis of five intrinsic motivation variables (all variables loaded on the same factor; loadings > .67). No significant difference was detected using any of the tests (all *p*s > .159).

[8] Test assumptions of homogeneity were met (Bartlett's test: *p*s > .140). Test assumptions of normality were violated for perceived learning and negative affect (*p*s < .001). ANCOVAs are robust with respect to the normality violation (Levy, 1980), but the results can be biased if normality is violated due to the presence of outliers (see Stevens, 2012). The results did not change when the analysis was re-run with outliers removed (perceived difficulty: $p = .018$; other variables: *p*s > .724) and when the data were analyzed using a Kruskal-Wallis test (perceived difficulty: $p = .019$; other variables: *p*s > .210).

is that the gamification had no detectable effect on negative affect and perceived learning and a moderate to high effect on perceived difficulty.

Even at a purely descriptive level, no consistent advantages (or disadvantages) were conferred by the gamified simulation version.[9] Apparently, the effects of the gamification were very limited. Therefore, we did not proceed with the analysis of whether the effects of the gamification on learning outcomes were mediated by intrinsic motivation factors.

We also explored whether the participants' areas of study (computer science vs. others, i.e., primarily social sciences) was a moderator. Two-way ANCOVAs with age as a covariate and participants' background as a factor showed that across the whole sample, the computer science students consistently outperformed the other students in all the tests (ps < .002): they were more interested after the tutorial (p = .003); their induced positive affect, flow, enjoyment and learning involvement were higher (p < .035); and they perceived the simulation to be easier (p = .001). Only for negative affect (p = .263) and perceived learning (p = .212), was no significant difference between the two groups of learners detected. However, condition × study background interaction was not significant for any of the dependent variables (ps > .118). Descriptively, there appears to be some interactions, the most salient one being for negative affect: induced negative affect was lower among the computer science students but only in the non-gamified conditions (see Table 7). These apparent non-significant interactions are either noise or missed true effects due to the relatively small number of participants per cell (~16).

[9] Also, because there were no significant between-group differences between the two control conditions in any of the dependent variables (all ps > .103 for two-sample t-tests), one can analyze differences between the G condition and the combined control conditions (we are aware of the unequal sample sizes in these tests). Again, there was no between-group difference (t-test: ps > .147; Wilcoxon signed-rank test: ps > .114) except for a medium to large effect on the perceived difficulty ($t(92) = -3.11$; p = .003; d = -0.67, 95 % CI [-1.12, -0.23]; W = 625; p = .009).

--- Insert Table 7 around here ---

Discussion and Conclusion

This study sought to explore the net effects of an added game goal, increased freedom of choice, points, virtual currency and praise (all combined) on intrinsic motivation factors, learning outcomes, and subjective evaluation among college learners studying a complex process in a computerized simulation. In doing so, we investigated edu-gamification in a new context. Neither beneficial nor detrimental effects of these gamification elements on retention and transfer test scores, initial interest, induced positive affect, flow, learning involvement and enjoyment were revealed. The only significant difference between the gamified simulation and the non-gamified simulation versions was in perceived difficulty. The gamified simulation was perceived to be easier than the non-gamified versions. For all the dependent variables, the participants' background (i.e., students of computer science vs. social sciences and art) was not found to be a significant moderator. Of the game elements used in this study, the participants perceived the clear game-like goal most positively.

Theoretical Perspectives

In this paper, we present an integrated view of how self-determination theory (SDT) (Deci and Ryan, 1985), cognitive-affective theory of learning from media (CATLM) (Moreno, 2005), and cognitive load theory (Sweller et al., 2011) together predict how several key game elements influence learners' intrinsic motivation, unnecessary cognitive load, and learning outcomes in edu-gamification contexts (Table 1). These theories imply that specific game elements or combinations thereof may be advantageous to learning under appropriate

circumstances, but enhanced learning cannot generally be expected, as demonstrated by the present findings.

On a general level, the reasons for the present null results are most likely as follows. Some game elements might not increase (or might even decrease) intrinsic motivation in the first place, as indicated by the overall lack of between-group differences in intrinsic motivation variables. If intrinsic motivation had increased for a particular learner, the positive effect of motivation on enhanced cognitive processing (implied by the CATLM) could be countered by increased distraction (predicted by cognitive load theory). The net impact of these effects on learning outcomes appears to be neutral.

On the level of individual elements, the potential positive effects of a clear goal and increased choice (implied by SDT because of support for the need for competence and autonomy) could be countered by increased distraction due to the inevitable presence of extraneous details in the user interface of the gamified simulation. Also, the positive effects of choice in our gamified version might have been quite subtle, because the learners chose between solving tasks in the task-solving part. Solving tasks offered plenty of choices by itself (in all three comparison conditions) and thus also increased the learners' feelings of autonomy.

The points were generally rated mildly positive and thus they were probably not perceived by learners as controlling "expected tangible rewards" (which tend to undermine intrinsic motivation, according to SDT) but rather as informative feedback. However, the learners could obtain feedback whenever they needed it by clicking on the "Assessment" button. The positive effect of points, such as that of choice, was probably thus too subtle or countered by increased distraction. Distraction was most likely caused by the presence of user

interface elements showing the number of achieved points and/or the learner's irrelevant thoughts related to the points.

The learners also appreciated that there was an "appropriate" amount of praise, which means that the praise also probably did not undermine intrinsic motivation (e.g., it could have been viewed by learners as childish). However, even though the praise contained learning-related information, it might have brought little added value compared to the control versions of the instructions. Its effect was again probably too subtle or was countered by increased distraction.

Why was the gamified simulation perceived to be easier? One possibility is that the presence of game elements "seduced" the learners to believe so. For example, in the context of emotional design in multimedia learning, certain "motivationally enhanced" manipulations also appear to lower ratings of difficulty. These include changing the style of instructional texts from formal to conversational (see Ginns et al., 2013) or adding facial anthropomorphisms to non-human graphical elements (e.g., Plass et al., 2014). In human-computer interaction research, Tractinsky, Katz, and Ikar (2000) reported strong correlations between a system's perceived aesthetics and perceived ease of use. Examining this issue in future research would be useful.

Practical and Methodological Implications

On the practical level, when considering the results of previous studies, our findings support the emerging notion that one should be cautious when gamifying an educational experience. One should carefully think about which approach may work for the target audience and consider the pros and cons regarding particular students. For instance, de-Marcos and colleagues (2016) showed that when a university course was augmented with

game-based elements *and* a social platform, it was beneficial for the target audience. Sandberg and colleagues (2014) showed that combining a gamification approach with adaptive difficulty was beneficial for children in a vocabulary learning application. Apparently, certain combinations of game-based elements, possibly with some non-game-based elements, can work in specific contexts. One should be careful when attempting to generalize the results of a study to different audiences: for instance, one may expect different effects among different age groups (e.g., primary school children vs. adolescents vs. university learners).

All of this also helps to highlight three methodological implications. First, gamification studies without control groups tend to report "promising" findings (Dicheva et al., 2015), while comparative studies (this one, studies reviewed under the heading *Edu-gamification Studies with Comparative Design*, and some studies discussed in the review by Ortiz-Rojas and colleagues, 2017) deliver predominantly mixed/null results. Thus, either interpretations of results of the former types of studies tends to be positively biased, or these studies use better gamification approaches than do studies with control groups. The implication is that studies with control groups that would use ecologically valid interventions should start to dominate the research field. Second, it seems so far that large effect sizes cannot be automatically expected. Researchers should thus consider using larger samples and/or more intensive modifications of learning experiences. Third, researchers should also start considering participant characteristics and educational contexts as moderating variables. Participant age is not the only variable that may influence outcomes (cf. Buckley et al., 2016). For instance, it seems reasonable that competitiveness (as a stable trait of participants; Houston et al., 2002; Harris and Houston, 2010) can moderate the effects of competition-based gamification approaches on both learning and affective-motivational outcomes (cf. Brom et al., 2016). Another possible moderator, especially when considering digital learning

materials, can be attitudes toward ICT (cf., e.g., Curtois et al., 2014) or students' gaming frequency (cf. Davis, Sridharan, Koepke, Singh, & Boiko, 2018). This study clearly showed that the computer science students performed better and liked the learning experience more than did the other students (who might have less positive attitudes toward ICT). No significant study background × condition interaction was revealed, but this lack may be due to the small sample size. For much-needed research on moderation effects, larger samples are also essential. When more results are available, the findings may eventually uncover which gamification approaches (and when and for whom) are beneficial or detrimental in educational contexts.

Limitations

This study is not without limitations. First, as already suggested, the study might have missed a true effect due to its small sample size (which is, however, quite typical for an experimental, multimedia learning study, i.e., ~30 per condition; cf., e.g., Fraenkel et al., 2012, p. 103). Second, we cannot exclude the possibility that some of our elements could have been implemented in a way that would have been more beneficial for learners. Third, we did not measure certain variables supposed to be influenced by the game elements, such as perceived choice, perceived autonomy/competence, and cognitive load. We measured only theoretically posited consequents (i.e., intrinsic motivation and learning outcomes; see Figure 1). Cognitive load/engagement is notoriously problematic to measure (e.g., Brünken, Plass, & Leutner, 2003; Brünken, Seufert, & Paas, 2010; de Jong, 2010). The other variables were not measured because of the already too long questionnaires, but assessing them at the cost of having fewer intrinsic motivation proxies is an option to consider in future studies. Fourth, because we investigated the combined effects of several game-based elements, we might have missed an individual element's effect due to a countering effect of another element.

Methodological "purity" advises that we investigate the effects of individual elements. However, there is a trade-off between doing so and ecological validity. Looking at single game-based elements (e.g., in our case, assigning points without converting them to virtual currency at the end and without stating a clear game-like goal) may reduce the believability of the intervention in the case of interconnected elements. This approach can artificially undermine intrinsic motivation. As a partial remedy, we asked the participants about their opinions regarding the individual game-based elements. This practice is one we recommend. At the same time, too many game-based elements should not be combined in a single manipulation. We recommend focusing on "minimal sets" of these elements such that removing a single element would artificially reduce the believability of the intervention.

In our opinion, these limitations do not undermine this study's key finding: that neither beneficial nor detrimental effects of certain game-based elements on learning outcomes and intrinsic motivation factors were revealed.

Concluding Remarks

The null results gained in this study generally corroborate earlier findings from comparative edu-gamification studies that the alleged benefits of gamification in education are disputable and that one should be careful when gamifying educational methods or materials. This interim conclusion should be treated with caution because comparative studies of edu-gamification approaches have thus far been limited, and each of them used a somewhat different setting and/or gamification approach. Thus, their results may not be directly comparable, let alone broadly generalizable. The question of how to gamify education and when (if at all) remains open. The theories introduced in our work (under the

heading *Theoretical Standpoints*) and their implications can assist researchers and designers in selecting the most promising game elements to be researched and eventually used.

References

Adams, D. M., Mayer, R. E., MacNamara, A., Koenig, A., & Wainess, R. (2012). Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of educational psychology, 104*(1), 235-249.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with Massive Online Courses *Proceedings of the 23rd International Conference on World Wide Web* (pp. 687-698): ACM.

Brom, C., Bromová, E., Děchtěrenko, F., Buchtová, M., & Pergel, M. (2014). Personalized messages in a brewery educational simulation: Is the personalization principle less robust than previously thought?. *Computers & Education, 72*, 339-366.

Brom, C., Buchtová, M., Šisler, V., Děchtěrenko, F., Palme, R., & Glenk, L. M. (2014). Flow, social interaction anxiety and salivary cortisol responses in serious games: A quasi-experimental study. *Computers & Education, 79*, 69-100.

Brom, C., Děchtěrenko, F. (2015). Mathematical Self-efficacy as a Determinant of Successful Learning of Mental Models from Computerized Materials. In *ECGBL-9th Proceedings of European Conference on Game Based Learning* (pp. 89-97).

Barata, G., Gama, S., Jorge, J., & Gonçalves, D. (2013). Improving Participation and Learning with Gamification *Proceedings of the First International Conference on Gameful Design, Research, and Applications* (pp. 10-17): ACM.

Brom, C., Šisler, V., Slussareff, M., Selmbacherová, T., & Hlávka, Z. (2016). You like it, you learn it: affectivity and learning in competitive social role play gaming. *International Journal of Computer-Supported Collaborative Learning, 11*(3), 313-348.

Brom, C., Děchtěrenko, F., Frollová, N., Stárková, T., Bromová, E., & D'Mello, S. K. (2017). Enjoyment or involvement? Affective-motivational mediation during learning from a complex computerized simulation. *Computers & Education, 114*, 236-254.

Buckley, P., Doyle, E., & O'Mahoney, A. (2016). Individualising Gamification: Investigating how Learning Styles Impact Upon Gamification *10th European Conference on Games Based Learning: ECGBL 2016* (pp. 82-88): Academic Conferences and Publishing International.

Cameron, J., Banko, K. M., & Pierce, W. D. (2001). Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *The Behavior Analyst, 24*(1), 1-44.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning a systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79-122.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.): Hillsdale, NJ: Erlbaum.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of educational psychology, 88*(4), 715-730.

Courtois, C., Montrieux, H., De Grove, F., Raes, A., De Marez, L., & Schellens, T. (2014). Student acceptance of tablet devices in secondary education: A three-wave longitudinal cross-lagged case study. *Computers in Human behavior, 35*, 278-286. doi:10.1016/j.chb.2014.03.017

Darejeh, A., & Salim, S. S. (2016). Gamification Solutions to Enhance Software User Engagement—A Systematic Review. *International Journal of Human–Computer Interaction, 32*(8), 613-642. doi:10.1080/10447318.2016.1183330

Davis, K., Sridharan, H., Koepke, L., Singh, S., & Boiko, R. (2018). Learning and engagement in a gamified course: Investigating the effects of student characteristics. *Journal of Computer Assisted Learning*. Advance online publication. doi:10.1111/jcal.12254.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science, 38*(2), 105-134.

de-Marcos, L., Domínguez, A., Saenz-de-Navarrete, J., & Pagés, C. (2014). An empirical study comparing gamification and social networking on e-learning. *Computers & Education, 75*, 82-91.

de-Marcos, L., Garcia-Lopez, E., & Garcia-Cabot, A. (2016). On the effectiveness of game-like and social approaches in learning: Comparing educational gaming, gamification & social networking. *Computers & Education, 95*, 99-113.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin, 125*(6), 627-668.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Denny, P. (2013). The effect of virtual achievements on student engagement *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 763-772): ACM.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From Game Design Elements to Gamefulness: Defining "Gamification" *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments* (pp. 9-15): ACM.

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: a systematic mapping study. *Educational Technology & Society, 18*(3), 1-14.

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.

Fortier, M. S., Vallerand, R. J., & Guay, F. (1995). Academic motivation and school performance: Toward a structural model. *Contemporary educational psychology, 20*(3), 257-274.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.): McGraw-Hill New York.

Garner, R., Brown, R., Sanders, S., & Menke, D. J. (1992). "Seductive Details" and Learning from Text *The role of interest in learning and development* (pp. 239-254).

Gauthier, A., Corrin, M., & Jenkinson, J. (2015). Exploring the influence of game design on learning and voluntary use in an online vascular anatomy study aid. *Computers & Education, 87*, 24-34.

Ginns, P., Martin, A. J., & Marsh, H. W. (2013). Designing instructional text in a conversational style: a meta-analysis. *Educational Psychology Review, 25*(4), 445-472.

Hakulinen, L., Auvinen, T., & Korhonen, A. (2013). Empirical Study on the Effect of Achievement Badges in TRAKLA2 Online Learning Environment *Learning and Teaching in Computing and Engineering (LaTiCE), 2013* (pp. 47-54): IEEE.

Halan, S., Rossen, B., Cendan, J., & Lok, B. (2010). High score!-motivation strategies for user participation in virtual human development *International Conference on Intelligent Virtual Agents* (pp. 482-488): Springer.

Hamari, J. (2017). Do badges increase user activity? A field experiment on the effects of gamification. *Computers in Human behavior, 71*, 469-478.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does Gamification Work? - A Literature Review of Empirical Studies on Gamification *2014 47th Hawaii International Conference on System Sciences* (pp. 3025-3034): IEEE.

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education, 80*, 152-161.

Harris, P. B., & Houston, J. M. (2010). A reliability analysis of the revised competitiveness index. *Psychological reports, 106*(3), 870-874.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Heidig, S., & Clarebout, G. (2011). Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review, 6*(1), 27-54.

Heidig, S., Müller, J., & Reichelt, M. (2015). Emotional design in multimedia learning: Differentiation on relevant design features and their effects on emotions and learning. *Computers in Human Behavior*, 44, 81-95.

Houston, J., Harris, P., McIntire, S., & Francis, D. (2002). Revising the competitiveness index using factor analysis. *Psychological reports, 90*(1), 31-34.

Isen, A. M., & Reeve, J. (2005). The influence of positive affect on intrinsic and extrinsic motivation: Facilitating enjoyment of play, responsible work behavior, and self-control. *Motivation and emotion, 29*(4), 295-323.

Johnson, D. W., & Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Edina, MN: Interaction Book Company.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational researcher, 38*(5), 365-379.

Johnson-Glenberg, M. C., & Megowan-Romanowicz, C. (2017). Embodied science and mixed reality: How gesture and motion capture affect physics education. *Cognitive research: principles and implications, 2*(1), doi.org/10.1186/s41235-017-0060-9

Katz, B., Jaeggi, S., Buschkuehl, M., Stegman, A., & Shah, P. (2014). Differential effect of motivational features on training improvements in school-based cognitive training. *Frontiers in human neuroscience, 8*, 242. doi:10.3389/fnhum.2014.00242

Ke, F. (2008). Computer games application within alternative classroom goal structures: cognitive, metacognitive, and affective evaluation. *Educational Technology Research and Development, 56*(5-6), 539-556.

Landers, R. N., & Callan, R. C. (2011). Casual Social Games as Serious Games: The Psychology of Gamification in Undergraduate Education and Employee Training *Serious Games and Edutainment Applications* (pp. 399-423): Springer.

Levy, K. J. (1980). A Monte Carlo study of analysis of covariance under violations of the assumptions of normality and equal regression slopes. *Educational and Psychological Measurement, 40*(4), 835-840.

Mayer, R. E. (2009). *Multimedia Learning* (2nd ed.): Cambridge University Press.

Mayer, R. E. (2014). *Computer Games for Learning: An Evidence-Based Approach.*: The MIT Press.

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport, 60*(1), 48-58.

McGonigal, J. (2011). *Reality is Broken: Why Games Make us Better and How They Can Change the World*: Penguin.

Moreno, R. (2005). Instructional technology: Promise and pitfalls. *Technology-based education: Bringing researchers and practitioners together* (pp. 1-19): Information Age Publishing.

Morschheuser, B., Hamari, J., & Koivisto, J. (2016). Gamification in crowdsourcing: A review *Proceedings of 49th Hawaii International Conference on System Sciences* (pp. 4375-4384).

Morschheuser, B., Hamari, J., Werder, K., & Abe, J. (2017). How to gamify? A method for designing gamification *Proceedings of the 50th Hawaii International Conference on System Sciences* (pp. 1298-1307).

Ortiz-Rojas, M., Chiluiza, K., & Valcke, M. (2017). Gamification and Learning Performance: A Systematic Review of the Literature. *European Conference on Games Based Learning* (pp. 515-522): Academic Conferences International Limited.

Park, B., Flowerday, T., & Brünken, R. (2015). Cognitive and affective effects of seductive details in multimedia learning. *Computers in Human behavior, 44*, 267-278.

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: a meta-analysis of research findings. *Psychological bulletin, 134*(2), 270-300.

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic Emotions and Student Engagement. In *Handbook of Research on Student Engagement* (pp. 259-282): Springer Science+Business Media.

Peng, W., Lin, J. H., Pfeiffer, K. A., & Winn, B. (2012). Need satisfaction supportive game features as motivational determinants: An experimental study of a self-determination theory guided exergame. *Media Psychology, 15*(2), 175-196.

Plass, J. L., Heidig, S., Hayward, E. O., Homer, B. D., & Um, E. (2014). Emotional design in multimedia learning: Effects of shape and color on affect and learning. *Learning and Instruction, 29*, 128-140. doi:10.1016/j.learninstruc.2013.02.006

Plass, J. L., O'Keefe, P. A., Homer, B. D., Case, J., Hayward, E. O., Stein, M., & Perlin, K. (2013). The Impact of Individual, Competitive, and Collaborative Mathematics Game Play on Learning, Performance, and Motivation. *Journal of educational psychology, 105*(4), 1050-1066. doi:10.1037/a0032688

Qin, Z., Johnson, D. W., & Johnson, R. T. (1995). Cooperative versus competitive efforts and problem solving. *Review of Educational Research, 65*(2), 129-143.

Rheinberg, F. (2004). *Motivationsdiagnostik [Motivation diagnosis]*: Gottingen: Hogrefe

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen [QCM: A questionnaire to assess current motivation in learning situations]. *Diagnostica, 47*, 57-66.

Rheinberg, F., Vollmeyer, R., & Engeser, S. (2003). Die Erfassung des Flow-Erlebens [in German]. In J. Steinsmeier-Pelster & F. Rheinberg (Eds.), *Diagnostik von Motivation und Selbstkonzept* (pp. 261-279): Hogrefe.

R Core Team (2016). A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ (Accessed 15 August 2017).

Sabourin, J. L., & Lester, J. C. (2014). Affect and Engagement in Game-BasedLearning Environments. *IEEE Transactions on Affective Computing, 5*(1), 45-56.

Sandberg, J., Maris, M., & Hoogendoorn, P. (2014). The added value of a gaming context and intelligent adaptation for a mobile learning application for vocabulary learning. *Computers & Education, 76*, 119-130.

Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading, 3*(3), 257-279.

Schraw, G., Bruning, R., & Svoboda, C. (1995). Sources of situational interest. *Journal of Literacy Research, 27*(1), 1-17.

Stárková, T., Lukavský, J., Javora, O., & Brom, C. (submitted 13-Jul-2018) Anthropomorphisms in Multimedia Learning: An Eye-tracking Replication Study with Null Results. Submitted manuscript.

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*: Routledge.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers, 13*(2), 127-145.

van Roy, R., & Zaman, B. (2017). Why Gamification Fails in Education-And How to Make it Successful: Introducing Nine Gamification Heuristics Based on Self-Determination Theory. In M. Ma & A. Oikonomou (Eds.), *Serious Games and Edutainment Applications, Volume II* (pp. 485-509): Springer International Publishing AG.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology, 54*(6), 1063-1070.

Wilensky, U. (1999) NetLogo, Center for Connected Learning and Computer-Based Modeling, Northwestern University, available: http://ccl.northwestern.edu/netlogo/ [accessed 14.8.2017].

Wouters, P., & van Oostendorp, H. (2017). Overview of instructional techniques to facilitate learning and motivation of serious games. In *Instructional techniques to facilitate learning and motivation of serious games* (pp. 1-16): Springer.

Tables

Table 1. Effects of the game elements examined in this study according to common assumptions, cognitive load theory and self-determination theory.

Game element	Common assumptions	Cognitive load theory	Self-determination theory
Game goal	enhances motivation; increases challenge; structures learning (+)	if it evokes additional decision-making efforts or irrelevant thoughts or is connected to additional interface elements, it increases unnecessary load (–); structuring learning can reduce unnecessary load (+)	if it improves structuring of learning activities, it can support the need for competence and/or autonomy (+)
Increased choice	enhances motivation (+)	if it is accompanied by additional interface elements or evokes irrelevant thoughts, it increases unnecessary load (–)	supports the need for autonomy (+)
Points, badges	represents feedback; can structure learning; enhances motivation (+)	accompanying interface elements increase unnecessary load; competitive pressure can evoke learning-irrelevant thoughts, increasing unnecessary load (–); structuring learning can reduce unnecessary load (+)	can undermine intrinsic motivation; thwarts the need for autonomy (–); if they serve as competence feedback, they can support the need for competence (+); too few points/badges may thwart the need for competence (–)
Praise (as a type of verbal reward)	represents immediate feedback; enhances motivation (+)	if it does not provide learning-relevant information, it increases unnecessary load (–)	supports the need for competence (+)

Note: (+) sign denotes potential to enhance learning and (–) sign denotes potential to hamper learning.

Table 2. Differences between conditions.

Control 1 Non-gamified formal	Control 2 Non-gamified conversational	Experimental Gamified
instructions in a formal style	instructions in a conversational style, simple narrative	instructions in a conversational style, simple narrative game goal, increased freedom of choice praise, increased feedback points, virtual currency

Table 3. Examples of two instructions from the tutorial part.

	Instruction #6	Instruction #7
Process instruction	**Excellent! <u>This is exactly how I imagined it.</u>** Because the brewing tank holds 1000 liters of water, you$_{T}$ had to add 150 kg of malt (10 x 15 kg) to it in order to brew 10-degree beer. **<u>There is the right amount of malt in the tank, which means you paid attention and I should therefore reward you. You earn 2000 points.</u>**[a]	Now you$_{T}$ heat the product to 75 DEGREES Centigrade. This is the temperature at which enzymes BEST CONVERT starches into sugars. There are also more complex methods of brewing that allow for better-tasting beer**, but I don't use them when making beer**.
Tutorial instruction	Look$_{T}$ into the brewing tank. Starches are shown inside (blue) along with enzymes (pink) and bacteria (blue and white). For now, the brewing tank contains no sugar. Click$_{T}$ ">>" **and** you$_{T}$ **will find out what happens next**.	Set$_{T}$ the right temperature. Then, look$_{T}$ at the "Infusion" button. The button can be clicked on or off: in doing so you$_{T}$ either start$_{T}$ or stop$_{T}$ the infusion process. Now try$_{T}$ to use the button several times to either start or stop the simulation. In addition, notice$_{T}$ that the TIME INDICATOR, below the image panel, shows the time that has elapsed SINCE THE PHASE BEGAN. Let$_{T}$ the infusion run for 5 to 10 minutes and then stop$_{T}$ the simulation and click$_{T}$ ">>".

Note: Some words were highlighted in the original texts using capital letters. Additions present only in the NC and G versions are shown in boldface. Additions present only in the G version are in underlined boldface. Unlike the English language, the Czech language features two syntactic forms of second-person singular pronouns. One is more informal, and the other more formal. Respective changes are denoted by the subscript $_{T}$. See Brom et al., (2014) for details.

[a]The learner added malt in the previous step.

Table 4. Control variables (means, standard deviations in brackets) and ANOVA results.

Variable	Condition			ANOVA[a, b]			
	NF	NC	G	F	*df*	η_p^2	95 % CI η^2
Age	23.09 (2.48)	22.29 (1.87)	23.84 (3.00)	3.16*	2, 95	0.06	[0, 0.16]
Energy	3.33 (1.45)	2.85 (1.35)	2.69 (1.15)	1.97	2, 91	0.04	[0, 0.12]
Prior attitude	2.82 (1.16)	2.88 (1.15)	2.44 (0.93)	1.35	2, 91	0.03	[0, 0.09]
Freq. videogames	1.70 (0.88)	1.76 (1.07)	1.87 (0.96)	0.26	2, 95	0.01	[0, 0.02]
Freq. exp. games	2.30 (1.29)	2.32 (1.07)	2.13 (0.88)	0.30	2, 95	0.01	[0, 0.03]
Math knowledge	4.39 (1.48)	4.44 (1.19)	4.29 (1.37)	0.11	2, 95	0.00	[0, 0.01]
ICT knowledge	4.76 (1.30)	4.79 (1.12)	4.9 (1.16)	0.13	2, 95	0.00	[0, 0.01]
Mental models	5.82 (1.67)	5.82 (1.68)	5.48 (1.96)	0.38	2, 95	0.01	[0, 0.04]
Initial anxiety	9.00 (5.01)	9.53 (4.88)	9.30 (4.11)	0.11	2, 94	0.00	[0, 0.01]
Graphing skills	5.82 (2.34)	6.65 (2.14)	6.80 (2.04)	1.89	2, 94	0.04	[0, 0.13]
Time on task	126.92 (23.23)	132.40 (22.66)	130.40 (20.35)	0.52	2, 95	0.01	[0, 0.05]
Self-assessed prior knowledge	3.61 (3.47)	4.40 (2.85)	4.52 (3.99)	0.67	2, 95	0.01	[0, 0.05]

* $p < .05$

[a]Test assumptions of homogeneity were met, except for age (Bartlett's test: $p = .033$), but ANOVAs are robust to the homogeneity violation as long as groups are of roughly equal size (Bathke, 2004). This is the present case. The normality assumption was violated for all the variables (Shapiro-Wilk test: ps $< .05$). ANCOVAs are robust to the normality violation (Levy, 1980), but the results can be biased if normality is violated due to the presence of outliers (see Stevens, 2012). We detected one outlier for age (Grubb's test: $p = .013$) and re-ran the test with the outlier removed. A marginally significant difference was still detected ($p = .092$). We also re-ran all the tests using a non-parametric Kruskal-Wallis test: no significance changed (age: $p = .045$; other ps $> .156$).

[b]The confidence interval for η_p^2 was computed using a bootstrapping technique ($N = 1000$).

Table 5. Correlation matrix (Pearson's *r*). The number in brackets denotes the number of observations used for correlation.

	Age	Initial interest	Learning involvement	Positive affect	Negative affect	Flow	Enjoyment	Perceived learning	Perceived difficulty	Retention test – immed.	Retention test – delayed	Transfer test – immed.
Initial interest	-0.09 (98)	–										
Learning involvement	-0.04 (98)	0.65*** (98)	–									
Positive affect	-0.19 (98)	0.52*** (98)	0.64*** (98)	–								
Negative affect	-0.07 (98)	0.05 (98)	-0.29** (98)	0.03 (98)	–							
Flow	0.01 (95)	0.48*** (95)	0.83*** (95)	0.62*** (95)	-0.38*** (95)	–						
Enjoyment	-0.27**[a] (98)	0.56*** (98)	0.67*** (98)	0.62*** (98)	-0.12[a] (98)	0.60*** (95)	–					
Perceived learning	-0.29**[a] (98)	0.34*** (98)	0.37*** (98)	0.45*** (98)	0.05 (98)	0.25* (95)	0.42*** (98)	–				
Perceived difficulty	-0.12 (94)	-0.24* (94)	-0.41*** (94)	-0.17 (94)	0.25* (94)	-0.40*** (91)	-0.16 (94)	-0.07 (94)	–			
Retention test – immediate	0.02 (98)	0.18 (98)	0.36*** (98)	0.23* (98)	-0.19 (98)	0.31** (95)	0.10 (98)	0.10 (98)	-0.30** (94)	–		
Retention test – delayed	0.20 (97)	0.32** (97)	0.37*** (97)	0.25* (97)	-0.16 (97)	0.40*** (94)	0.11 (97)	0.13 (97)	-0.41*** (93)	0.55*** (97)	–	
Transfer test – immediate	0.13 (96)	0.08 (96)	0.26* (96)	0.18 (96)	-0.14 (96)	0.27** (94)	0.13 (96)	0.05 (96)	-0.31** (92)	0.60*** (96)	0.55*** (95)	–
Transfer test – delayed	0.33** (96)	0.22* (96)	0.31** (96)	0.18 (96)	-0.07 (96)	0.30** (93)	0.05 (96)	0.02 (96)	-0.36*** (92)	0.45*** (96)	0.67*** (96)	0.61*** (94)

* $p < .05$; ** $p < .01$; *** $p < .001$

[a]Spearman's rank correlation coefficients are similar, with the following notable exceptions: ρ (enjoyment, age) = -.18 ($p < .1$); ρ (enjoyment, negative affect) = -.20*; ρ (perceived learning, age) = -.18 ($p < .1$).

Table 6. Manipulation check variables: means, standard deviations and one-sample t-test for deviation from the midpoint concerning valence.

Variable	Scale	Neutral midpoint	Mean (SD)	t[d]	*df*
Praise	1 – 5[a]	3	3.06 (0.42)	0.68	26
Points					
- importance	1 – 7[b]	-	4.89 (1.51)	-	-
- valence	1 – 7[c]	4	5.73 (1.15)	7.97***	27
Money					
- importance	1 – 7[b]	-	4.39 (1.93)	-	-
- valence	1 – 7[c]	4	5.06 (1.12)	5.28***	30
Goal					
- importance	1 – 7[b]	-	6.23 (0.84)	-	-
- valence	1 – 7[c]	4	6.00 (0.93)	11.96***	30

*** $p < .001$

[a]1 = much less; 5 = much more

[b]1 = very small/none; 7 = very large

[c]1 = very negative; 7 = very positive

[d]Because normality assumptions were violated (ps < .003), we re-ran the tests using the Wilcoxon Signed-Rank test. No significance changed (praise: $p = .596$; other ps < .001).

Table 7. Means and SDs for the participants, split based on their area of study.

Variable	Condition								
	NF			NC			G		
	All students	Computer science	Others	All students	Computer science	Others	All students	Computer science	Others
n	33	17	16	34	17	17	31	16	15
Initial interest	26.58 (5.07)	28.00 (4.49)	25.06 (5.36)	25.68 (5.13)	27.41 (5.04)	23.94 (4.75)	27.68 (4.48)	28.50 (4.15)	26.80 (4.78)
Learning involvement	44.71 (5.89)	48.02 (4.26)	41.18 (5.37)	45.29 (6.63)	47.53 (6.87)	43.06 (5.73)	46.61 (5.53)	48.71 (3.98)	44.36 (6.17)
Positive affect	30.20 (5.88)	31.41 (5.47)	28.91 (6.20)	32.82 (7.03)	33.35 (7.21)	32.29 (7.02)	32.26 (7.49)	34.47 (7.44)	29.90 (7.03)
Negative affect	13.71 (3.76)	12.71 (2.76)	14.78 (4.43)	14.53 (5.28)	13.06 (4.56)	16.00 (5.67)	13.45 (4.22)	14.22 (5.50)	12.63 (2.10)
Flow	54.11 (8.28)	57.21 (6.56)	50.60 (8.83)	56.27 (8.28)	59.94 (8.39)	52.59 (6.53)	57.23 (7.33)	60.22 (4.64)	54.03 (8.41)
Enjoyment	4.85 (0.98)	5.06 (0.63)	4.62 (1.23)	4.88 (0.78)	5.03 (0.70)	4.74 (0.85)	4.94 (0.87)	5.09 (0.76)	4.77 (0.98)
Perceived learning	4.52 (0.89)	4.53 (1.04)	4.50 (0.73)	4.66 (0.80)	4.68 (0.90)	4.65 (0.70)	4.65 (0.89)	4.84 (0.81)	4.43 (0.94)
Perceived difficulty	3.45 (0.64)	3.21 (0.51)	3.70 (0.67)	3.30 (0.53)	3.13 (0.36)	3.47 (0.64)	2.96 (0.62)	2.74 (0.71)	3.23 (0.37)
Retention test									
- immediate	24.72 (5.11)	26.74 (3.13)	22.58 (5.98)	25.53 (4.07)	26.82 (3.42)	24.24 (4.35)	24.82 (3.43)	25.55 (2.65)	24.05 (4.06)
- delayed	17.78 (6.37)	20.24 (5.15)	15.17 (6.65)	18.62 (6.40)	21.24 (6.50)	16.00 (5.27)	19.88 (5.60)	21.59 (5.74)	17.93 (4.92)
Transfer test									
- immediate	0.27 (0.96)	0.52 (0.66)	-0.02 (1.17)	0.53 (0.82)	0.93 (0.77)	0.12 (0.67)	0.18 (0.88)	0.55 (0.88)	-0.19 (0.72)
- delayed	-0.48 (0.89)	-0.12 (0.75)	-0.89 (0.87)	-0.29 (1.01)	0.17 (1.17)	-0.74 (0.55)	-0.22 (1.09)	0.31 (0.92)	-0.84 (0.96)

Note: higher values mean “more”, including perceived difficulty (more difficult) and negative affect (a higher negative affect).

Figures

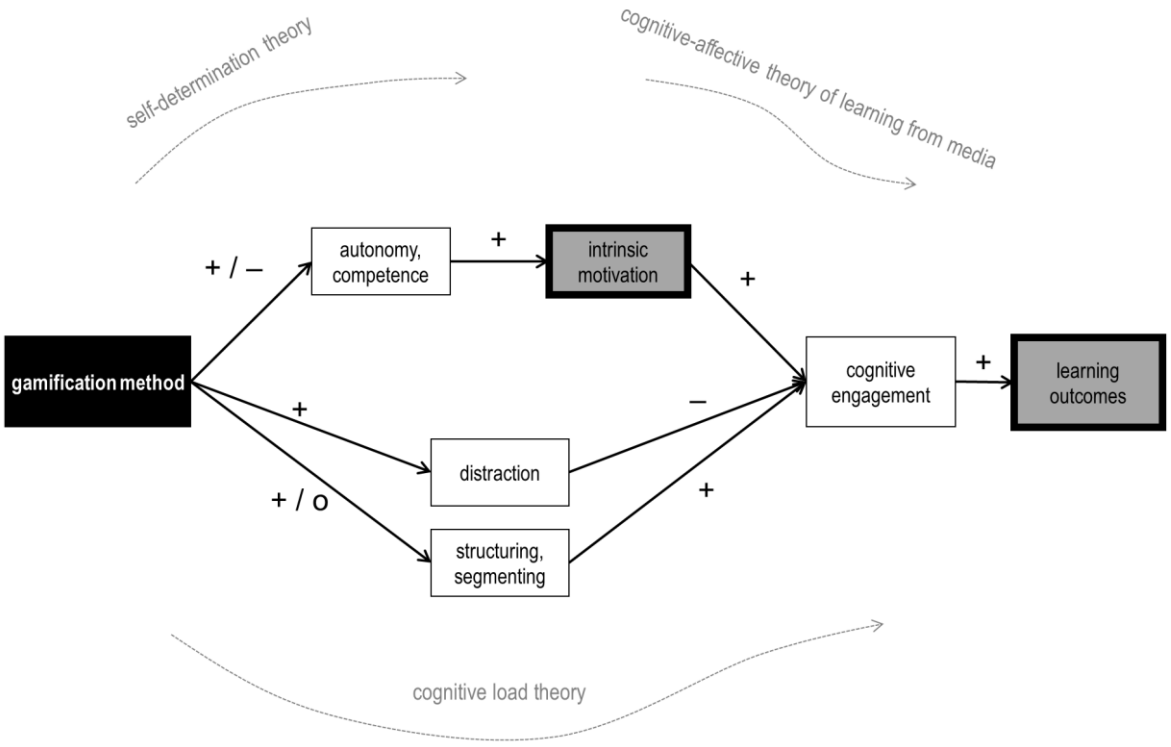

Figure 1. Theoretical predictions derived from SDT, CATLM, and cognitive load theory. Positive (+), negative (–), and no (o) influences are depicted. Measured constructs are in gray boxes.

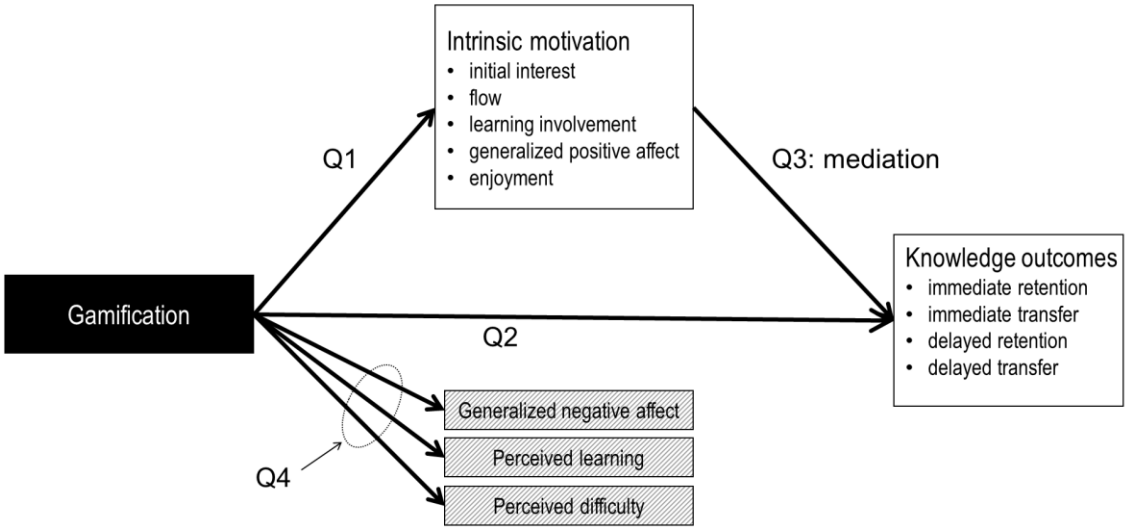

Figure 2. Manipulation (black box), main dependent (white boxes) and supplementary dependent (hatched boxes) variables used in this study. Arrows represent key investigated connections (with corresponding research questions indicated). Intrinsic motivation is measured via five proxy variables.

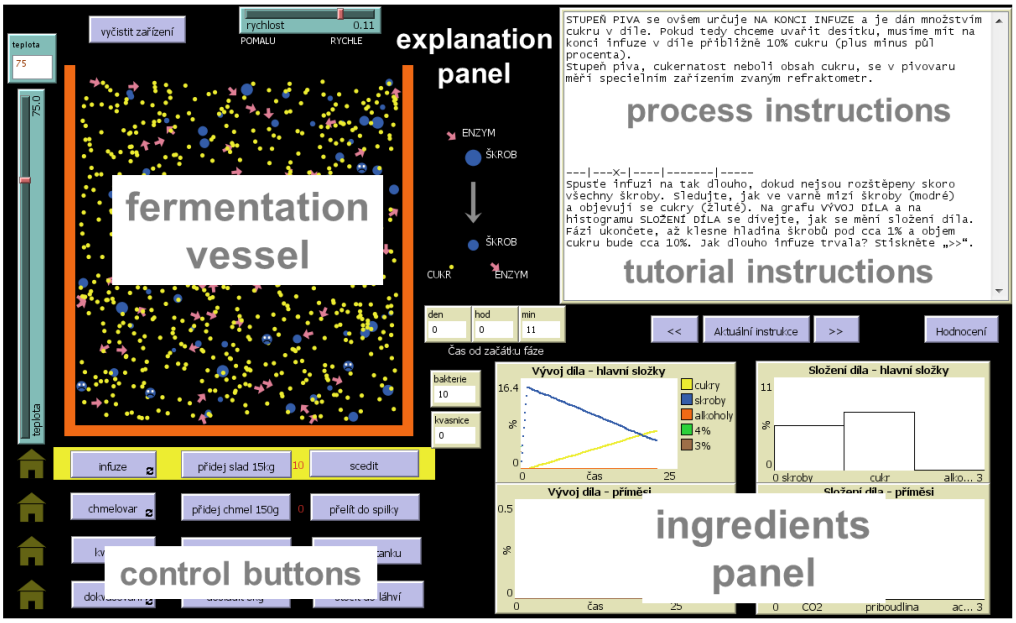

Figure 3. Simulation screenshot in the NF version. The interface is exactly the same in the NC version, except for the instructional texts (i.e., their style). The fermentation vessel, control buttons, explanation panel, instructional screens, and panels with graphs and histograms are labeled. The slider for controlling the temperature is on the far left (set to 75 degrees Celsius). The learner can scroll through the instructions using the three buttons below the instructional screens. On the right side of these buttons, there is the assessment button. The slider for controlling speed and the button for starting the entire simulation are located above the fermentation vessel. All text is in Czech and is shown for illustrative purposes only (i.e., to demonstrate the user interface's layout).

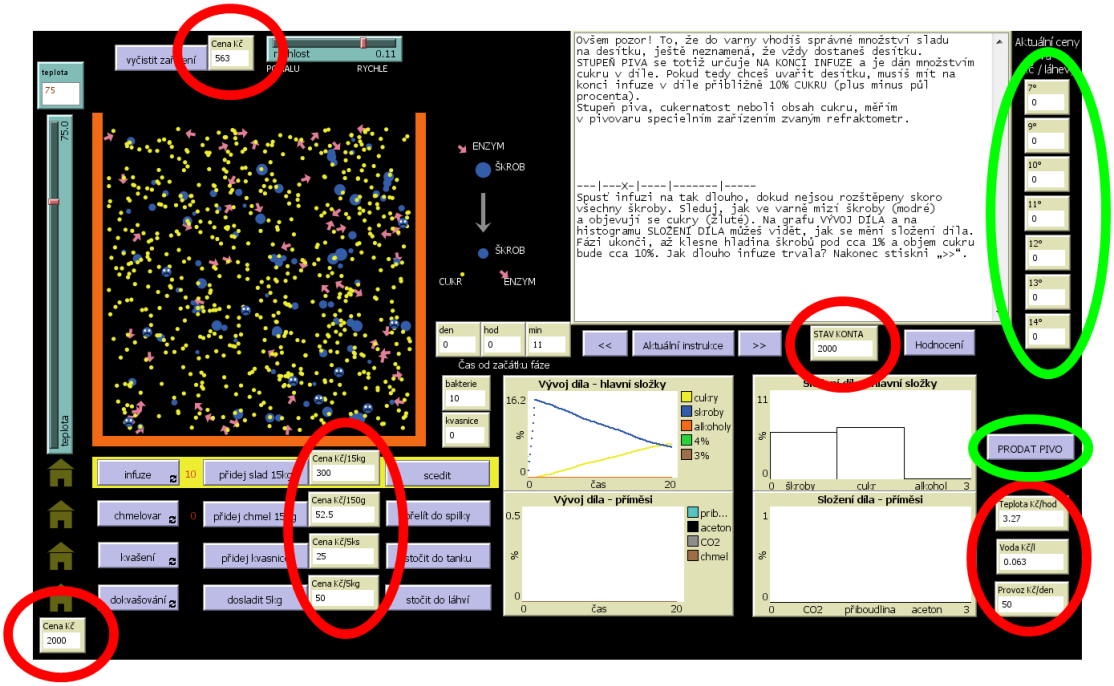

Figure 4. Simulation screenshot in the G version. The pricing of various ingredients and amounts of energy and water are noted within red ovals. The wholesale price and the button for selling beer are noted within green ovals.

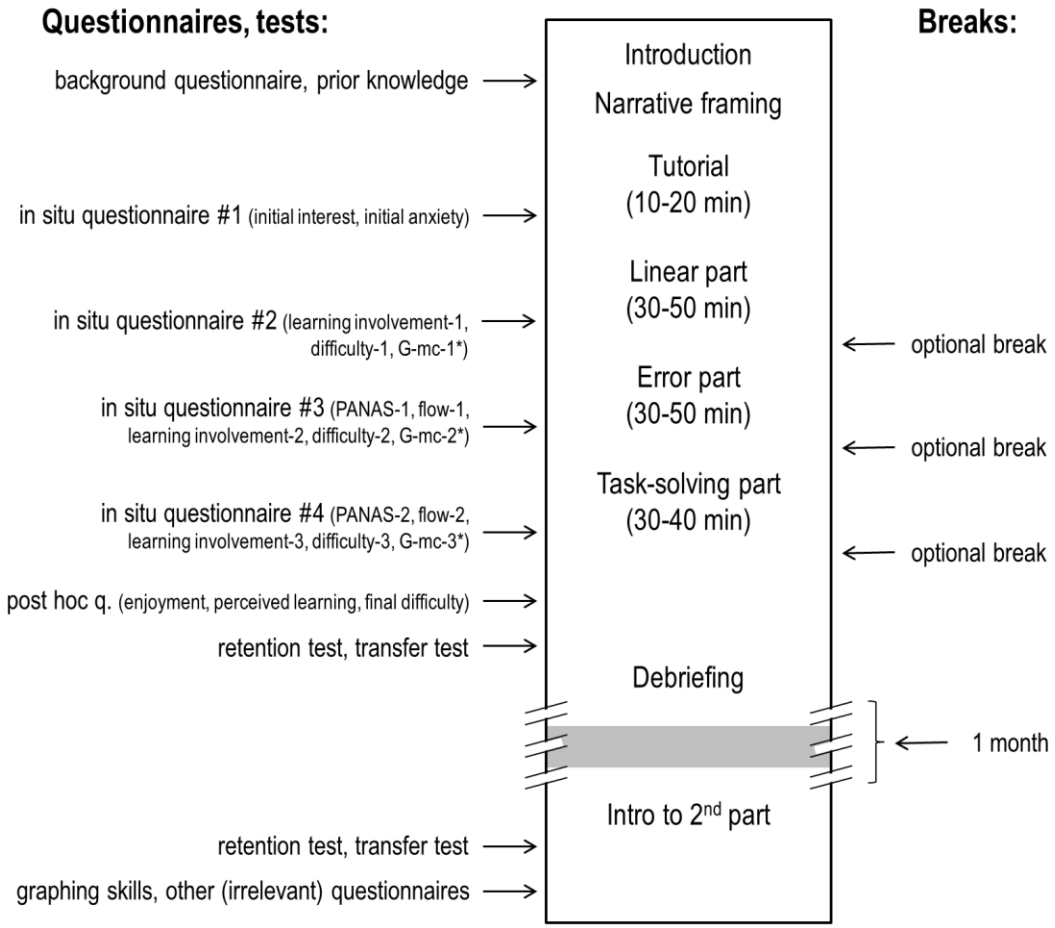

Figure 5. Schedule of the experiment. (* G-mc = Gamified manipulation check questions)

Supplementary Material

1. Constructs measured by questionnaires during the experiment

Construct	When measured	Questions	Scale
Self-assessed prior knowledge) (α = .68[a])	prior to the experiment	• Q1a: My relatives (or I personally) brew beer. • Q1b: I have taken part in an excursion to a brewery. • Q1c: We learnt about beer brewing in school. • Q1d: I know what *Saccharomyces cerevisiae* is. • Q1e: I know how *Lactobacillus* can influence beer. • Q1f: I know why malt is added to beer before yeast.	dichotomous: agree – not agree
		• Q2: Please write down whether you have ever tried to learn about the topic of beer brewing. If so, when and where?	open-ended
		• Q3: Should you be asked to explain why and when alcohol is created during the beer brewing process, would you consider yourself to be:	4-point ordinal item (*1) I don't know, so far I have had no interest in this topic; 2) beginner, I know something about the topic; 3) intermediate; 4) advanced, I know quite a lot about the topic*.)
		• Q4: Can you explain why a morning headache can be worse when you drink non-alcoholic beer rather than alcoholic beer the evening before?	6-point Likert item *(1 - definitely yes; 6 - definitely no)*
		• Q5: How often do you discuss the topic of beer brewing with your friends or family?	6-point Likert item *(1 - always; 6 – never)*
		• Q6 – 8: Check to indicate your knowledge of beer brewing [Q6] / wine-making [Q7] / whiskey production [Q8].	6-point Likert item *(1 - very good; 6 - very weak)*

Self-assessed knowledge of mathematics	prior to the experiment	• Check one of the following to indicate your knowledge of mathematics.	6-point Likert item ((*1 - very good; 6 - very weak*)
Self-assessed ICT skills	prior to the experiment	• Check one of the following to indicate your knowledge of ICT.	6-point Likert item (*1 - very good; 6 - very weak*)
Frequency of playing live action experiential /simulation games	prior to the experiment	• How often do you play experiential and/or simulation games or tabletop role-playing games (e.g. LARPs, simulations of medieval battles, outdoor puzzle hunts, AD&D, etc.)?	5-point ordinal item (*1 – never or I don't know what these terms mean; 2) once or twice so far; 3) approx. once a year; 4) more than once a year, but less than once a month; 5) at least once a month on average.*)
Self-assessed ability of acquiring mental models	prior to the experiment	• Imagine you will be examined on the history of shipping traffic in the 19th century. A week before the exam, the examiner proposes you that you can learn just one of the following two things: a) the names of British steamboats from the second half of the 19th century, including their displacement and their propeller type, or b) how these steamboats' propellers work. There are over 60 of steamboats and five functionally-distinct types of propellers. What would you prefer to learn?	7-point Likert item (*1 - I strongly prefer the names of the steamboats, including their displacement and propeller type; 7 - I strongly prefer to learn how the propellers work*)
Energy	prior to the experiment	• How alert do you feel this morning? • How do you feel overall right now?	two 7-point Likert items (*1 – very well*; *7 – very bad)*
Prior attitude	prior to the experiment	• My thoughts pertaining to this experiment are:	7-point Likert item (*1 – very positive*; *7 – very negative)*
Initial anxiety (three questions from the Questionnaire on Current Motivation; Rheinberg et al., 2001) (α = .81)	1st in situ; after tutorial	• When I think about the task, I feel somewhat concerned. • I am afraid I will make a fool of myself. • I think I won't do well at the task.	three 7-point Likert items (*1 – don't agree at all*; *7 – I completely agree)*
Graphing skills (shortened version; McKenzie & Padilla, 1986) (α = .76)	in the delayed testing session (a month after the intervention)	9 items, see McKenzie & Padilla (1986)	nine multiple choice items
Initial interest (five questions from the	1st in situ; after	• Today's topic seems very interesting to me.	five 7-point Likert items (*1 – don't*

Questionnaire on Current Motivation; Rheinberg et al., 2001) (α = .82)	tutorial	• I am eager to see how I will perform on today's task. • I'm really going to try as hard as I can on this task. • While doing this task I will enjoy discovering how to brew beer. • I would work on this task even in my free time (if I have the instructional animation).	*agree at all*; *7 – I completely agree)*
Generalized positive affect (i.e., the positive scale of PANAS; Watson et al., 1988) (α = .87, .88)	3rd and 4th in situ; after the error and the task-solving parts	10 items, see Watson et al. (1988) with the following initial instruction: • Mark to what extent you experience these feelings at this moment: [the list of 10 feelings; e.g., interested, active, alert, excited].	ten 5-point Likert items (*1 – very slightly or not at all*; *5 – extremely)*
Flow (Flow Short Scale; Rheinberg et al., 2003) (α = .93, .90)	3rd and 4th in situ; after the error and the task-solving parts	10 items, see Rheinberg et al. (2003) e.g. • I do not notice time passing. • I feel I have everything under control. • I am completely lost in thought.	ten 7- point Likert items (*1 – definitely no*; *7 – definitely yes)*
Learning involvement (inspired by Schraw et al., 1995; Isen & Reeve, 2005) (α = .86, .88, .81)	2nd – 4th in situ; after the linear, the error and the task-solving parts	• So far, I have enjoyed brewing beer. • I always knew what to do next. • I always knew how to complete the assigned tasks. • I'm tired. • I'm looking forward to the next part [the 4th in situ administration: I'd like to continue in brewing beer] • I focused on brewing beer. • I think I am doing well so far. • I was careful and conscientious when completing the tasks.	eight 7-point Likert items (*1 – definitely no*; *7 – definitely yes)*
Enjoyment	post hoc	• I enjoyed doing this activity • I would describe this activity as very interesting	two 6-point Likert items (*1 – very much*; *6 – very little)*; reverse coded
Perceived difficulty (α = .70)	2nd – 4th in situ, post hoc	• The difficulty of the simulation [4th in situ administration: task-solving] meets my expectations.	four 7-point Likert items (*1 – very easy*; *7 – very difficult)*
Perceived learning	post hoc	• How much do you know about brewing beer?	two 6-point Likert items (*1 – very*

		• Do you think that you have learnt something about brewing beer? •	*much*; *6 – very little)*; reverse coded
Negative affect (i.e., the negative scale of PANAS; Watson et al., 1988) (α = .84, .88)	3rd and 4th in situ; after the error and the task-solving parts	10 items, see Watson et al. (1988) with the following initial instruction: • Mark to what extent you experience these feelings at this moment: [the list of 10 feelings; e.g. irritable, distressed, upset].	ten 5-point Likert items (*1 – very slightly or not at all*; *5 – extremely)*
Manipulation check	2nd – 3rd in situ	• For you personally, would it be better if the grandpa praised you:	5-point Likert item (*1 – much less often; 5 – much more often)*
		• For you personally, how important was it that the grandpa awarded you points? (rating importance and valence)	two 7-point Likert items (importance: *1 - very little; 7 – very much;* valence: *1 – very negative; 7 – very positive)*
	4th in situ	• For you personally, how important was it that money was part of the game? (rating importance and valence)	two 7-point Likert items (importance: *1 - very little; 7 – very much;* valence: *1 – very negative; 7 – very positive)*
		• For you personally, how important was it that you had to achieve a game goal? (rating importance and valence)	two 7-point Likert items (importance: *1 - very little; 7 – very much;* valence: *1 – very negative; 7 – very positive)*

2. Knowledge tests – question examples:

Retention, e.g.:

- Write down names of the four main phases of beer brewing in the correct order, as you learned today.
- In what phase or phases of beer brewing are enzymes present during the whole phase?
- Please, describe how temperature is being changed during the whole process.
- Please explain what happens during the fermentation phase and what main products are created during this phase. Imagine you are writing a short encyclopedia entry for beginners. (open-ended)

Transfer, e.g.:

- Why does the chance that the product will spoil increase, if we cannot manage a stable temperature during the whole fermentation phase? **Explain in detail.**
- We got rid of bacteria during the boiling phase. However, after the conditioning, the product still contains acetone (which is a product of bacteria). When and how could acetone have got into the beer? Write down **every possibility** you can imagine.
- "How would you adjust the lager tank so that it can be used for fermentation? Write down **all possibilities** you can think of and **explain** why these changes would be needed." [emphasis always as in the original]

  Note: emphasis in the original.

For retention, α was .66 for the immediate test and .73 for the delayed test. For transfer, α were .80 and .63 for the immediate tests and .82 and .75 for the delayed tests.

References

Isen, A. M., & Reeve, J. (2005). The influence of positive affect on intrinsic and extrinsic motivation: Facilitating enjoyment of play, responsible work behavior, and self-control. *Motivation and emotion, 29*(4), 295-323.

McKenzie, D. L., & Padilla, M. J. (1986). The construction and validation of the test of graphing in science (TOGS). *Journal of Research in Science Teaching*, *23*(7), 571-579.

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen [QCM: A questionnaire to assess current motivation in learning situations]. *Diagnostica, 47*, 57-66.

Rheinberg, F., Vollmeyer, R., & Engeser, S. (2003). Die Erfassung des Flow-Erlebens [in German]. In J. Steinsmeier-Pelster & F. Rheinberg (Eds.), *Diagnostik von Motivation und Selbstkonzept* (pp. 261-279): Hogrefe.

Schraw, G., Bruning, R., & Svoboda, C. (1995). Sources of situational interest. *Journal of Literacy Research, 27*(1), 1-17.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology, 54*(6), 1063-1070.