# NOTICE



# CITATION

Brom, C., Bromová, E., Děchtěrenko, F., Buchtová, M., Pergel, M. (2013, Dec 10) Personalized Messages in a Brewery Educational Simulation: Is the Personalization Principle Less Robust than Previously Thought? *Computers and Education*. Advance online publication. doi: 10.1016/j.compedu.2013.11.013

# Title page

## *1) Full title*

Personalized Messages in a Brewery Educational Simulation: Is the Personalization Principle Less Robust than Previously Thought?

## *2) Authors*

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

Edita Bromová
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00, Prague, the Czech Republic
edita@email.cz

Filip Děchtěrenko
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00, Prague, the Czech Republic
filip.dechterenko@gmail.com

Michaela Buchtová
Faculty of Arts, Charles University in Prague
U kříže 8, 15800 Prague 5, Czech Republic
michaela.buchtova@ff.cuni.cz

Martin Pergel
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00, Prague, the Czech Republic
perm@kam.mff.cuni.cz

## 3) Corresponding author

Cyril Brom

Faculty of Mathematics and Physics, Charles University in Prague, Room 312, Malostranske Namesti 25, Prague, 11800, Czech Republic.

E-mail: brom@ksvi.mff.cuni.cz

Tel: (420) 221 914 216; Fax: (420) 221 914 281

## 4) Source of funding

Cyril Brom was partially supported by project nr. P407/12/P152 supported by Czech Grant Science Foundation (GA ČR). Michaela Buchtová was partially supported by projects Digital Technologies in Education (GAUK 581012) and Digital Simulations in Lifelong Learning (VG012 161873). Filip Děchtěrenko was partially supported by a student grant GAUK no. 68413.

## 5) Acknowledgement

We thank Tereza Stárková who helped with data collection and Jiří Lukavský for helpful comments on earlier versions of the manuscript. We also thank the staff of Laboratory of Behavioural and Linguistic studies who helped us with management of the subject pool and provided places for part of the experiment. We thank Jan L. Plass for comments on an early phase of our study and Bruce M. McLaren for comments on his work.

The human data were collected with APA ethical principles in mind.

# Personalized Messages in a Brewery Educational Simulation: Is the Personalization Principle Less Robust than Previously Thought?


## Abstract

The personalization principle, one of the design principles of multimedia learning, states that people learn better from multimedia presentations when instructions are in a conversational style rather than a formal style, possibly due to learners' increased interest. This principle was shown to be robust in short interventions that could be completed within minutes or a few dozen minutes; however, complex digital simulations and games that support the acquisition of complex mental models usually take longer to complete. In this study, we investigate the personalization principle in a new context: in an interactive simulation on the topic of beer brewing, which lasts 2-3 hours. Instructions were presented in the Czech language, either in a personalized style, where learners were addressed conversationally by "their grandpa, an owner of the family brewery," or in a non-personalized, more formal style without the grandpa. In Experiment 1, 26 college students, who interacted with both simulation versions, expressed on average a preference for the personalized version of the simulation. However, some of them worried that personalization could distract them. In Experiment 2 with a between-subject design, the knowledge of 75 predominantly college students was tested by means of retention and transfer tests immediately after completing the simulation and also a month later. Contrary to most previous works, our results showed no difference between the personalized and non-personalized groups in learning achievement, despite the fact that

learners who received the personalized treatment voluntarily spent about 20% more time on the simulation. We also applied various measures of the learner's affective state, including Flow Short Scale and PANAS, but – again – no between-group differences were observed. These results indicate that personalization is not always beneficial to learning, which raises important questions for future research. Additional findings suggest that the simulation, no matter the treatment type, was most beneficial to learners with high mathematical abilities and who play computer games frequently, and also to those who liked the simulation more.

## Keywords

simulations; interactive learning environments; serious games; evaluation of CAL systems; media in education; personalization principle; beer brewing; mental models

## 1. Introduction

One useful feature of *computer-based simulations* is that they enable students to observe a computational model of a complex phenomenon, interact with it and actively inspect its underlying causalities by investigating the consequences of their actions. This can help students develop a so-called *mental model* of the phenomenon; that is, an internal representation of possible behavior of the device/system being modeled and the possible evolution of situations and problems (Johnson-Laird, 1983; Gentner & Stevens, 1983). This supports the ability to draw inferences and make predictions about reality (cf. Papert, 1993).

An old idea on how to improve learning (in general) is to make the educational process engaging (e.g., Comenius, 1627 - 1633/1657/1967). The more the learners are interested, the

more they invest energy into learning and thus the more they learn. The caveat is that what causes the extra interest can also distract them, reducing their learning gains. Thus, if we have two versions of the same educational simulation, the one that has a higher ability to increase the learners' engagement and – at the same time – delivers higher engagement through fewer extraneous details should be, according to the idea, a more useful pedagogical tool as concerns the acquisition of mental models.

How can we increase learners' engagement with as few extraneous details as possible? The well-known *cognitive theory of multimedia learning* (CTLM, Mayer, 2001) offers us several principles that support effective learning from multimedia materials; including educational simulations and games (Mayer, 2011). Of these principles, the one that presents a possible answer to our question, and is supported empirically, is the so-called *personalization principle*. This principle states that "people learn better from multimedia presentations when words are in conversational style rather than formal style" (Mayer, 2001, p. 242). At the same time, an expansion of CTLM, *cognitive-affective theory of learning with media* (CATLM, Moreno, 2005; Moreno & Mayer, 2007) offers us a theory-grounded articulation of the idea from the previous paragraph and, based on that articulation, also one possible explanation for why the personalization principle might function.

According to CATLM, in a slightly simplified way, mental models are constructed in the learner's long-term memory by means of so-called *generative cognitive processing*. When a higher cognitive capacity for the generative cognitive processing is available, the better the mental model is constructed. Sadly, the total cognitive capacity of a learner does not equal the cognitive capacity for generative cognitive processing (Fig. 1a). The CATLM actually assumes the following trade-off: on the one hand, motivational factors can increase the total cognitive capacity (or the lack of the learner's motivation may fail to engage the learner in

generative processing even when cognitive capacity is available). On the other hand, the capacity available for generative cognitive processing may be *reduced* by the processing of so-called *extraneous details* that are not directly related to the learning content (Fig. 1b). However, at the same time, the extraneous details may help to make the learning materials engaging; thereby contributing to an increase in the learner's total cognitive capacity or recruiting generative cognitive processing (Fig. 1c).

Why might the personalization principle function? Instructions in conversational style may create a sense of social presence, which may lead to the learner's increased interest (Moreno & Mayer, 2004). If learners feel they are in a conversation with a partner, they may work harder (Beck & et al., 1996; cited from Clark & Mayer, 2011, p. 184). If that outweighs the possibly distracting effects of the personalization (compared to a "non-personalized" version of the text), the CATLM would predict that learners will invest more of their cognitive capacity into generative cognitive processing, thus leading to better learning.[1] We call that prediction motivation → learning framework.

--- Insert Fig. 1 about here ---

[1] Note that different explanations of how the personalization principle may function also exist, e.g. (Moreno & Mayer, 2000; Günizi, 2010), such as improved coding due to self-referential language used by personalized texts.

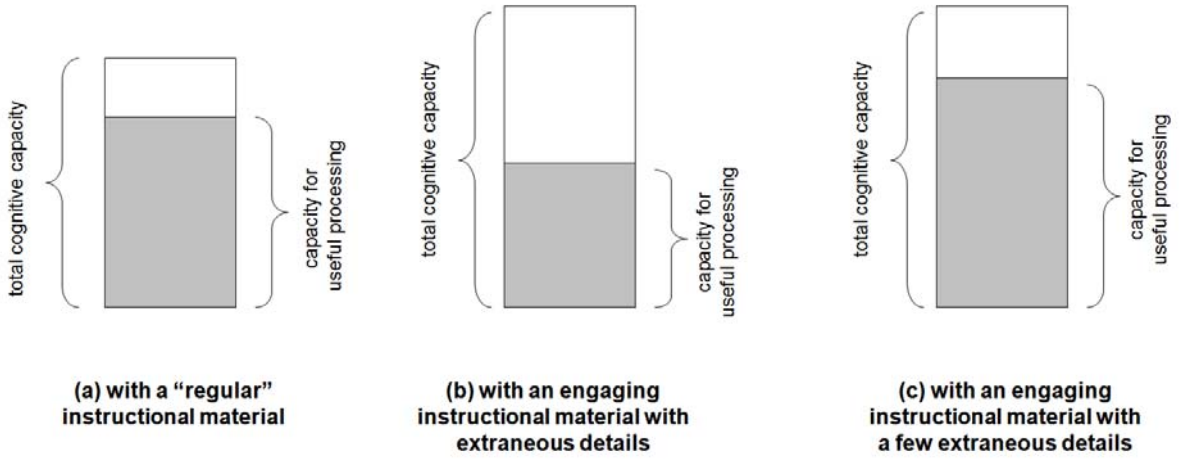


Figure 1: Learning can be compromised with instructional materials containing extraneous details. While total cognitive capacity is lower without an engaging material (a) than with engaging material (b, c) due to the material's motivational elements, the capacity for useful processing (that causes learning) can be lower with the engaging material if too many extraneous details are present (b).

Note, however, that the CATLM does not predict that *every* personalized text leads to better learning outcome. It merely creates an explanatory framework that is able to explain why materials with texts in conversational style *can* be better for learning than texts that are not in conversational style. However, the same framework is also able to explain the *reverse effect*; that is, a situation where personalized instructions, as opposed to non-personalized instructions, *compromise* learning (due to a distraction). It is worth mentioning that the personalization principle has not yet been demonstrated in a treatment lasting longer than about half an hour. In fact, to our best knowledge, it was only demonstrated with learning materials that supported, more or less, the acquisition of a mental model of a single cause-and-effect phenomenon. That objective can usually be achieved within minutes or a few

dozen minutes. However, many educational simulation, but also complex instructional tutorials and educational games last longer.

The best support for the personalization principle comes from the original work of Moreno & Mayer (2000). In this work they demonstrated the principle in two experiments using 140 seconds long animation of lighting formation and in three experiments using a roughly half-hour-long educational micro-game, Design-A-Plant (Lester, Stone & Stelling, 1999) on the topic of botanical anatomy; a game featuring a pedagogic agent. Then they replicated the principle using a modification of the latter treatment (Moreno & Mayer, 2004) and also by using a 60-second-long, narrated animation explaining the functioning of the human respiratory system (Mayer et al., 2004). Recently, Günizi (2010) replicated it using less-than-half-hour-long Turkish materials on stellar death that included texts, animations and static pictures. However, Doolittle (2010) tried to demonstrate the principle in a 2.5 hour-long, multimedia tutorial on teaching the high-level skill of critical historical reasoning, which included gathering, analyzing and interpreting historical sources, and he failed. One of the groups using non-personalized materials even outperformed a group using personalized materials. The effect was also not demonstrated in the study by McLaren et al. (2006) using a roughly hour-long, web-based tutorial on stoichiometry. That study was less controlled than other studies and around half of the participants may not have been native English speakers (the texts were in English).

To our knowledge, none of the above mentioned studies investigated directly the question of whether personalization could possibly have distracting effects. However, a so-called expertise-reversal effect was detected, for instance, in a study investigating a similar principle – the politeness principle (McLaren et al., 2011a). This effect is a pattern in which low prior knowledge learners benefit from the treatment (e.g., a more polite version of a web-based

tutorial), but not high prior knowledge learners. Thus, there is some evidence warning us that personalization may not perhaps be beneficial to all learners.

The main goal of the present study is to investigate an important boundary condition of the personalization principle: the complexity of the phenomenon being modeled. The *complexity*, as operationalized in this study, entails two things. First, the length of exposure to the treatment should be in terms of hours and not minutes. We are interested in long treatments because participants may get tired of or become bored by a too long intervention. Second, the educational objective of the treatment should be the acquisition of a *complex* cause-and-effect mental model, which means that the phenomenon being modeled can be decomposed to causally connected sub-phenomena, leading to the notion that the whole mental model comprises sub-models of these sub-phenomena connected by causal links. We are interested in this situation because if some learners are distracted by the personalization, this may be more easily detected when the modeled phenomenon is complex rather than simple. For the purpose of this experiment, we have developed an interactive educational simulation of the beer brewing process, which takes students 2 - 3 hours to complete and is offered in the Czech language[2]. If the personalization effect is not demonstrated, the study can also pave the way for future studies investigating under what conditions, and for what type of learners, the conversational style is a distracting factor.

We now introduce the goals and hypotheses of our study. Then we explain our brewery simulation. Afterwards, we present two experiments of this study. The paper concludes with a general discussion.

[2] Upon request, the simulation is also available in English for research purposes.

# 2. Goals and Hypotheses

In this study we have four different hypotheses, introduced in Sections 2.1 (Hypothesis 1), 2.2 (Hypothesis 2), and 2.3 (Hypotheses 3, 4).

## *2.1 Goal 1: Application of the Motivation → Learning Framework to the Personalization Principle*

The study consists of two experiments. Both employ two different versions of the same simulation: one with personalized instructional texts situated within a story about a brewery (P Version) and the other using the same simulation with non-personalized texts and without any story (N Version). The first experiment investigates if learners prefer the P Version of the simulation to the N Version, if given a choice. It also reports learners' opinions regarding the personalization, focusing on its possible distracting effects. The second is the main experiment that uses between-subject design and compares motivation and learning achievement of learners from two different groups: the group that uses the P Version of the simulation (P Group) and the group that uses the N Version of the simulation (N Group). In the main experiment, our primary dependent variable of interest is users' score on knowledge *transfer tests*, which measure learners' ability to apply what has been learnt in a new situation (Mayer, 2001). Our secondary dependent variables are related to the learner's affective state and they are measured by several means, including PANAS (Positive and Negative Affect Schedule; Watson et al., 1988), Flow Short Scale (Rheinberg et al., 2003) and length of exposure to the simulation. As an abbreviation, we will use the umbrella term "motivation" to denote these affective variables. This term is simplistic but intuitive for a broad spectrum of readers.

Our *Hypothesis 1* is: *The P Group's average score in transfer tests is higher than the N Group's average score ($P_{trans} > N_{trans}$) and – at the same time – the P Group's average motivation is higher than the N Group's average motivation ($P_{mot} > N_{mot}$).* Thus, in terms of the motivation → learning framework, our Hypothesis 1 states that texts in conversational style will motivate learners and therefore contribute to an increase in their learning gain, while possible distraction, due to extraneous details produced by the personalization, will be negligible. We base this hypothesis on the outcomes of previous studies (Moreno & Mayer, 2000; Mayer et al., 2004; Moreno & Mayer, 2004; Günizi, 2010) that demonstrated the personalization effect using short treatments concerning the acquisition of mental models for single cause-and-effect phenomena. If the hypothesis is confirmed, then personalization principle boundaries can be extended to *longer* treatments concerning acquisition of mental models for *complex* cause-and-effect phenomena.

However, studies by Doolittle (2010) and McLaren et al. (2006) warn us that we may not demonstrate the personalization effect. Thus, we have to design the experiment so that other outcomes can also be interpreted. The key possible outcomes we have to be able to account for come through the following dependent variables: learning scores and motivation. At the same time, the outcome can also inform us indirectly of the magnitude of possible distraction caused by one of the simulation versions (Common sense suggests that the P Version could serve as a distractor, mainly because it contains additional text compared to the N Version. However, in principle, it is also possible that the N Version is more distracting than the P Version, because learners' cognitive apparati may be more "tuned" to the P Version).

We now outline the main possible outcomes of the study explicated by means of the motivation → prediction framework. If multiple explanations are possible, we present the most parsimonious one:

1. $P_{trans} > N_{trans}$ and $P_{mot} > N_{mot}$; this is Hypothesis 1.

2. If $P_{trans} < N_{trans}$ and $P_{mot} < N_{mot}$, personalization had a detrimental effect on learners' motivation and therefore contributed to a decrease in their learning gain.

3. If $P_{trans} <= N_{trans}$ and $P_{mot} > N_{mot}$, personalization motivated learners but any positive learning effect that that could have had was outweighed by distractions due to personalization.

4. If $P_{trans} => N_{trans}$ and $P_{mot} < N_{mot}$, personalization had a detrimental effect on learners' motivation, but any negative effect that could have had was outweighed by an unknown factor (for example: non-personalized texts featured more unwanted extraneous details than personalized texts did).

5. If $P_{trans} = N_{trans}$ and $P_{mot} = N_{mot}$, personalization brings no motivational benefits and thus the learning gain differences are negligible.

6. If $P_{trans} > N_{trans}$ and $P_{mot} = N_{mot}$, personalization had a positive effect on learning gains; either due to an unknown factor or due to its positive influence on motivation that we were unable to detect.

7. If $P_{trans} < N_{trans}$ and $P_{mot} = N_{mot}$, personalization compromised learning due to either an unknown factor or due to its negative influence on motivation that we were unable to detect.

Anyway, according to CATML, we should find *a positive relation between motivation and learning effects*; assuming the compromising effect of the processing of extraneous details on the capacity for useful processing is, more or less, constant within a group undergoing the

same type of treatment. Note that this is not a hypothesis of this study; rather it is a justification for the usage of the motivation → learning framework for explanatory purposes.

## 2.2 Goal 2: Investigation of the Personalization Principle's Lasting Effects

Information about the *lasting* effects of a learning intervention is generally more useful than information about the immediate effect. Sadly, to our knowledge, lasting effects of the personalization have not been investigated. Yet we know from old reviews of studies on learning effects of educational (often non-computer based) simulation games, i.e. studies that (usually) compared simulation games to "traditional teaching methods", that simulation games might actually improve retention or understanding, as measured by delayed post-tests, despite the fact that no immediate effect is found (Pierfy, 1977; Randel et al., 1992; see also Brom et al., 2011, who summarized little additional evidence concerning educational computer games). Because the possible explanation of this effect involves the games' high motivational factor in combination with "hands-on-experience," both of which also play a role in the present study, this work's second goal is to collect data a month after the intervention; giving us an immediate post-test/delayed-post-test design for the main experiment.

Based on the evidence mentioned above, our *Hypothesis 2* is: *Between-group differences in transfer test scores detected in delayed tests will be larger than the differences in immediate tests and they will favor the P Group. If no between-group differences are found in immediate transfer tests, we predict differences in delayed transfer tests in favor of the P Group.*

## 2.3. Goal 3: Investigation of What Personal Characteristics Moderate Learning and Motivational Outcomes

It is a common practice to use undergraduates studying psychology as participants in educational research. It is also typical to focus only on the effect of one or two independent variables in studies investigating the learning effects of instructional innovations. However, in the context of educational simulations (and games) it is important not only to determine what aspects of interventions are beneficial to learning, but also what aspects work well for which participants (Tobias et al., 2011). Indeed, we can expect between-subject differences; for instance, regarding the amount of *a priori* knowledge (e.g., Tobias et al., 2011, p. 201; cf. also the expertise-reversal effect, e.g., McLarren et al., 2011a). For that reason, we use a heterogeneous sample: undergraduates with diverse study backgrounds; including computer science, physics, mathematics, language studies, psychology, arts, new media studies, librarianship and psychology (and a few others). At the same time, we use low prior knowledge learners in order to: a) isolate the effect of the key participants' characteristics relevant for Hypotheses 3 and 4 mentioned below (without exploring their interaction with the amount of learners' prior knowledge); and b) to explore if personalization could have distracting effects on some learners in the context of a complex simulation when taking the expertise-reversal effect away.

Our *Hypotheses 3* and *4* concern the individual characteristics of participants and they are:

3) *Participants who have higher mathematical abilities will acquire mental models of complex mechanistic phenomena (exemplified in the brewing process model) better than participants less able in mathematics.* The rationale is that people with higher mathematical abilities may be more able than people with lower mathematical abilities to acquire complex

mental models in general (but may be less able, for instance, to memorize facts). Mathematical abilities and self-ratings of mathematical abilities were shown to predict performance in some complex skill acquisition tasks, e.g. in an air traffic controller simulation tasks (Ackerman et al., 1995). However, studies that would demonstrate analogical results concerning acquisition of complex mental models are unknown to us.

4) *Participants who play computer games or live action role-playing games or social role-playing games often (but not card games and other table games) will acquire mental models of complex mechanistic phenomena (exemplified in the brewing process model) using a motivating and complex educational simulation better than participants who play this kind of games rarely or never.* The rationale is twofold: first, frequent players can enjoy the simulation more and thus, according to the motivation → learning framework, would learn more; second, frequent players can find it easier to interface with a complex simulation, whose controls resemble those of a computer game, and this can leave more of their cognitive capacity for processing educational content as opposed to learning how to control the simulation (cf. Mayer's pre-training principle; Mayer, 2001, chap. 10).

## 3. Brewery Educational Simulation

Our main goal was to investigate the personalization effect in the context of a longer educational simulation, whose educational objective is the acquisition of a complex, cause-and-effect mental model. To achieve that objective, we needed to find a phenomenon reasonably complex and motivating enough for the participants to learn it. At the same time, it had to be one about which participants had a low a priori knowledge, because our goal was not to investigate interactions with prior knowledge. We picked the process of beer brewing and designed and developed its educational simulation; including textual instructions. Note

that beer brewing is a complex process consisting of several sub-processes. Additionally, beer brewing is a source of national pride in the Czech Republic and a personally meaningful task for many Czechs. Thus high motivation is to be expected. Finally, many people do not actually know how to brew beer (low a priori knowledge was confirmed as detailed in Sec. 5.2.1).

We now provide an overview of the beer brewing process (Sec. 3.1), give detail on its simulation (Sec. 3.2) and contrast our personalized version to the non-personalized version (Sec. 3.3). We remark that the results of our study, which we explain in detail later in this paper, do not always agree with past results. Therefore, we describe the simulation in detail in order to enable the contrasting of our simulation to other interventions; for the purposes of future studies and reviews.

## *3.1 Beer Brewing Process*

The beer-brewing process is a rather complicated one. This is because more than 2,000 sensorically active chemical compounds were isolated in the beer (Esslinger, 2009). Thus we focused only on the main topics that can be used, for example, by a home-brewer as part of his/her first steps. We focused only on the technology used for bottom-fermented lagers of the pilsner type. We omitted all the processes that can be outsourced (without loss of quality); like malt barley preparation. Our overall aim was to keep the whole simulation sufficiently complex so as to achieve our objective.

To briefly describe the process, it consists of mashing, lautering, boiling, whirlpooling, wort cooling, (cool) fermenting and conditioning. As lautering and whirlpooling are procedures demanding mainly manual (technical) abilities, they cannot be taught through a computer simulation. So we focused on the remaining phases only: i.e., on *mashing*, *boiling*, *fermenting*

and *conditioning*. Filtering and packaging are also not simulated. The mashing was simplified and so we implemented infusion mashing. The boiling process is restricted to the addition of hops. The fermentation and conditioning simulates the life of the yeast in the wort.

We focused on typical points in each of the procedures, during which errors can be made and where the learner's knowledge of the process can be assessed easily in *post hoc* testing. During the infusion phase, we simulate how the enzymes in the malt break down the starch into sugars with respect to the temperature; having one universal sugar. During the boiling phase, we only check whether it lasts long enough and how the addition of hops corresponds to a particular recipe. During the fermentation phase, the yeast (*Saccharomyces cerevisiae*) converts sugars to alcohol and other chemicals (fusel). We simulate the activity of yeast with respect to the temperature; including the yeast's thermal shocks and its change in metabolism. We also simulate parasitic bacteria falling into the fermentation tank (*Clostridium acetobutylicum*). From the yeast's by-products we focus only on ethanol, $CO_2$ and generic fusel. For the bacteria we focus on acetone. During the conditioning phase, a slight change in taste should take place. We omit this moment, but it does not matter too much as we simulate the $CO_2$ concentration (whose absence creates a worse beer) and the fusel concentration (which increases if the yeast produces $CO_2$ too quickly). Before the conditioning phase, it is possible to make a sugar-surrogation.

## *3.2 Beer Brewing Simulation*

We developed the educational simulation modeling the brewing process, as described in Sec. 3.1, using the Netlogo toolkit (Wilensky, 1999). The process model and textual instructions were consulted with an expert on beer brewing.

The simulation's graphical interface (Figure 2) consists of the following:

- textual instructions,
- animation panel showing the content of the fermentation vessel,
- supplementary explanation panel relaying the meaning of graphical elements,
- four panels with graphs and histograms showing the amount of ingredients in the product,
- a timer showing time elapsed since the beginning of the current phase of brewing,
- an adjustable thermometer,
- two panels showing the “number” of bacteria and yeast in the fermentation vessel (i.e. showing a numerical value),
- a button for clearing the fermentation vessel (starting the simulation),
- twelve buttons for controlling the process,
- three buttons for navigating through the instructions,
- four buttons for restarting the current phase (one button for each of the four phases; the buttons feature a house image pictogram, symbolizing the “home” command),
- one button for showing the product quality assessment for the current phase;
- a slider controlling the speed of the simulation.

--- Insert Figure 2 here ---

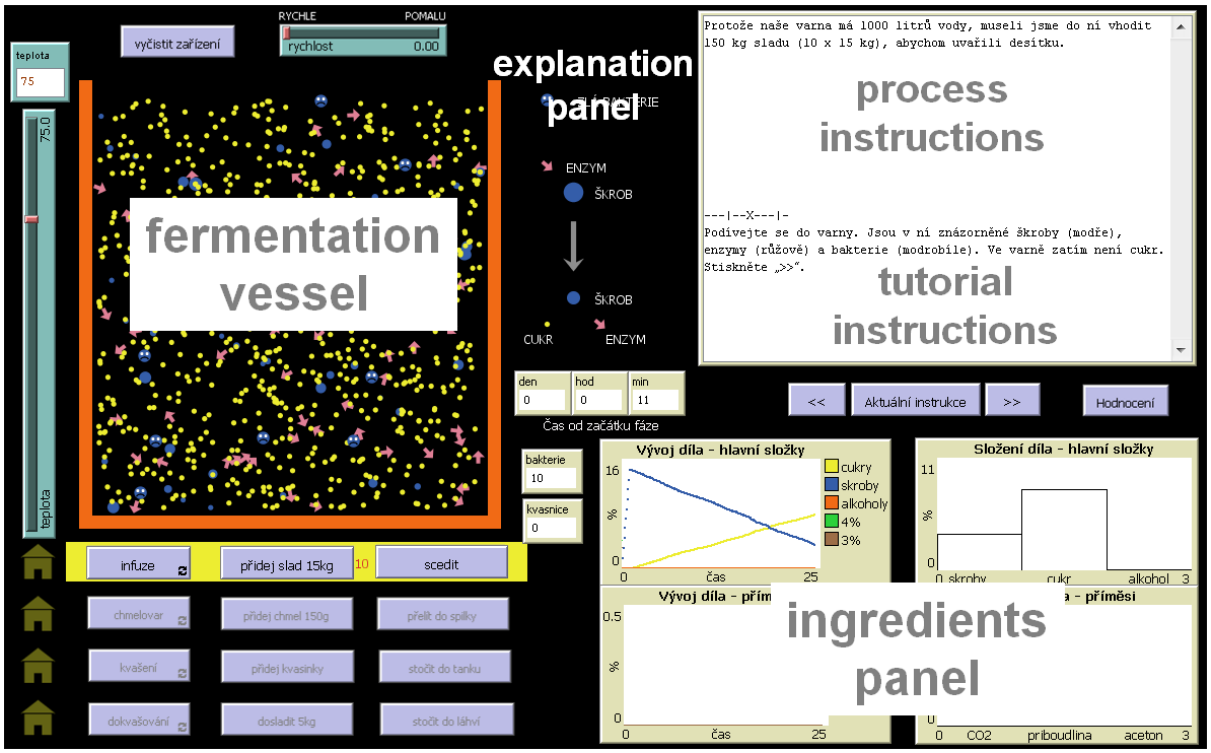


Figure 2: The simulation screenshot. The main elements of the graphical interface are described.

The application has been designed to allow learners to proceed at their own pace. There are two types of instructions: *process instructions* that describe the beer brewing process *per se* and *tutorial instructions* that tell the learner what to do next. Process instructions are depicted together with tutorial instructions; the former placed above the latter. The learner first studies

the process instructions and then follows the commands in the tutorial instructions. The tutorial instructions typically tell the learner what button to click and why, where to focus his/her attention, etc. This usually results in users running the simulation for a while and inspecting the consequences.

Unlike the English language, the Czech language features two syntactic forms of singular second-person pronouns, one being more informal ("*You* [informal] *can do it.*" is translated to "*Ty to můžeš udělat.*", where "*ty*" is, syntactically, the singular form of "*you*") and the other more formal ("*You* [formal] *can do it.*" is translated to "*Vy to můžete udělat.*", where "*vy*" is, syntactically, the plural form of "*you*"). The former is often used when interlocutors are familiar with each other, e.g. friends or family members, the latter when they are not, e.g. teacher - student, sales clerk - shopper. The non-personalized version (the N Version) is based on the formal form, which we will denote as the V-form or "$you_V$" according to Brown and Gilman (1960). The personalized version (the P Version) is based on the informal form, which we will denote as the T-form or "$you_T$". Similarly, Czech verbs have two syntactically different forms of the second-person singular, including imperatives; one informal and the other formal.

An example of two consecutive non-personalized instructions is shown on Tab. 1, along with the corresponding screenshot (Figure 2). Note that the use of the passive voice is rare in the Czech language, even when instructions are given in the non-personalized form. Take, for example, Tab. 1, Instruction #6; it would be unusual to state in Czech: "*Because the brewing tank holds 1000 liters of water, 150 kg of malt (10 x 15 kg) had to be added in order to brew 10-degree beer.*" Instead the active plural form is used: "*...we had to add...*" Thus, the N

Version uses texts with V-forms and, more often than not, the active voice instead of the passive voice.[3]

--- Insert Table 1 around here ---

Table 1: An example of two process instructions and their corresponding tutorial instructions from the N Version. The corresponding screenshot is depicted in Figure 2.

[3] Note that Günizi (2010) used a somewhat similar approach in her study of the personalization effect of Turkish instructions, supplementing a short set of animated images and pictures on the topic of stellar death. She used three different instructions: one based on T-form, one based on V-form and one using the third person without any comments directed to the learner (neutral form). Because it is not natural to use the neutral form in the Czech language, our N Version can be considered to be something in between Günizi's neutral form and her version using the V-form. Our changes made to turn the N Version into the P Version are detailed in Section 3.3 and Appendix A.

| | **Instruction #6** | **Instruction #7** |
|---|---|---|
| **Process instruction** | *Because the brewing tank holds 1000 liters of water, we had* [note: in the previous step] *to add 150 kg of malt (10 x 15 kg) to it in order to brew 10-degree beer.* | *Now we 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 we will not discuss them here.* |
| **Tutorial instruction** | *Look$_V$ 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 „>>“.* | *Set$_V$ the right temperature. Then look$_V$ at the „Infusion“ button. The button can be clicked on or off: in doing so you$_V$ either start$_V$ or stop$_V$ the infusion process. Now try$_V$ to click the button several times to either start or stop the simulation. Also notice$_V$ that the TIME INDICATOR, below the image panel, shows the time that has elapsed SINCE THE PHASE BEGAN. Let$_V$ the infusion run for 5 to 10 minutes and then stop$_V$ the simulation and click$_V$ „>>“.* |

The whole simulation has four parts (do not confuse them with the four phases of beer brewing: mashing, boiling, fermenting and conditioning). These parts are: a *tutorial*, a *linear* part, an *error part*, and *tasks*. They are detailed in Appendix A. Notably, in the final part, a learner uses the simulation to brew his/her beer of a specific type.

The learner has several means for controlling the simulation, depending on the phase. The simulation provides feedback via an "assessment" button. The user can also continuously monitor the amount of ingredients through the graphs, histograms and numerical panels.

The details of controlling the simulation are also outlined in Appendix A.

The application was developed with Mayer's principles of instructional design (Mayer, 2001) in mind. Namely, a pre-training principle has been applied by using the tutorial before the actual learning starts (i.e., before the linear part of the simulation starts). The spatial contiguity principle has been implemented by keeping the on-screen text as close to the fermentation vessel as possible and several captions appear directly in the vessel. The signaling principle has been implemented by using capital letters in the text to highlight keywords or ideas (see Table 1) and by changing the color in the fermentation vessel to emphasize changes related to a particular element. The coherence principle has been applied by keeping the graphics schematic and by removing unnecessary information from the instructional texts during a pilot study. Finally, according to the segmenting principle, the users proceed through the simulation at their own pace.

## *3.3 Personalized Version*

For the purpose of this study, the personalization has been operationalized as follows: a) an engaging background story is added, as detailed below; b) as a consequence, on-screen

instructions are given in a conversational style rather than a formal style, including changing V-forms to T-forms. Details are given in Appendix B. Otherwise, the personalized version of the simulation is exactly the same as the non-personalized version.

The objective of the background story is to justify and contextualize the personalization, while remaining relevant to the task performance. It is explained to learners verbally by the experimenter: "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 brewmaster. In the fifties, the communists confiscated your family brewery, but it was returned to your family after the Velvet Revolution in the nineties. Afterwards, your grandpa ran the brewery for about 20 years, but he is now 85 years old and he is looking for his successor. You are one of the people he has chosen to take on this role. This doesn't 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 ones interact with it the best they know how. Only then would he allow the very best candidate to be trained at the real brewery and possibly succeed him. Your grandpa will speak to you, via textual instructions, for the duration of the simulation. Everything written in the instructions is what your grandpa would say."[4]

Note that some past personalization studies investigated treatments without background stories (e.g., Mayer et al., 2004), others featured background stories in both the personalized as well as the non-personalized versions (e.g. Moreno & Mayer, 2000; Exp. 3 – 5) and yet

[4] This is common knowledge among Czech university students: communists would have confiscated the brewery around 1950. The Velvet Revolution occurred in 1989.

others introduced some narrative aspects only in the personalized version (Doolittle, 2010). Generally, narrative aspects tend to appear in longer treatments; probably because a longer exposure necessitates the contextualization of the personalization. Otherwise, the personalization might look unnatural (which may not be the case for interventions lasting a couple of minutes).

In about half of the personalization studies, researchers added some aspects of politeness to the personalized version. In our case, the P Version features little politeness, similarly to the studies of Moreno & Mayer (2000, Exp. 1 – 2) and Mayer et al. (2004). Our P Version is based rather on different aspects of conversational style, which are detailed in Appendix B. Generally, however, it is important to keep in mind that our operationalization of personalization is not directly comparable to all previous treatments using personalization (and not all the previous treatments are directly comparable to each other).

Table 2 presents a personalized version of the instructions shown for the N Version in Table 1.

--- Insert Table 2 around here ---

Table 2: Examples of two process instructions and the corresponding tutorial instructions in the P Version. Differences between the P and the N Version in the Table 1 are underlined.

| | **Instruction #6** | **Instruction #7** |
|---|---|---|
| **Process instruction** | *Excellent! 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.* | *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. Also 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$_T$ „>>".* |

The P Version has 6,750 words and 39,489 letters in total, including spaces and carriage returns, while the N Version has 6,138 words and 36,833 letters including spaces and carriage returns. This means that the P Version is nearly 10% longer when measured in terms of the number of words and around 7.2% longer when measured in terms of the number of letters.

# 4. Experiment 1: Will Students Prefer the Personalized Simulation?

Some past works demonstrated that university students tend to prefer polite (Mayer et al., 2006) or personalized versions of instructional materials (Moreno & Mayer, 2004), especially in combination with informal language (Günizi, 2010), but the results are not unequivocal (see Moreno & Mayer, 2000, Exp. 3 - 5; Mayer et al., 2004). For this reason, we aimed at demonstrating directly preference of university learners towards the P Version of our instructional intervention. We also wanted to pin down reasons behind the learners' preference, being interested in inter-individual differences. In Experiment 1, we let the participants interact with the first parts of both versions of our simulation and asked them what version they preferred and why.

## *4.1 Method*

### 4.1.1 Participants

The participants consisted of 26 university students: 19 students of humanities (namely new media or librarianship (10 males, 9 females)), and 7 students of technical disciplines (namely

computer science or mathematics or physics (6 males, 1 female)). All of these participants spoke the Czech or Slovak language fluently.[5]

### 4.1.2 Experimental Design and Procedure

Using within-subject design, we showed half the participants the P Version of the tutorial of the simulation, described in Sec. 3, and then the N Version of the tutorial for the same simulation. The order was reversed for the second half of the participants (i.e., the N Version came first, the P Version second). Thus we have P-N and N-P treatment types.

The participants with technical backgrounds were tested in groups of 1, 3 and 3 per session. The participants with backgrounds in humanities were tested in groups of 9 and 10 per session. The whole group was assigned the same treatment type. The groups were formed on a random basis and the assignment of the treatment type to a group was also random. Each participant was seated at his/her computer. The experiment was anonymous; participants were assigned numbers. Together, 14 participants received the P-N treatment type and 12 the N-P treatment type.

When the experiment started, participants were welcomed and explained that they would interact during the experiment with the first parts of two slightly different versions of the same educational simulation on the topic of beer brewing. They were also informed that a short questionnaire would be administered at the end of the simulation. They were told nothing specific about the personalization of the instructional texts at the beginning.

---

[5] Note that Slovak language is very close to Czech language. Many Slovak students study in the Czech Republic and it is generally no problem for Slovak students to understand or even speak Czech.

Participants with the P-N treatment were then told the introductory story about the family brewery (Sec. 3; about 3 minutes) and explained how the simulation's interface works (about 2 minutes). They were instructed to read carefully both the process as well as the tutorial instructions. They were told not to try to read only the tutorial instructions, because the purpose of the simulation is to teach them the beer brewing process described mainly in the process instructions. Then they interacted with the P simulation tutorial at their own pace (about 10 - 15 minutes). Participants who finished sooner waited for the slower ones. As soon as the slowest participants finished (2 participants were interrupted at the 15-minute mark), the administrator ran the N Version tutorial for every participant and explained to them that they would now interact with the same simulation but with non-personalized textual instructions. They were to imagine that the simulation was not framed in the story of a family brewery. The learners could interact at their own pace with the N Version for 20 minutes; if they finished the tutorial, they could start the linear part of the N Version (as detailed in Sec. 3 and Appendix A). The participants were interrupted at the 20-minute mark and a short experience questionnaire was administered to them (5 minutes).

The procedure was the same for the N-P participants, except that they were first introduced to the simulation interface using the N Version (about 2 minutes) and then they interacted with the N Version tutorial (up to 15 minutes; 1 participant was interrupted at the 15-minute mark). They were then told the story about the family brewery (around 3 minutes), and afterwards took part in the P Version tutorial (and possibly with its linear part – 20 minutes). Finally, the experience questionnaire was again administered (around 5 minutes).

### 4.1.3 Materials, Apparatus

The pen-and-pencil questionnaire asked the participants:

a) the following two questions with a 6-point Likert scale: “How did you like today’s extract from the lesson about beer brewing?” (*1 - very much; 6 - very little*), and “If you had the possibility, would you like to continue interfacing with the simulation?” (*1 - definitely yes; 6 - definitely no*);

b) the following question with a 5-point Likert scale: “Should you continue interfacing with the simulation, would you prefer the version with the grandpa or without the grandpa?” (*1 - I definitely prefer the grandpa; 5 - I definitely prefer the version without the grandpa*; the question was scored as +2 (the grandpa) ... –2 (not the grandpa));

c) the following open-ended question: “Please explain briefly your answer to the previous question.”

The simulation is described in Sec. 3. It was run on notebooks or desktop PCs with at least 17"-wide screens.

## *4.2 Results and Discussion*

The primary question was whether the participants would prefer the P Version of the simulation. The results indicated that they did (*Mean*=0.5; *SD*=1.02; $t(25)=2.476$; $p=0.020$). No subject selected –2 (“definitely not with the grandpa”), while several selected +2 (“definitely with the grandpa”). We observed no noticeable differences between males and females or between the P-N and N-P treatment types. The preference trend was apparent both among students with humanities backgrounds as well as technical backgrounds.

The secondary question was whether we would find inter-individual differences explaining this preference. The open-ended question yielded interesting outcomes in this regard (Tab. 3).

In general, the students also liked the simulation ("like" question: *Mean*=2.04; *SD*=1.15; "continue" question: *Mean*=1.65; *SD*=1.2) and the differences between the P-N and N-P treatments, between males and females, and between learners with different study backgrounds were again negligible.

Thus, we presented additional evidence supporting the idea that university learners, on average, prefer personalized studying materials. More importantly, what is new in our study are qualitative data (Tab. 3), which directly supports the disconcerting idea that personalization may increase the motivation of learners but can also distract them. In other words, the personalization may work for some learners but not for others. This idea is important for our Hypothesis 1.

--- Insert Table 3 here ---

Tab 3: All relevant participants' explanations of their preferences. Our notes are given in square brackets. The number of times a comment recurs is shown in the third column.

| **Participant's preference** | **Comment** | **Background** |
|---|---|---|
| (+2 ... "definitely the grandpa";<br>-2 ... "definitely not the | | (T - technical; H - humanities) |

| | | |
|---|---|---|
| grandpa") | | (numbers) |
| +2 | There's a story there, so it's more engaging. | T |
| +2 | The instructions are more understandable. | H (2x) |
| +2 | The immersion into the simulation is greater. | H |
| +2 | The grandpa said "excellent" - he praised me. | H |
| +1 | Because he [the grandpa] praised me. | H |
| +1 | It is [the P Version] more personal and likable. | H (2x) |
| +1 | Even though I know it's only a game, I have a higher motivation to study. [note the participant used the word "game"; this word was intentionally avoided by the experimenter during the whole session] | T |
| 0 | It was [the difference] only for motivational purposes, to make people to | H |

| | | |
|---|---|---|
| | try harder, was it not? | |
| 0 | The difference was negligible. / I didn't notice large differences. | T (2 x)<br>H (3 x) |
| -1 | I prefer the formal style, because I'm used to it. | T |
| -1 | The extra comments [made by the grandpa] have no purpose, I prefer a technical manual. | T |
| -1 | I was distracted by the story ... I concentrated less on the instructions and more on the story. | H |
| -1 | The instructions without the grandpa were clearer, more precise (for me). | H |

# 5. Experiment 2: Will Students Learn Better with Personalized Instructions and What Influences the Learning Outcome?

Experiment 1 indicated that if learners had the possibility to choose between the P and N Versions of our simulation, based on 15-20 minutes long experience with each, more of them would opt for the P Version rather than the N Version. However, that does not automatically

mean that they would have learnt more from the P Version. Nor does it mean that if we ask them how motivating their experience has been after they just finish the whole intervention, which lasts 2-3 hours, that they would still think that the P Version is better than the N Version. It was our goal in Experiment 2, the study's main experiment, to investigate the actual learning outcomes of the two versions of the simulation and the impact of motivational and other variables on learning outcomes; i.e. to collect data to verify our hypotheses outlined in Section 2.

## *5.1 Method*

### 5.1.1 Experimental Design

The study used between-subject design with two groups and it compared learning from a personalized version of the beer brewing simulation (P Group) to the same simulation with non-personalized instructions (N Group). The participants a) assessed subjectively their prior knowledge of beer brewing and general alcohol production in a pre-questionnaire, which also yielded biographical data, their self-evaluation of mathematical knowledge and their frequency of playing computer, board and non-computer simulation games (and also some other data). The participants' actual performance was measured by three means: b) by their achievement in solving tasks embedded in the simulation (to brew a particular beer type), and c) by their achievement in retention and transfer tests collected immediately after the intervention and d) a month after the intervention. We also administered e) inventories yielding information about participants' flow experience and affective state during the intervention, f) questionnaires in which participants evaluated subjectively their simulation experience; immediately after the treatment as well as a month after the treatment, and g) a test on graphing in science a month after the treatment. We also controlled for the time needed

to finish individual parts of the simulation (i.e., the tutorial, the linear part, the error part and the tasks), and the time needed to complete tests and questionnaires.

The main dependent variables were performance in simulation tasks, scores in the retention and transfer tests and participants' subjective evaluation of the simulation experience. The main moderator variables were self-evaluation of one's mathematical knowledge, the frequency of playing computer games and non-computer simulation games, and time needed to finish the simulation not including the tasks (i.e. the first three parts).

Concerning qualitative data, after completing the one-month delayed post-tests, we let half of the participants perform yet another task embedded in the simulation environment. Then we conducted a brief interview about how hard it was for them to finish the task. We conducted a longer interview with the second half of the participants, aimed at gathering information about their overall perception of the simulation and about their opinion concerning the personalization.

In the first session, participants were tested in groups of between 1 and 7 persons per session. We had 22 groups in total. All participants from one group had the same simulation version, either the P or the N Version. The assignment was random.

One-month delayed post-tests were administered to between 1 and 4 participants per session. Together we had 38 groups for the delayed post-tests administration. These post-test groups combined participants originally having P and N treatments, but the participants did not have the opportunity to discuss that difference until the test session ended.

### 5.1.2 Participants

The participants were 73 university students with intentionally diverse backgrounds; including computer science, physics, mathematics, language studies, psychology, arts, new media studies, librarianship and psychology (and a few others), plus 2 non-university students (75 in total; 5 of them did not attend the delayed post-tests session). The reason for having participants with diverse backgrounds was the fact that we were investigating the impact of individual participants' characteristics on learning outcomes and motivation (Hypotheses 3, 4). All of the participants spoke the Czech or Slovak languages fluently. Two additional participants were excluded from the study for not being fluent in one of these two languages. The age of participants ranged from 18 to 31 (*Mean*=22.08; *SD*=2.32), and one outlier was 40 years old. The breakdown of participants is given in Table 4.

For all participants studying arts and for most participants studying computer science and new media, experiment participation was part of their assignment for the course Computer Games Development at Charles University in Prague (CUP) (around 20 participants). Psychology and language studies students were recruited from the participant pool of students at the CUP Faculty of Arts and received course credit for their participation (around 25 students). The remaining students volunteered for the experiment and received 400 CZK (around 25 USD) as compensation when they accomplished their one-month delayed post-tests. Around half of these remaining participants were from CUP. The data collection took place from November 2012 to April 2013. As detailed below, all of the participants can be considered to have low prior knowledge of the topic of beer brewing.

--- Insert Table 4 here ---

Table 4: Breakdown of participants according to research group, gender and study background.

| **Students** | | **Group** | |
|---|---|---|---|
| | | **P** | **N** |
| Technical (computer science, physics, mathematics) | Males | 11 | 11 |
| | Females | 5 | 3 |
| All others | Males | 8 | 9 |
| | Females | 12 | 16 |
| Total | Males | 19 | 20 |
| | Females | 17 | 19 |

### 5.1.3 Materials - Questionnaires

For each participant, the pen-and-pencil materials consisted of several tests typed on A4 paper sheets, as detailed in Table 5.[6]

--- Insert Table 5 here ---

Table 5: The names of the tests, time of their administration, the number of questions, the number of A4 sheets for each test, and time needed to complete each test. Note that times needed to complete the tests are actually means of real values: the time to complete the tests was a controlled variable. For instance, on average, the Flow questionnaire 2 really took participants a shorter time to complete than the Flow questionnaire 1, probably due to the familiarity effect. [7]

---

[6] Our pilot experiments indicated that participants prefer to fill in tests by hand rather than using their electronic counterpart, i.e. by typing on a keyboard.

[7] In the test session a month after the original experiment, four extra psychological inventories were administered. They are not listed in Table 5. These questionnaires are irrelevant to the present study and it took participants 14 minutes on average to complete all of them. Two were administered between the Transfer test 2 and the Graphing test; the other two at the end.

| Name | When administered | Number of questions | Number of A4 sheets | Time to complete |
|---|---|---|---|---|
| Pre-questionnaire | Immediately before the intervention | 21 | 3 | Approx. 7 minutes |
| Flow questionnaire 1 | The 1st one right after the participant has finished the error part | 16 | 1 | Approx. 3 minutes |
| PANAS 1 | The 2nd one right after the participant has finished the error part | 20 | 1 | Approx. 3 minutes |
| Flow questionnaire 2 | The 1st one right after the participant has finished solving the simulation tasks | The same as Flow 1 | The same as Flow 1 | Approx. 2 minutes |
| PANAS 2 | The 2nd one right after the participant has finished solving the simulation tasks | The same as PANAS 1 | The same as PANAS 1 | Approx 2 minutes |
| Motivation questionnaire 1 | The 1st one right after the intervention | 8 | 1 | Approx. 3 minutes |
| Retention test 1 | The 2nd one right after the intervention | 2 versions; both with 11 questions | 2 | Approx. 9 minutes |
| Transfer test 1 | The 3rd one right after the intervention | 2 versions; one with 8 and one with 6 questions | 8 or 6, each typed on a separate A4 sheet | Approx. 20 minutes |
| Motivation questionnaire 2 | The 1st one a month later | 14 | 2 | Approx. 4 minutes |

| | | | | |
|---|---|---|---|---|
| Retention test 2 | The 2nd one a month later | The same as Retention test 1 | The same as Retention test 1 | Approx. 8 minutes |
| Transfer test 2 | The 3rd one a month later | The same as Transfer test 1 | The same as Transfer test 1 | Approx. 18 minutes |
| Test of graphing in science (shortened to TOGS) | The 6th one a month later | 9 | 3 | Exactly 5 minutes |

*Pre-questionnaire.*

The purpose of the Pre-questionnaire was to solicit information about participants' gender, age and field of study. We also measured participants' frequency of using computers and playing computer games on a 4-point scale ranging from "*1) less than one hour a week*" *to* "*4) more than 10 hours a week*"; frequency of playing board/card games on a 4-point scale ("*1) never or less than once a year*"; "*4) at least once a month*"); and frequency of playing experiential and/or simulation games or tabletop role-playing games on a 5-point scale ("*1) never or I don't know what these terms mean*"; "*5) at least once a month on average*."). The last two variables are denoted as *Frequency of playing board games* and *Frequency of LARP playing* respectively.

Participants' self-assessed knowledge of mathematics was measured with one item with 6-point Likert scale (*1 - very good; 6 - very weak*) and participants' self-perceived ability to acquire mental models of mechanisms and processes with one item with 7-point Likert scale (*1 - very weak; 7 - very good*).

Finally, because we did not opt for real knowledge pre-tests in order to avoid cuing participants on what they should remember (see, e.g., Judd et al., 1991), we included eight

questions to measure indirectly participants' knowledge of beer brewing and making alcohol. This is also similar to how many other multimedia learning studies estimate prior knowledge.

The exact wording of each question is described in Appendix C.

*Flow Questionnaire and PANAS*

To measure the participants' experience of flow when interacting with the simulation, we administered a Flow Short Scale (Rheinberg et al., 2003; see also Engeser & Rheinberg, 2008). In this study, we report the data from its first subscale measuring components of flow experience with ten 7-point Likert items. The questionnaire was administered twice: after the error part (Flow 1) and after solving the tasks (Flow 2). The Cronbach alpha was 0.86 for the Flow 1 and 0.89 for the Flow 2.

To obtain information about participants' affective state when interacting with the simulation, we administered PANAS (Positive and Negative Affect Schedule; Watson et al., 1988), which consists of two mood scales; one for positive and the other for negative affect. Each scale consists of ten 5-point Likert items. The questionnaire was administered twice: immediately after both the Flow questionnaires. The Cronbach alpha was 0.86 for the positive scale and 0.81 for the negative scale of PANAS 1, while the values were 0.90 for positive scale and 0.83 for negative scale of PANAS 2.

*Retention and Transfer Tests*

We had two very similar versions of the Retention test; both with 11 questions. The tests differed mainly in their wording and ordering of questions. Each participant was given one version during the immediate testing session and the other during the delayed testing session. The order in which tests were administered was counterbalanced across participants, i.e. half of the participants were given the first version in the immediate testing session and the second version in the delayed testing session, and vice versa. The test contained nine short-answer

questions, such as “Write down names of the four main phases of beer brewing in the correct order, as you learnt today.” or “In what phase or phases of beer brewing are enzymes present during the whole phase?” The test also contained one multiple choice question 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.”

We had two versions of the Transfer test; one with 6 open-ended questions and one with 8 open-ended questions. The questions differed in the two versions, but they were paired across the versions and the paired questions tested similar knowledge. One question in the shorter version was “paired” to three questions in the longer version. In terms of Mayer’s research (2001, p. 39), we used all kinds of transfer questions: conceptual, prediction, redesign and troubleshooting questions. Examples of questions include: “Why does the chance that the product will spoil increase, if we cannot manage a stable temperature over the whole fermentation phase? **Explain in detail.**” or “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.” or “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]. Each question was typed on a separate A4 sheet of paper. Below the text for each question, 6 - 10 blank lines were included: space for writing the answer. Participants had an allotted time to complete each question, ranging from 2 to 5 minutes. That time was typed above each question.

We made both Retention and Transfer tests and iteratively refined them during a pilot study. During the first phase of the pilot, some questions were removed and some modified. During the second phase of the pilot, we administered final versions of the tests to a) *naive*

participants and to b) *fully informed* participants, who interacted with the simulation as in Experiment 2. However, when answering the tests, the latter group could use all the information from the simulation. More specifically, the simulation was available to them while they filled in the tests and they were instructed to use it. After finishing the tests, this group of participants was given a 15-30 minute break and then instructed to return to the tests and to the simulation. They were told to improve their answers so as to achieve the highest score possible. These participants' study backgrounds were similar to those in the participant sample for Experiment 2. Twenty-two naive participants completed the Retention test with the average score 18.8% of the maximum possible score (*SD* = 9%) while 12 fully informed participants achieved the average score 94.5% (*SD* = 3.1%). Forty-two naive participants (including naive participants who filled the Retention test) completed each three to five randomly picked questions from the Transfer test with the average score 9.4% (*SD* = 11%) and 12 fully informed participants (the same who filled in the Retention test) completed both versions of the Transfer test, i.e. 14 questions (8 + 6) in total, with the average score 69.4% (*SD* = 10.7%). The reason why naive participants did not complete the whole transfer test was that filling in transfer tests was found to be boring for these participants so we could not administer more questions to a single person due to the low-stake test problem. Still, each question in the final Transfer test was filled in by at least eight naive participants.

*Motivation Questionnaires*

The Motivation questionnaire 1, administered immediately after the treatment, asked participants to rate their self-perception of acquired knowledge using two questions (denoted as *Knowledge 1* and *Learn 1*); interest using one question (*Like 1*); and perception of difficulty of the learning from the materials using two questions (jointly denoted as *Hard*). The questions had 6-point Likert scale (*1 - very much / very good*; *6 - very little / very weak*). Three additional questions, irrelevant to the present study, were included in the questionnaire.

The Motivation questionnaire 2, administered a month after the treatment, asked participants to rate their self-perception of acquired knowledge using two questions (*Knowledge 2* and *Learn 2*); interest using one question (*Like 2*); and motivation using one question (*Motivation*). The former three questions were paired with appropriate questions from the Motivation questionnaire 1. The *Motivation* question was included only in the Motivation questionnaire 2, because we did not feel it appropriate to include it right after the treatment. The questions had 6-point Likert scale (*1 - very much / very good*; *6 - very little / very weak*). Frequency of drinking beer and other alcoholic beverages was measured on a 4-point scale ranging from "*1) never or less than once a year*" to "*4) more than once a week*". To keep the Pre-questionnaire reasonably short, we included these two questions in the Motivation questionnaire 2. Frequency of participants seeking out information or talking about about beer brewing was measured on a 4-point scale ranging from "*1) never*" to "*4) more than four times*". The former question was supplemented with the open-ended question: "If so, what information did you look for? .....". Five additional questions irrelevant to the present study were also included in the questionnaire.

The exact wording of questions is described in Appendix C.

*Graphing Test*

The test of graphing skills was administered a month after the treatment. We used Questions 1-4, 12-14, 16, 17 from the test of graphing in science (TOGS), originally intended for high school students ($7^{th}$ -$12^{th}$ grade) (McKenzie & Padilla, 1986). We chose the respective questions as those most closely reflecting the skills that participants would need to read the graphs/histograms in our simulation. Based on a pilot with 7 university students different from the study's participants, we set the completion time for the test to five minutes. We used the fixed time limit in Experiment 2 to make the test challenging enough for an audience older

than that of the original TOGS. (That turned out to have some limitations, as discussed in Section 5.2.3.)

### 5.1.4 Materials - the Intervention, the Apparatus

The participants used the simulation on the topic of beer brewing we developed as described in Sec. 3. The participants interacted with all the different parts: the tutorial, the linear part, the error part, and they had to complete 1 - 4 tasks. The simulation was run on notebooks or desktop PCs with at least 17"-wide screens. Each participant was seated at a separate computer. Each computer had two blank A4 sheets of paper and a pen in front of it.

### 5.1.5 Procedure

*The First Session.*

Each session started between 9 am and 10:40 am after all participants arrived and were seated.

First, the participants were told that they would interact at their own pace (for about 2-3 hours) with an educational simulation on the topic of beer brewing and fill in several questionnaires and tests. They were told that it would be possible to have short breaks during the experiment. The true purpose of the experiment was not revealed to the participants (until data from all participants were collected), but they were informed that we were investigating what they would learn from the simulation. They were also informed that the experiment would have two parts, the second about a month later, but they were not told what would happen during the second part.

Second, the participants were given informed consent and assigned numbers to keep their data anonymous. They were also asked to write down their nicknames, or their real names if they so wished, on a special list next to their numbers. The purpose of this list was to assign the participants the same numbers a month later.

Third, participants completed Pre-questionnaire at their own pace. The time to complete it was measured; as well as time to complete all other questionnaires and tests.

Fourth, for every P Group, participants were told the story of the family brewery as detailed in Sec. 3 (about 3 minutes). This part was missing for the N Groups. Then the simulation's interface was explained to them (about 2 minutes). As in Experiment 1, they were instructed, all together, to read carefully both the process as well as the tutorial instructions (see Sec. 3). They were told not to try and skip the process instructions, because the purpose of the simulation is to teach them the beer brewing process, which is described mainly therein. They were informed that they could make their own notes, if they preferred writing down information as part of the learning process (our pilots indicated that some participants would prefer to do so, while others would not).

Fifth, they interfaced with the tutorial at their own pace. Time to completion was measured, as well as time to completion for all the other simulation parts and tasks.

Sixth, when a participant finished, he/she started to interface with the linear part of the simulation at his/her own pace. After he/she finished, the participant was offered an optional break.

Seventh, the participant interacted with the error part of the simulation at his/her own pace. He/she was given the Flow questionnaire 1 and then the PANAS 1. After completing the questionnaires, the participant was offered an optional break.

Eighth, the participant was given a set of tasks. When solving a task, the participant could use his/her notes and the process instructions of the simulation (but the tutorial instructions were not displayed). When a task was finished, the next one was assigned immediately without any break. The last task was assigned, at the latest, during the 29$^{th}$ minute after the first task started. Help was offered to participants who had been solving tasks for more than 50 minutes

(3 cases), or who had been working on the first task for more than 40 minutes (2 cases). The tasks were always given in the same order and they were as follows:

1. Please brew 13-degree beer in the simulation environment.
2. Please brew 10-degree beer that contains 5-6% sugar.
3. Please brew 11-degree beer that is spoiled (contains acetone).
4. Please brew a drinkable 10-degree beer in less than 50 days.

Immediately after finishing the last assigned task, the participant completed the Flow questionnaire 2 and then the PANAS 2 (with the instruction that the questionnaires relate only to the task-solving phase). When the participant completed both the questionnaires, he/she was offered an optional break. The average time for all breaks taken during the whole session was 8.1 minutes across all participants (*SD*=5.6).

Finally, the test session started. Participants began by completing the Motivation questionnaire 1 at their own pace. Then the Retention test was given with the instruction to complete it within 7-8 minutes and to use all that time. Afterwards, six or eight questions of the Transfer test (depending on the test's variant) were presented; one at a time. Each question was typed on a separate A4 sheet of paper. The time allotted for completing a question ranged from 2 to 5 minutes, as indicated on each sheet of paper. Participants were instructed to complete their answers within that time limit, and to use the full time allotted. The participants could not return to previously answered questions. The order in which the questions were distributed was randomized. If a participant went overtime with the Retention test or a question from the Transfer test, the experimenter approached him/her to collect the test/question, but let the participant finish, if he/she had anything to add. If he/she insisted, the participant could also return the sheet with the answer before the time limit was up. Note that the time allotted for completing the tests was not fixed strictly, because a substantial amount

of handwriting was involved. It was supposed that participants write at different paces. Note also that we worried that participants would be too tired to answer the tests after 2-3 hours, but the pilot proved that this was not the case when breaks were offered to the participants as described above.

When the test session ended, we thanked the participants and asked them not to discuss details of this experiment with any person.

*The Second Session.*

The second session was usually conducted four weeks later and it lasted about an hour. Ninety percent of sessions started before noon. When all participants (1-4) in the session arrived, they were welcomed and seated (with the exception of a few late-comers). The purpose of the second session was explained to them. They were assigned the numbers given to them at the beginning of the first session and distributed questionnaires and tests in the order shown in the Table 5. All questionnaires, except for the graphing test, were completed at the participants' own pace. The graphing test was collected after 5 minutes (note that it contained multiple-choice or short-answer questions, thus the amount of handwriting was minimal). The course of the administration of the Retention and the Transfer test was the same as during the first session. Every participant was given a test version other than the one he/she received in the first session.

After completing the tests and questionnaires, a randomly-selected half of the participants performed yet another task embedded in the simulation environment: to brew 9-degree beer. They then underwent a brief interview about their experience when solving the task. The other half underwent a more complex, structured interview.

### 5.1.6 Scoring

Raw scores, or reverse scores if appropriate (Self-assessed knowledge of mathematics, Knowledge 1, Knowledge 2, Learn 1, Learn 2, Like 1, Like 2, Motivation, Hard), were used for all Pre-questionnaire questions and questions from both Motivation questionnaires, except for the eight questions related to participants' self-evaluation of a priori knowledge of beer brewing. The *beer brewing a priori knowledge score* (BB-a-priori score) was calculated according to the method described in Appendix D. The minimum score was 0, the maximum score was 32.

To elucidate Hypotheses 3 and 4, we created a priori one composite called *Gamers score* as follows: $(6 - M) + CG + 0.5L$, where $M$ is the true score of Self-assessed knowledge of mathematics, $CG$ is Frequency of playing computer games and $L$ is Frequency of LARPs playing[8].

Flow questionnaires were analyzed through T-norms provided within standardized Flow Short scale (Rheinberg, 2004). Concerning both scales of PANAS, raw scores will be reported. In the TOGS test, the participant was given one point for correctly answering each of the nine questions, giving us scale 0 .. 9, expressed in percentages in the remaining text.

A scorer unaware of the treatment condition scored both Retention tests according to an answer key prepared during pilots. For every question, the maximum possible score was 1 - 6 points, while the maximum possible score for the whole test was 31 points. The granularity was set to 0.5 point to enable assessment of partially correct answers. Recall that the

[8] We think it can be useful to remark that a slang term "Geek score" may capture the underlying meaning better than "gamers score," but we avoid it due to its negative connotations.

Retention test contained one open-ended question. The exact wording was not required in the answer on that question. Instead, the scorer assigned 1 point for every key "idea unit" out of 6 possible idea units (that were part of the answer key) or 0.5 point for partially correct "idea unit."

Two independent scorers unaware of the treatment condition scored independently both Transfer tests based on a key prepared during the pilots. The key contained important "idea units" for which 1 point would be rewarded and less important "idea units" for 0.5 point. Partially correct "idea units" were rewarded 0.5 or 0.25 points, respectively. The answers that could be derived based on common knowledge were not considered as useful "idea units," were not present in the answer key and therefore were rewarded 0 points. The disagreements between the scorers were resolved through consensus. The maximum possible score for a question was 1 - 6 points with the maximum possible score 17 points for the first variant of the whole Transfer test and 18 points for the second variant.

In the last phase of the experiment, participants had to complete several tasks. The *Task score* variable was calculated as the sum of a) the number of tasks started before the 30th minute after the start of the first task and b) evaluation of the quality of the beer produced by the tasks (based on the automatic assessment given by the simulation). Participants could be assigned up to four tasks and could be given 0 .. 3 points for the beer quality with the granularity 0.5, which gave us possible task score ranging from 0 to 7.

When reporting *Time spent on the simulation*, we used the total time spent on the tutorial plus the linear part plus the error part of the simulation, i.e. without breaks and without time needed for solving the tasks. An additional measure is *Questionnaire time*. It amounts to the sum of the times to complete the Pre-questionnaire, both PANASes and both Flow

questionnaires. These questionnaires involved substantial amount of reading but minimal amount of handwriting.

### 5.1.7 Data Analysis

Data were analyzed in statistical program R 3.0.0 (R Core Team, 2013). Differences between the P and the N group were tested using Welch's t-test (Welch,1947). We decided to use to Welch's approximation because of possibly unequal variances of groups. For paired wise comparisons in Sec. 5.2.4, we used paired t-test. The effect size for t-tests was expressed by Cohen's *d* and it was classified into negligible (Cohen's $d < 0.2$), small (Cohen's $d < 0.5$), medium (Cohen's $d < 0.8$) and large (Cohen's $d \geq 0.8$). Effect sizes for correlation coefficient were classified into none ($r < 0.1$), small ($r < 0.3$), medium ($r < 0.5$) and large ($r \geq 0.5$) (Cohen, 1988). All correlations were expressed by Pearson correlation coefficient. In Sections 5.2.3 and 5.2.4 we split participants into two groups divided by median and then compared means by Welch's t-test. This approach can be more informative about differences between groups than a simple correlation.

We had two versions of the Retention test and the Transfer test. In order to ensure comparability between the versions, we z-transformed scores from both versions of the tests so that we can omit influence of the test version on performance. This standardization could tell us how many standard deviations were scores of participants in each variant of the Retention and the Transfer test away from the sample mean, which was mapped to zero. The Immediate and the Delayed tests were transformed separately. We will use variables *Retention Test 1* and *2*, and *Transfer Test 1* and *2* to denote scores from the respective immediate (1) and delayed (2) tests (see Tab. 5). We will use variables *Retention Test* and *Transfer Test* to denote the average score a participant achieved in the respective immediate and delayed test, i.e. *Retention Test* = (*Retention Test 1* + *Retention Test 2*) / 2; *Transfer test* is

defined analogically. In Sec. 5.2.2, raw test scores will also be presented for illustrative purposes.

We analyzed influence of several variables yielded by the Motivation questionnaires, namely *Like* variable (the average value of Like 1 and Like 2), *Learnt* variable (the average value of Learnt 1 and Learnt 2), and *Hard* variable (the average value of both Hard questions from the Motivation questionnaire 1). Concerning the Flow and the PANAS, we also averaged data from both measurements, i.e. over the Flow 1 and the Flow 2, and over the PANAS 1 and the PANAS 2. We will denote positive scale of PANAS as *PANAS+* and negative scale as *PANAS–*. Concerning other questions, such as *Frequency of computer use*, we used raw scores, or reverse scores (see the beginning of Sec. 5.1.6).

## 5.2 Results

### *5.2.1 Participants' Characteristics*

We analyzed data from 75 participants in total. None of the participants had to be excluded due to incorrect data. Five participants did not come to post-test, so we excluded those data in pairwise test (e.g. correlation analysis). We could not obtain Task score for two participants due to technical problems, thus Task score is reported for 73 participants.

We compared participants' characteristics between the P and the N Group. As shown in Tab. 6, there were no significant differences, thus we can assume that the groups were sampled equally.

The Frequency of computer use approached ceiling and was excluded from the further analysis. Very low BB-a-priori score indicated that all of our participants had low prior knowledge.

The data showed that the P Group participants spent highly significantly more time on the simulation than the N Group participants (large effect size) and that the difference was evenly distributed among the three simulation parts: the tutorial, the linear part and the error part. There was no between-group difference in Time spent on the tasks, which was expected due to the experimental protocol (see Sec 5.1.5). The P Group participants covered significantly more A4 paper sheets with notes (medium effect size). Because the simulation contained a large amount of texts, we could assume that slowly reading participants interacted with the simulation longer than quickly reading participants. The speed of reading could be detected by measuring how long did it take a participant to complete questionnaires with multiple choice and short answer questions (as opposed to open-ended questions, which require the substantial amount of handwriting), as captured by Questionnaire time variable. Indeed, we found a strong positive correlation between Questionnaire time and Time spent on the simulation ($r$=0.50; $p$<.001; large effect size). However, no difference was found between Questionnaire time of the P and the N Groups; again suggesting that the groups were sampled equally.

--- Insert Table 6 about here ---

Table 6: Differences between the P Version (P) and the N Version (N) of the simulation. There is background information about participants in the first part of the table and intervention variables in the second part of the table. Significant differences are denoted by **bold**.

| | P | | N | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *t* | *df* | *p* | *d* |
| N | 36 | | 39 | | | | | |
| Age [years] | 22.61 | 4.02 | 22.05 | 1.92 | 0.76 | 49.3 | 0.451 | 0.18 |
| BB-a-priori-score | 5.56 | 3.39 | 5.11 | 2.91 | 0.61 | 69.1 | 0.543 | 0.14 |
| Freq. computer use | 3.78 | 0.42 | 3.85 | 0.43 | -0.69 | 72.8 | 0.490 | -0.16 |
| Freq. computer games playing | 1.69 | 0.79 | 1.67 | 0.87 | 0.15 | 73.0 | 0.885 | 0.03 |
| Freq. LARP playing | 2.67 | 1.33 | 2.90 | 1.39 | -0.73 | 72.9 | 0.465 | -0.17 |
| Freq. board games playing | 3.03 | 0.97 | 2.79 | 0.89 | 1.08 | 71.1 | 0.285 | 0.25 |
| Freq. beer drinking | 3.06 | 1.03 | 2.80 | 1.02 | 1.05 | 68.0 | 0.298 | 0.25 |
| Freq. alcohol drinking | 2.91 | 0.82 | 2.63 | 0.73 | 1.54 | 67.2 | 0.128 | 0.37 |
| **Time spent on the simul.** | 112.31 | 21.90 | 93.03 | 19.35 | 4.03 | 70.1 | **<0.001** | **0.94** |
| **Time spent on the tutor.** | 16.30 | 3.99 | 13.23 | 4.51 | 2.85 | 72.9 | **0.006** | **0.66** |
| **Time spent on the linear part** | 46.97 | 12.19 | 38.92 | 10.30 | 3.80 | 68.8 | **0.003** | **0.72** |
| **Time spent on the errors** | 49.31 | 9.99 | 40.87 | 8.55 | 3.91 | 69.2 | **<0.001** | **0.91** |
| Time spent on the tasks | 38.56 | 7.20 | 38.08 | 5.83 | 0.31 | 68.0 | 0.755 | 0.07 |
| Questionnaire time | 19.11 | 4.75 | 18.13 | 3.56 | 1.01 | 64.7 | 0.318 | 0.24 |
| **Nr. of A4 pages written** | 1.17 | 0.92 | 0.56 | 0.72 | 3.13 | 64.0 | **0.003** | **0.74** |
| Nr. of tasks completed | 2.14 | 0.64 | 2.15 | 0.74 | -0.09 | 72.6 | 0.926 | -0.02 |

### *5.2.2 Does Personalization Promote Learning?*

As can be seen in Table 7, we found between-group difference neither in Transfer nor in Retention test variables, nor in the immediate–delayed test score differences. There was also no between-group difference in Task score. For illustrative purposes, Figures 3 and 4 show raw scores from all tests, including naive and fully informed participants (see Sec. 5.1.3).

--- Insert Table 7 about here ---

Table 7: Differences in performance between the P Version (P) and the N Version of the simulation. Data are given in z-scores except for Task score.

| | P | | N | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *t* | *df* | *p* | *d* |
| Retention Test 1 | 0.09 | 0.82 | -0.08 | 1.13 | 0.73 | 69.3 | 0.468 | 0.17 |
| Transfer Test 1 | -0.04 | 0.96 | 0.04 | 1.03 | -0.36 | 73.0 | 0.719 | -0.08 |
| Retention Test 2 | 0.05 | 0.94 | -0.05 | 1.06 | 0.43 | 67.0 | 0.665 | 0.10 |
| Transfer Test 2 | -0.08 | 1.05 | 0.08 | 0.94 | -0.65 | 67.2 | 0.518 | -0.16 |
| Diff Retention | 0.03 | 0.74 | -0.03 | 1.22 | 0.23 | 56.0 | 0.819 | 0.05 |
| Diff Transfer | 0.06 | 0.90 | -0.06 | 1.10 | 0.53 | 65.5 | 0.600 | 0.13 |
| Task score | 4.94 | 0.76 | 4.75 | 1.15 | 0.85 | 64.8 | 0.398 | 0.20 |

--- Insert Figure 3 about here ---

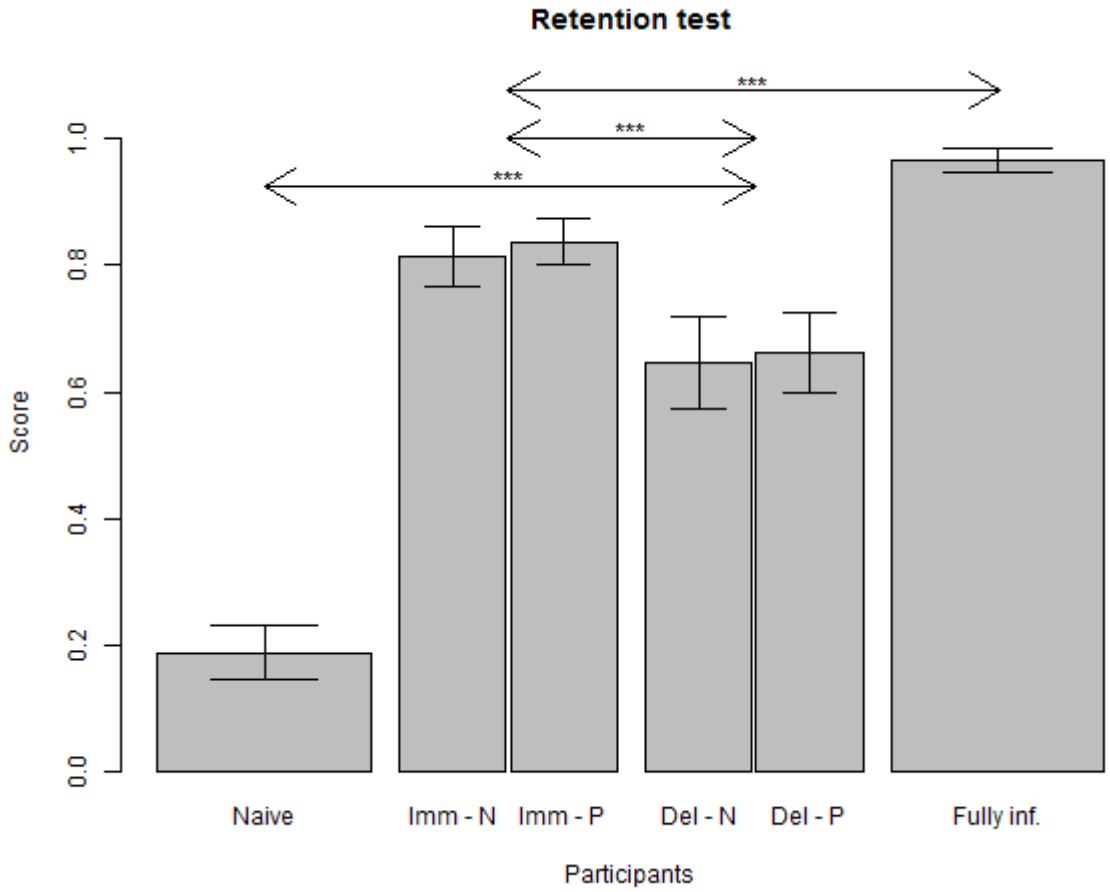


Figure 3: Means of raw scores from the Retention test expressed as a fraction of the maximum possible score. Means of the naive and the fully informed participants' scores as well as means of scores of the experimental participants achieved in the Immediate and the Delayed tests are depicted. Error bars represent standard errors of the mean. Significance is denoted by stars.

--- Insert Figure 4 about here ---

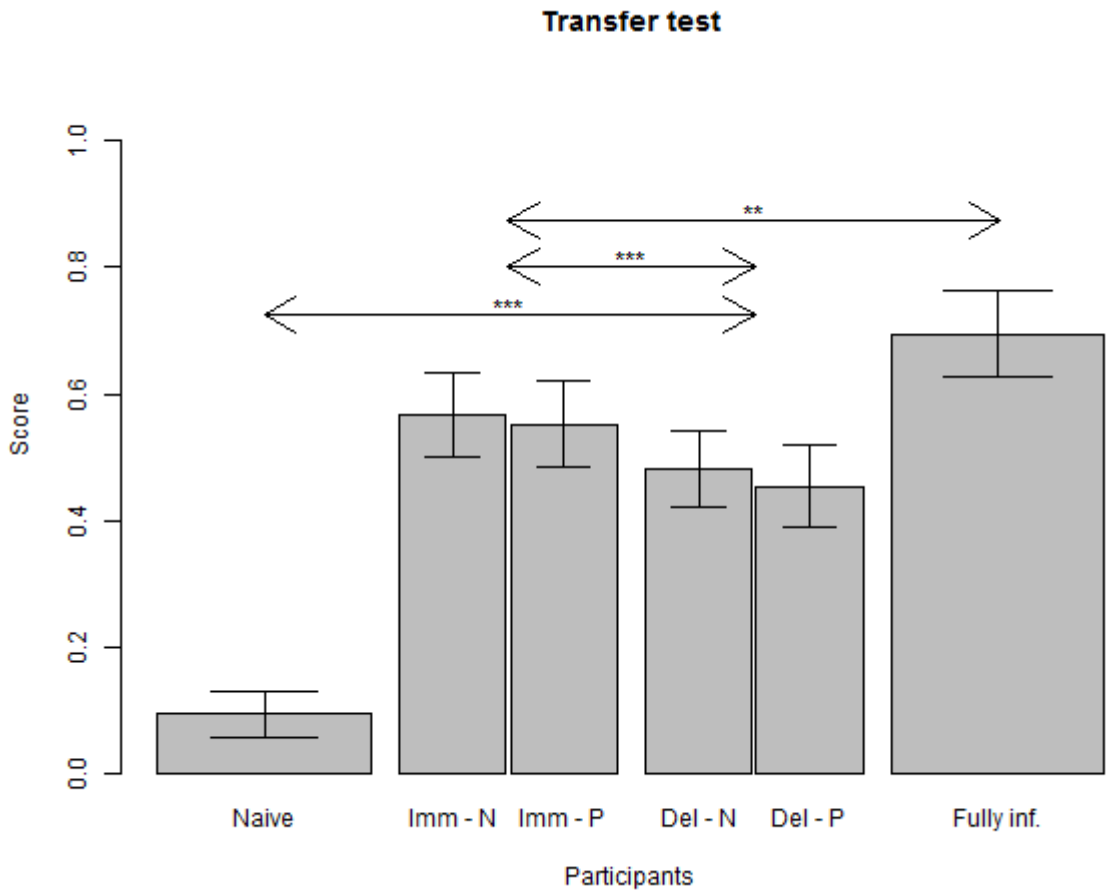


Figure 4: Means of raw scores from the Transfer test expressed as a fraction of the maximum possible score. Means of the naive and the fully informed participants' scores as well as means of scores of the experimental participants achieved in the Immediate and the Delayed tests are depicted. Error bars represent standard errors of the mean. Significance is denoted by stars.

Participants could seek some information between the main experiment and the post-test. Only 11 participants, evenly spread in the two conditions, sought for information (all of them

"once or twice"). Qualitative data indicated that five of them sought for information not contained in our simulation (mostly related to "brewing different types of beer"), four for possibly relevant information (e.g., "home-brewing") and two did not explain themselves. There was also no between-group difference in Frequency of talking about beer brewing between the main experiment and the post-test ($t(67.9)=-0.6$; $p=.551$; $d=-0.14$). Thus, we omit these factors as covariate.

Inspection of the relationship between the test scores showed that there was a strong correlation between Retention and Transfer test variables ($r=0.60$; $p<.001$; large effect size) as well as correlation between these variables and Task score ($r=0.41$; $p<.001$ for Retention test; $r=0.49$; $p<.001$ for Transfer test; medium effect sizes). Especially the latter correlation indicates that knowledge of the beer brewing mental model is manifested both in the real task performance as well as in the Transfer test score. For that reasons and also due to brevity, we will not report Task score data in the remaining text but data related to the Transfer test scores only (and the Retention test scores data too since it can be assumed that the Retention tests also capture factual knowledge beyond the mental model).

While Time spent on the simulation was correlated with Questionnaire time as already shown in Sec. 5.2.1, it was correlated neither with Retention test variable ($r=-.07$; $p=.50$) nor with Transfer test variable ($r=-.12$; $p=.30$). That suggests that the longer time spent on the intervention may not necessarily lead to better knowledge acquisition.

The fact that we found no between-group difference in the test scores could be attributed to the floor or the ceiling effect. Therefore, we compared the raw tests scores of the experimental participants to the raw scores achieved by the naive and the fully informed participants, as introduced in Sec. 5.1.3 (Fig. 3, 4). The difference between the naive participants and the

experimental participants was significant on both the Delayed retention test ($t$(74.96)=15.05; $p$<.001; $d$=2.61) and the Delayed transfer test ($t$(109.54)=13.33; $p$<.001; $d$=2.34). The difference between the fully informed participants and the experimental participants was significant on the Immediate retention test ($t$(72.73)=8.03; $p$<.001; $d$=1.15) as well as the Immediate transfer test ($t$(25.99)=3.45; $p$=.002; $d$=0.69). We remark that the score difference between the Immediate and the Delayed tests is also significant for both tests (Retention: $t$(69)=9.23; $p$<.001; $d$=1.03, Transfer: $t$(69)=5.24; $p$<.001; $d$=0.48).

Finally, all the P Group participants who undergone the final interview except of one remembered the story about the grandpa, suggesting they were not oblivious to the story. This fits well with the results of Experiment 1. Roughly one third of the P Group participants had positive comments ("the story was motivating", "the application would be mildly worse for learning without the grandpa", etc.), one third had neutral comments ("[the simulation] would be the same without the grandpa" etc.) and one third had no comments regarding whether the simulation would be better or worse without the grandpa. No-one had a negative comment; the most negative comment was: "I was aware of the grandpa, but he was irrelevant [for the purpose of learning], I don't think [his story] influenced me in any way."

### *5.2.3 Does a Higher Mathematical Knowledge or the Frequency of Playing Games Lead to a Better Mental Models Acquisition?*

To elucidate Hypotheses 3 and 4, we inspected relations between test scores and Self-assessed mathematical knowledge (Math score), Frequency of computer games playing, LARP playing and board/card games playing, TOGS score, Self-perceived ability of acquiring mental models (Mental models score), and Gamers score composite. Recall that Gamers score had been formulated a priori before the experiment started.

First, differences in all the described characteristics were non-significant between the P and the N Conditions; as can be seen in Table 8.

--- Insert Table 8 about here ---

Table 8: Differences between the P Version (P) and the N Version (N) of the simulation in key characteristics.

| | P | | N | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *t* | *df* | *p* | *d* |
| Math score | 3.92 | 1.57 | 3.74 | 1.43 | 0.51 | 70.4 | 0.609 | 0.12 |
| Freq. computer games playing | 1.69 | 0.79 | 1.67 | 0.87 | 0.15 | 73.0 | 0.885 | 0.03 |
| Freq. LARP playing | 2.67 | 1.33 | 2.90 | 1.39 | -0.73 | 72.9 | 0.465 | -0.17 |
| Freq. board games playing | 3.03 | 0.97 | 2.79 | 0.89 | 1.08 | 71.1 | 0.285 | 0.25 |
| TOGS score | 0.76 | 0.24 | 0.77 | 0.24 | -0.06 | 68.0 | 0.956 | -0.01 |
| Mental models score | 6.17 | 1.46 | 5.82 | 1.71 | 0.94 | 72.6 | 0.349 | 0.22 |
| Gamers score | 5.94 | 2.02 | 5.83 | 2.17 | 0.24 | 72.0 | 0.813 | 0.06 |

As can be seen in Table 9, significant differences between participants with Low and High Transfer test scores were achieved in Frequency of playing computer games, Mental models score, TOGS score and Gamers score. The effect sizes were in medium to large range. Differences in Math score are only marginally significant; however, we also inspected differences in the mean scores of the moderator variables between the Low and the High groups when the groups were formed using the Immediate transfer test score only vs. the Delayed transfer test scores only (as opposed to their averages; as depicted in Table 9 and

described in Sec. 5.1.7). For all but one moderator variables, the Low/High-group differences were similar in all the three cases. The notable exception is Math score: the differences in Math score between the High and the Low performing groups in the Immediate transfer test are significant ($t$(70.23)=2.46; $p$=.016; $d$=0.57) as well as analogical differences when the High and the Low groups were formed using the Delayed transfer tests ($t$(66.9)=3.05; $p$=.003; $d$=0.73).

--- Insert Table 9 about here ---

Table 9: Differences between the high scoring and the low scoring participants in the Transfer test in several characteristics. The High and Low groups were formed using Transfer test variable. Significant differences are denoted by **bold** and trends by *italic*.

| Transfer test | High | | Low | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *t* | *df* | *P* | *d* |
| *Math score* | 4.24 | 1.60 | 3.60 | 1.35 | 1.78 | 64.6 | *0.080* | *0.43* |
| **Freq. computer games playing** | 2.03 | 0.89 | 1.43 | 0.65 | 3.21 | 62.4 | **0.002** | **0.77** |
| Freq. LARP playing | 3.00 | 1.35 | 2.74 | 1.31 | 0.81 | 67.9 | 0.422 | 0.19 |
| Freq. board games playing | 3.00 | 0.80 | 2.89 | 0.96 | 0.54 | 65.9 | 0.592 | 0.13 |
| **TOGS score** | 0.87 | 0.17 | 0.66 | 0.26 | 4.01 | 58.3 | **0.002** | **0.96** |
| **Mental models score** | 6.49 | 0.74 | 5.69 | 1.84 | 2.38 | 44.7 | **0.022** | **0.57** |
| **Gamers score** | 6.75 | 2.15 | 5.40 | 1.67 | 2.90 | 62.2 | **0.005** | **0.70** |

While we did not expect analogical differences concerning the Retention test, we surprisingly found the same relation between TOGS score and Frequency of playing computer games, on the one hand, and Retention test variable on the other (Table 10). However, other differences

were not significant as expected. (The difference in Math score was also not significant when only the Immediate or Delayed version of the Retention test was considered.)

--- Insert Table 10 about here ---

Table 10: Differences between high scoring and low scoring participants in the Retention test in several characteristics. The High and Low groups were formed using Retention test variable. Significant differences are denoted by **bold** and trends by *italic*.

| Retention test | High | | Low | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *t* | *df* | *p* | *d* |
| Math score | 4.21 | 1.45 | 3.63 | 1.52 | 1.62 | 67.0 | 0.111 | 0.39 |
| **Freq. computer games playing** | 1.94 | 0.91 | 1.51 | 0.70 | 2.21 | 64.0 | **0.031** | **0.53** |
| Freq. LARP games playing | 2.77 | 1.29 | 2.97 | 1.38 | -0.63 | 67.6 | 0.533 | -0.15 |
| Freq. board games playing | 2.97 | 0.82 | 2.91 | 0.95 | 0.27 | 66.6 | 0.788 | 0.06 |
| **TOGS score** | 0.86 | 0.20 | 0.67 | 0.24 | 3.71 | 65.7 | **<0.001** | **0.89** |
| Mental models score | 6.31 | 1.11 | 5.86 | 1.72 | 1.32 | 58.0 | 0.191 | 0.32 |
| *Gamers score* | 6.51 | 1.96 | 5.63 | 2.02 | 1.85 | 67.0 | *0.069* | *0.44* |

For exploratory purposes, we report the correlation matrix for characteristics in Tables 9 and 10 (Table 11). Correlations are not particularly strong, yet there is a significant correlation among all of the following variables: Math score, TOGS score and Mental models score, suggesting existence of a common denominator. Significance or trend is also achieved among

these variables and Frequency of playing computer games. Relatively low correlation between Math score and TOGS score can be, at least partially, explained by our informal observation that some participants with study backgrounds in computer science or physics who reported high self-assessed mathematical skills did not score well in the test of graphing due to the time limit to complete that test. These participants were "slow readers," because their Questionnaire time was high. Additionally, the mean of Mental models score is relatively high (5.98 on scale 1-7), making this questions less sensitive due to approaching the ceiling effect.

--- Insert Table 11 about here ---

Table 11: Correlation matrix of Mathematical score, Frequency of playing computer games, Frequency of LARP playing, Frequency of playing board/card games, TOGS score, Mental models score and Gamers score composite. Significant differences ($\alpha$=0.05) are highlighted in **bold**. Recall that Gamers score is a composite of Mathematical score, Frequency of playing computer games and Frequency of LARP playing.

| | Math | Games | LARP | Board | TOGS | Men. Models | Gamers score |
|---|---|---|---|---|---|---|---|
| Math score | - | | | | | | |
| Freq. comp. games pl. | **0.26** | - | | | | | |
| Freq. LARP playing | 0.03 | **0.23** | - | | | | |
| Freq. board games pl. | 0.08 | 0.07 | **0.39** | - | | | |
| TOGS score | **0.26** | 0.20 | -0.04 | 0.08 | - | | |
| Mental models score | **0.32** | **0.28** | **0.33** | 0.20 | **0.24** | - | |
| Gamers score | **0.83** | **0.66** | **0.44** | 0.21 | **0.46** | **0.26** | - |

Interestingly, a higher Gamers score predicts a higher Immediate transfer test score, but not vice versa (Fig. 5). This relation is similar when the Delayed transfer test is considered instead of the Immediate transfer test. That comforts us that a complex mental model can be mastered even by people who do not regularly play games and/or self-assess their mathematical knowledge high.

--- Insert Figure 5 about here ---

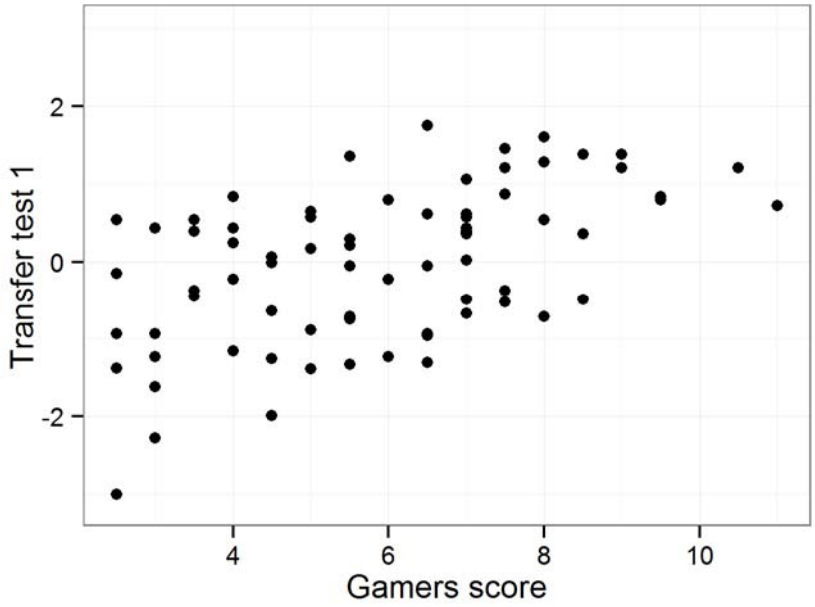


Figure 5: The relation between Gamers score and Transfer test 1 variable. Note there are nearly no points below the diagonal.

### *5.2.4 Do Affective Variables Relate to Learning Outcomes?*

Variability in learning performance can be explained by differences in motivational characteristics. However, as summarized in Tab. 12, none of the measured characteristics differed significantly between the groups.

--- Insert Table 12 about here ---

Table 12: Differences between the P Version (P) and the N Version (N) of the simulation in motivational characteristics.

| | P | | N | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *t* | *df* | *p* | *d* |
| Like | 5.11 | 0.65 | 5.17 | 0.75 | -0.34 | 66.8 | 0.734 | -0.08 |
| Learnt | 4.94 | 0.85 | 4.90 | 0.86 | 0.21 | 68.0 | 0.835 | 0.05 |
| Motivation | 5.00 | 0.94 | 4.69 | 0.90 | 1.43 | 67.9 | 0.158 | 0.34 |
| Flow | 55.43 | 7.06 | 55.71 | 8.12 | -0.16 | 72.7 | 0.876 | -0.04 |
| PANAS+ | 32.89 | 6.73 | 31.26 | 7.28 | 0.99 | 71.0 | 0.326 | 0.23 |
| PANAS– | 14.24 | 4.00 | 13.86 | 3.89 | 0.42 | 70.2 | 0.676 | 0.10 |
| Hard | 2.86 | 0.85 | 2.67 | 0.89 | 0.97 | 72.9 | 0.337 | 0.22 |

Participants' self-assessed knowledge of beer brewing was small before the intervention (Q6 question; *Mean*=1.78; *SD*=0.78) but it substantially increased after the intervention (Knowledge 1 question; *Mean*=4.36; *SD*=0.86). It decreased after a month but still remained high (Knowledge 2 question; *Mean*=3.51; *SD*=1.11).

The relation between the affective dimension and learning was investigated using the median split technique. Several affective variables were found to relate to Retention (Tab. 13) and Transfer (Tab. 14) test variables. In particular, concerning both tests, high scoring participants were more often in the flow state, they liked the simulation more, their PANAS+ score was higher, they thought they learnt more and they found the simulation easier. All differences were at least marginally significant and most effect sizes were in the medium to large range.

On the other hand, no differences were found concerning PANAS–, indicating that a low performance was not connected to distress or unpleasurable experience.

--- Insert Table 13 about here ---

Table 13: Differences between high scoring and low scoring participants in the Retention test in motivational characteristics. The High and Low groups were formed using Retention test variable. Significant differences are depicted using **bold** and trends using *italic*. Note that the lower the score of Hard question was, the less difficult the simulation was.

| Retention test | High | | Low | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *t* | *df* | *p* | *d* |
| **Like** | 5.31 | 0.69 | 4.97 | 0.67 | 2.11 | 68.0 | **0.039** | **0.50** |
| **Learnt** | 5.30 | 0.66 | 4.54 | 0.86 | 4.14 | 63.5 | **<0.001** | **0.99** |
| **Motivation** | 5.14 | 0.88 | 4.54 | 0.89 | 2.84 | 68.0 | **0.006** | **0.68** |
| *Hard* | 2.56 | 0.75 | 2.96 | 0.97 | -1.94 | 63.9 | *0.057* | *-0.46* |
| **Flow** | 58.39 | 5.91 | 53.31 | 7.31 | 3.19 | 65.1 | **0.002** | **0.76** |
| **PANAS+** | 34.35 | 6.57 | 30.63 | 6.65 | 2.32 | 65.8 | **0.024** | **0.56** |
| PANAS– | 13.70 | 3.57 | 14.07 | 4.23 | -0.40 | 65.2 | 0.694 | -0.10 |

--- Insert Table 14 about here ---

Table 14: Differences between high scoring and low scoring participants in the Transfer test in motivational characteristics. The High and Low groups were formed using Transfer test variable. Significant differences are depicted using **bold** and trends using *italic*.

| Transfer test | High | | Low | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *t* | *df* | *p* | *d* |
| **Like** | 5.33 | 0.55 | 4.96 | 0.78 | 2.30 | 61.4 | **0.025** | **0.55** |
| **Learnt** | 5.16 | 0.78 | 4.69 | 0.86 | 2.40 | 67.4 | **0.019** | **0.57** |
| *Motivation* | 5.03 | 0.95 | 4.66 | 0.87 | 1.70 | 67.5 | *0.094* | *0.41* |
| **Hard** | 2.44 | 0.75 | 3.07 | 0.90 | -3.18 | 65.7 | **0.002** | **-0.76** |
| **Flow** | 58.51 | 6.29 | 53.19 | 6.89 | 3.38 | 67.4 | **0.001** | **0.81** |
| *PANAS+* | 34.00 | 6.78 | 30.96 | 6.63 | 1.87 | 65.6 | *0.066* | *0.45* |
| PANAS– | 13.65 | 3.75 | 14.11 | 4.08 | -0.49 | 66.0 | 0.628 | -0.12 |

Due to Hypothesis 3 and 4, we also studied how these factors differ between participants with low and high Gamers score (Tab. 15). The high scoring participants were more often in the flow state and they found the simulation easier; with effect sizes in the medium to large range. Somewhat smaller, marginally significant difference was also found for Like question.

--- Insert Table 15 about here ---

Table 15: Differences between participants with high and low Gamers score in motivational characteristics. Significant differences are depicted using **bold** and trends using *italic*.

| Gamers score | High | | Low | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *t* | *df* | *p* | *d* |
| *Like* | 5.30 | 0.60 | 4.97 | 0.77 | 1.99 | 62.3 | *0.052* | *0.48* |
| Learnt | 4.99 | 0.97 | 4.87 | 0.72 | 0.57 | 62.6 | 0.568 | 0.14 |
| Motivation | 4.94 | 0.94 | 4.74 | 0.93 | 0.92 | 67.0 | 0.360 | 0.22 |
| **Hard** | 2.43 | 0.70 | 3.09 | 0.89 | -3.58 | 70.8 | **<0.001** | **-0.82** |
| **Flow** | 57.86 | 7.26 | 53.55 | 7.46 | 2.51 | 71.5 | **0.014** | **0.58** |
| PANAS+ | 32.56 | 6.38 | 31.50 | 7.71 | 0.63 | 68.8 | 0.528 | 0.15 |
| PANAS– | 13.84 | 4.11 | 14.31 | 3.80 | -0.50 | 68.7 | 0.618 | -0.12 |

For exploratory purposes, we also report the correlation matrix for all motivational characteristics (Table 16). The strongest correlation is for PANAS+ with Motivation, Flow, Like and Learnt variables (medium to large effect). Flow score correlates with all other characteristics (medium to large effect). That suggests that the affective variables share a common denominator. PANAS– is largely orthogonal to PANAS+ (Watson et al., 1988), therefore, it is not surprising we found no correlation between PANAS– and PANAS+, and subsequently, between PANAS– and Like/Motivation variables. The negative correlation between Hard variable and Flow score is also not surprising (Engeser & Rheinberg, 2008). It is consistent with this last result that we found negative correlation between PANAS– and Flow score; indicating that participants with a high PANAS– score probably did not experienced the flow state (and found the simulation difficult).

--- Insert Table 16 about here ---

Table 16: Correlation matrix of motivational characteristics. **Bold** values are significant for α=0.05. Scores from the Flow questionnaire correlate with all other characteristics (moderate to high effect). Note that the lower the score of Hard question was, the less difficult the simulation was.

| | Like | Learnt | Mot. | Hard | Flow | PANAS + | PANAS – |
|---|---|---|---|---|---|---|---|
| Like | - | | | | | | |
| Learnt | **0.55** | - | | | | | |
| Motivation | **0.39** | **0.57** | - | | | | |
| Hard | -0.18 | -0.12 | -0.07 | - | | | |
| Flow | **0.55** | **0.38** | **0.40** | **-0.40** | - | | |
| PANAS+ | **0.69** | **0.52** | **0.51** | -0.07 | **0.57** | - | |
| PANAS– | -0.18 | **-0.25** | -0.17 | **0.35** | **-0.49** | -0.17 | - |

As reported in Sec. 5.2.1, the P Group participants wrote significantly more pages of A4 paper sheets than the N Group participants. The number of A4 paper sheets covered with notes strongly correlates with Time spent on the simulation ($r$=0.59; $p$<.001), suggesting that the P Version could make the participants to learn harder. However, that did not result in a better learning outcome.

## *5.3 Discussion*

Experiment 2 was the main experiment of the present study and it investigated the impact of personalized instructions on the acquisition of a complex mental model by means of an educational simulation (on the topic of beer brewing) that it took about 2 - 3 hours to complete. In particular, the experiment investigated a) actual learning outcomes and how they relate to the participants' "motivation" (Goal 1), b) lasting effects of the personalization in

terms of transfer test achievements (Goal 2), and c) whether mathematical knowledge and frequency of game playing is related to learning outcomes (Goal 3). We intentionally focused on a *long-lasting* intervention and on a *complex* mental model acquisition because the personalization effect was repeatedly demonstrated in the past in experiments with short treatments (up to 30 minutes) aiming at a simple mental model acquisition.

### 5.3.1 Goal 1: Application of the Motivation → Learning Framework to the Personalization Principle

The differences between the P Group and the N Group in achievement in both the Transfer and the Retention test were negligible (Tab. 7). We also found no between-group difference in any of the affective variables (Tab. 12). Yet when the participants' test scores were compared to the scores of naive participants, we clearly see the improvement (Fig. 3, 4). The participants also felt they learnt a lot (Sec. 5.2.4). Thus, it follows that our data does not support Hypothesis 1, which states that the P Group participants would outperform the N Group participants and, at the same time, the P Group participants would be more motivated (i.e., $P_{trans} > N_{trans}$ & $P_{mot} > N_{mot}$).

We directly found consistent, moderate to high positive relation between affective variables and achievements in both of the tests (Tab. 13, 14). That strengthens our confidence in the motivation → learning framework and we need not hesitate to use it to explicate the study's findings[9]. In terms of this framework, the straightforward conclusion seems to be that the

[9] Mind however, that strictly speaking, we cannot determine, based on our data, if motivation increases learning gains, or vice versa, or if there is a mutual influence.

personalization, in the case of a 2-3 hours long simulation experience, brings no motivational benefits and thus the between-group learning gain differences are negligible, which is Case 5 from Sec. 2.1 (i.e., $P_{trans} = N_{trans}$ and $P_{mot} = N_{mot}$). The "complication" to this straightforward conclusion is that the P Group participants spent voluntarily about 20 % (about 19 minutes) more time on the simulation than the N Group participants, which is a large effect size (Tab. 6), yet that difference did not result in better learning outcomes. Why did the simulation take longer to complete? Because the texts of the P Version are less than 10 % longer than the N Version's texts (Sec. 3.3) and because the participants spent less than 50 % of time reading these texts (based on our informal observation), it seems that less than 1/4 of the difference can be attributed to longer reading. The rest of the difference can be partly explained by the fact that the P Group participants made, on average, about 0.5 A4 paper sheets more notes than the N Group participants (Tab. 6, Sec. 5.2.4); but it seems unlikely that making 0.5 A4 paper sheets of notes lasted the whole remaining part of the difference, nearly 15 minutes. Therefore, at least part of the difference was probably caused by repeating the simulation steps or running the simulation more slowly (recall that the current simulation's phase could be restarted and that the simulation speed could be adjusted – see Sec. 3.2 and Appendix A). That, including the difference in the amount of notes, could be attributed to a higher *carefulness* of the P Group participants caused by the personalization, a difference that could escape detection by our questionnaires. If a higher carefulness was really the culprit remains a question for future work; but if it was, it seems probable that any positive effect that it could have had must have been outweighed by distraction due to the personalization (compare this with Case 3 from Sec. 2.1 and also Fig. 1b). Indeed, the longer time needed to complete the P treatment could have been also partly caused because participants devoted part of their cognitive capacity to thinking about the grandpa's story. Alternatively, they could have been

distracted by a different aspect of the personalization; for instance, by filler conversational formulations (Point (6), Appendix B).

For these reasons, our first conclusion is that the personalization principle not only may not improve learning but it could be, in some conditions, detrimental to learning. This does not agree well with the majority of past results. This outcome makes investigation of boundary conditions of the personalization principle an important endeavor. We will return to this point in the general discussion in Sec. 6.

### 5.3.2 Goal 2: Investigation of Lasting Effects of the Personalization Principle

The between-group differences in achievement in the Delayed transfer test were negligible as well as differences in the decrement between achievement in the Immediate and the Delayed transfer tests (Tab. 7; Fig. 4). At the same time, participants were able to reconstruct substantial amount of information a month after the intervention (Fig. 4), suggesting a presence of a residual mental model in their minds. In other words, the results were not caused by the floor effect. On average, the participants talked about the topic "once or twice" between the first and the second testing session, and 11 participants (16%) sought "once or twice" for additional information between the sessions, suggesting their residual mental model could have been slightly improved due to memory rehearsing. However, there was no difference between the conditions concerning participants' information seeking behavior (Sec. 5.2.2). Considering these things together, Hypothesis 2 is not supported by our data; we found no differences in favor of the P Group in long-term.

### 5.3.3 Goal 3: Investigation of What Personal Characteristics Moderate Learning and Motivational Outcomes

*Mathematical abilities and mental models.*

Participants with higher Self-assessed mathematical skills, a higher TOGS score and a higher Self-assessed ability of acquiring mental models were more often in the group of participants scoring high in the Transfer test (Tab. 9, Sec. 5.2.3). The effect sizes were predominantly in medium to large range. There was also small to moderate correlation between these characteristics (Tab. 11), suggesting existence of a common denominator. Thus, Hypothesis 3 is supported by our data; at least in our case, participants with higher mathematical abilities were able to acquire a complex mental model using a complex simulation better than participants less able in mathematics.

Concerning the Retention test, the same difference was found only for the TOGS score (Tab. 10). For the TOGS score, we also see the largest effect size concerning both the tests (Tab. 9, 10). This may be explained as follows: mathematical abilities, in general, can be useful for general acquisition of complex mental models, but not necessarily for learning facts. In our case, higher graphing skills were also particularly useful for learning facts because reading graphs constituted a considerable portion of the simulation experience. All facts could be studied using the textual instructions only, but some facts comprising numbers, such as recommended levels of various beer constituents, could be strengthened in the memory when reading graphs/histograms.

*Playing games.*

Frequent computer game players were more often in the group of participants scoring high in the Transfer test than less frequent players; the effect size of the difference was medium (Tab. 9). Only negligible differences were found for frequent players of experiential simulation/tabletop role-playing games and board/card games. Thus, Hypothesis 4 is partly supported: those who play computer games often acquire a complex mental model better

using a complex educational simulation than participants who play computer games rarely or never (expected); the same does not seem to be the case for experiential simulation/tabletop role-playing games (not expected) and board/card games (expected).

Gamers score proved to be a useful construct (Tab. 9, 10, 15, Fig. 5). Importantly, it helped to reveal the "one-way" prediction that people with a high Gamers score would score high in the Transfer test, but not vice versa (Fig. 5). This relationship was only partly apparent when Gamers' score constituents were considered alone. A supplementary exploratory analysis indicated that this relationship might be caused because people with a high Gamers score found the simulation easier and to a somewhat lesser extent also likeable (Tab. 15), but more research would be needed to confirm this hypothesis. However, this construct is not without limitations. Recall that we *a priori* predicted that Frequency or LARP playing would have the smallest impact as concerns prediction of the actual achievement (see Sec. 5.1.6 for how the composite was computed); it turned out that the prediction value of this question was even smaller than we had anticipated (Tab. 9). Thus, the LARP sub-question did not contribute much to the Gamers score composite. Exploratory Principal Component Analysis may help to refine this for the purposes of future research.

# 6. General Discussion

The purpose of this study was to investigate the personalization effect in a new context; namely, within a 2-3 hour-long treatment, whose educational objective was the acquisition of a complex cause-and-effect mental model. Our underlying theoretical framework was the cognitive-affective theory of learning with media (CATLM, Moreno, 2005; Moreno & Mayer, 2007). Based on this theory and also based on past research results (Moreno & Mayer, 2000; Mayer et al., 2004; Moreno & Mayer, 2004; Günizi, 2010), we derived our main hypothesis

that personalization would motivate students to invest more of their cognitive capacity into processing of the learning materials and that distraction due to personalization would be minimal, leading to better learning. Based on past results from game-based learning literature (summarized in Brom et al., 2011; see also Wouters et al., 2013), we also predicted that the differences would be more pronounced in favor of the personalization condition a month after the treatment administration.

Our secondary interest was to investigate whether a relationship exists between learning outcomes, measured by transfer tests, on the one hand, and mathematical abilities and frequency of game playing on the other. This interest was motivated by the fact that knowledge about what user characteristics can determine learning outcomes is limited in the context of simulation/game-based learning (e.g., Tobias et al., 2011).

## *6.1 A Failure to Replicate the Personalization Effect*

### 6.1.1 Our Findings and Related Findings

Despite the fact that participants chose the personalized version of the simulation, when demonstrated both versions, each for 15-20 minutes, during Experiment 1, we failed to replicate the personalization effect in Experiment 2. There were no between-group differences in transfer test scores, nor in retention test scores.

Importantly, to our best knowledge, past research only demonstrated the personalization effect in treatments shorter than about 30 minutes (Moreno & Mayer, 2000; Mayer et al., 2004; Moreno & Mayer, 2004; Günizi, 2010). All of these had focused on mental models acquisition. Doolittle (2010) failed to replicate it both using a 2.5-hour-long tutorial that

taught high-level skills as well as in a complementary 3-minute-long narrated animation teaching high-level skills.

However, the failure to replicate the personalization effect might not be caused only by the longer treatment; there are more possible causes. An important piece of auxiliary evidence that we should consider comes from research investigating the benefits of using polite as opposed to direct language within educational material, a so-called "politeness principle." This principle can be considered as partly overlapping the personalization principle (see Appendix B, Point 5). A study of Wang et al. (2008) showed that university students who used a polite virtual tutor achieved higher scores than students using a direct tutor, within an engineering, on-line learning system, for a period of about 36 minutes. This difference was caused mostly by students without engineering backgrounds and with average computer skills (rather than students with engineering backgrounds and above average computer skills; but the total sample was only 37 students). McLarren et al. (2011a) showed that college students with low prior knowledge who learned to solve stoichiometry problems in chemistry with a polite web-based tutor, outperformed in problem-solving tests students learning with a tutor that used direct language. At the same time, the data showed a reverse trend when high prior knowledge students were considered (the treatment probably lasted 1-2 hours). This pattern is usually called an expertise-reversal effect. However, McLarren's team did not replicate these findings in a study with high school participants situated in a real classroom as opposed to a laboratory: there were no differences between the low prior or high prior knowledge students (McLarren et al., 2011b).

In the context of these three studies, it is important to highlight two things. First, all our students can be considered to have low prior knowledge of beer brewing. Second, we did not found, in our data, that interaction between treatment type and self-assessed mathematical

knowledge (which could be considered an analog to Wang et al.'s "engineering background" and computer literacy variables) influence achievement on transfer tests. Detailed results from the exploratory moderation analysis can be found in Appendix E.

### 6.1.2 Possible Explanations of the Failure to Replicate the Personalization Effect and Next Steps

(1) The most straightforward explanation is that the personalization/politeness principle becomes less robust for longer treatments; perhaps participants get tired of or become bored by it. We already pointed out this possibility in Introduction to this article. Note, however, Wang et al. (2008) actually argued that the opposite should be the case. McLarren's second study (McLarren et al., 2011b) and Doolittle's results concerning his short treatment (Doolittle, 2010) also indicate that the truth could actually be more complicated. In our opinion, this issue can be reconciled only when the length of the treatment is systematically investigated; possibly using the same content and participants with similar backgrounds. Related to that point is the question whether our personalized version enhanced the learners' mental capacity (via a higher engagement), as proposed in Figures 1b and 1c. Mental capacity is a multifaceted construct and we did not measure its aspects directly. Interestingly, indirect results (see, e.g., DeLeeuw & Mayer, 2008) are mixed. While in Experiment 1, the learners preferred the P Version, in Experiment 2 we found no between-group difference in affective variables, except for time exposure. It is thus possible that the personalized version could increase mental capacity during the first part of the interaction (e.g., during the first 30 minutes), but not afterwards. If that is the case, the personalization principle would indeed tend to be stronger in short applications.

(2) An interesting possibility, which crystallized during interviews in Experiment 2, is that the extent to which one imagines the learning content plays a role. It has been shown in the past that instructions to imagine concepts or procedures can facilitate learning compared to more conventional studying techniques (e.g., Cooper et al., 2001; Leahy & Sweller, 2004). In fact, most experiments, in which a personalization effect was demonstrated, used personalized instructions that – in our opinion – also facilitated imagination (most notably Moreno & Mayer, 2000, Exp. 1 & 2; Mayer et al., 2004; Günizi, 2010). At the same time, there was likely very limited use of imagination involved in our treatments, as well as in those of Doolittle (2010) and McLarren et al. (2011a; 2011b); in both the personalized/polite and non-personalized/direct versions. Is the personalization/politeness principle actually an "imagination principle"?

(3) It is possible that subtle differences in the level of personalization/politeness, or in the language used, caused the difference in learning outcomes (see Appendix B for details of our personalization). Indeed, Mayer noted that a "super-personalized" treatment used in an unpublished pilot study "did not improve test performance above the non-personalized treatment" (Mayer, 2001, p. 252). Günizi (2010) actually used two different personalized versions and they improved the test performance in different ways: both outperformed the non-personalized version, but the difference was significant in only one case. Finally, our personalized version featured a background story, which could, in some cases, motivate learners to participate longer in the intervention. However, it could also distract them from learning, given their limited cognitive capacity (see Tab. 1, Tab. 3, cf. Fig. 1c). In the past Mayer & Moreno (2000; Exp. 3 – 5) demonstrated the personalization effect with a treatment where *both* versions featured a background story. At the same time, a recent meta-analysis of the learning effects of serious games (Wouters et al., 2013) indicated that games *without* a

narrative are better than games with a narrative, but the difference is not significant and both game types are slightly, yet significantly, better than the "traditional" type of instruction. Thus, the "narrative issue" is still an open one. Future studies might wish to focus on the effect of different background stories on different types of learners.

(4) Another possibility is that the personalization principle may work for participants sharing some characteristics but not others. For instance, personalized instructions may be more distracting to some people (as also suggested by our Experiment 1). A high a priori knowledge is one possible culprit (McLarren et al., 2011a) and technical background/high computer literacy another (Wang & al, 2008). However, these may not be the primary factors as our Experiment 2 demonstrated. It is worth noting that the personalization/politeness effect was often demonstrated on students from psychology subjects pools (Moreno & Mayer, 2000; Mayer et al., 2004; Moreno & Mayer, 2004; Wang et al., 2008), but there are exceptions (Günizi, 2010; Mayer et al., 2004, Exp. 3). In a supplementary analysis, we did not find any influence of interaction between students' backgrounds (cf. Tab. 4) and the treatment condition on learning outcomes, even though students with technical backgrounds outperformed the other students. In the exploratory analysis only a small, marginally significant, interaction between the treatment type and Gamers score was revealed; and only for retention tests (Appendix E, Tab. E4).

(5) In a similar vein to (4), we can speculate that Czech study participants are used to the Czech schooling environment, which is more formal than the US schooling environment (one in which most previous studies' participants had probably grown up). Therefore, personalized treatment may serve as a distractor to a greater number of Czech learners than to US learners, because Czech learners, in general, are less used to a personalized approach to education than are US students (but bear in mind Günizi's study conducted in Turkey (Günizi, 2010)).

The last three cases, if confirmed, would mean that the personalization principle is very brittle. In our opinion, Explanations (1) and (2) are the most plausible. We have already started to clarify Possibility (2) by replicating the original study of Moreno & Mayer (2000, Exp. 1) with additional questionnaires assessing the extent to which participants imagine the learning content. Failure to replicate the original study would add evidence to support Explanation (5).

## *6.2 An Opposing Effect of Distraction and Motivation*

As already stated, Experiment 1 showed that participants would choose the personalized version of the simulation, when given a choice between the two versions. However, in Experiment 2, we found no between-group differences concerning affective variables, when assessed during the actual intervention (Flow, both PANAS subscales), and immediately or a month after the intervention (Hard, Like, Learnt, Motivation questions). Yet we found that P Group participants voluntarily spent about 20% more time on the simulation. We also found a positive moderate to high relation between test scores, on the one hand, and Flow scores, the positive PANAS subscale, and participants' self-reported simulation difficulty (reverse coded), likability, learning achievement and motivation to complete the intervention (Hard, Like, Learnt, Motivation questions) on the other. What is new in our study is the direct demonstration of the relationship between the affective variables and test achievements, and the usage of Flow questionnaire and PANAS.

Concerning affective variables, other works presented ambiguous findings. Some works presented null results (Moreno & Mayer, 2000, Exp. 3, 5; Mayer et al., 2004; Wang et al., 2008), other works some positive differences in favor of the personalized treatment (Moreno & Mayer, 2000, Exp. 4; Moreno & Mayer, 2004) or personalized instructions (Mayer et al.,

2006), and Günizi (2010), who used two different personalized treatments, reported mixed findings.

The inconsistent results regarding differences between the P and N treatments can be, in general, most likely attributed to two issues: a) limitations of the measurement instruments, b) different levels of distraction for different participants caused by personalization: as the CATLM framework clearly predicts, higher engagement may not always result in better learning if extraneous details are present (cf. Fig. 1 and see also Sec. 6.1.2, Point (1), (3)). Concerning Point (a), in the future it would be advantageous to use better measurement instruments (such as PANAS or Flow Short Scale) (cf. Wang et al., 2008). Concerning Point (b), our qualitative data from Experiment 1 indeed point to a possibly large difference among participants in their attitudes towards the personalization: while many welcome having the story involved, etc., several are afraid of the distraction (see Tab. 3). Could it be that the level of distraction is a function of some personal characteristics (other than self-assessed mathematical abilities and frequency of playing games)? In the future it would be useful to try to measure the amount of distraction.

Another issue is time spent on a self-paced intervention. Personalization, particularly if a story is involved, can motivate people to proceed more cautiously and therefore longer. However, the same effect can be caused by devoting more time to thinking about the story rather than deeper learning (which is, again, a distraction). These two possible causes could be connected: the detrimental effect of the greater distraction could be offset by the longer exposure (cf. Fig. 1b and 1c). That would also explain the absence of differences in learning outcomes in our Experiment 2. To our knowledge, of the self-paced studies, only Wang et al. (2008) measured treatment exposure time and they did not report data for individual conditions. In the future it would be nice to see if someone else replicates our findings

regarding differences in exposure time (duration). If differences are detected, it would be useful to explain what caused them, for instance, by measuring participants' self-reported carefulness, or, as already proposed, the amount of distraction. Otherwise it would be hard to explain the motivation–distraction tension. Note that such an approach can also be useful in any study on multimedia learning that compares two or more treatments that are supposed to promote motivation (and thereby learning outcomes) differently.

Finally, it is also possible that personalization is not (or not always) related to higher levels of motivation/interest. One of the different possible explanations of the personalization effect is imagination facilitation (Point (2) from Sec. 6.1.2). However, other possibilities exist such as improved coding due to self-referential language used in personalized instructions (cf. Moreno & Mayer, 2000; Günizi, 2010).

## *6.3 What Predicts Learning Outcomes?*

We found that the frequency of playing computer games positively predicts learning outcomes. The probable reason for this is that frequent game players are used to working with software that has a complex user interface. Thus they could master the interface of our simulation more easily than non-players did. This is actually not a very surprising outcome. Wang et al. (2008) found a similar correlation: in their study, students with above average computer skills performed, on average, better than students with average computer skills. Digital game-based learning literature regularly reports that it is difficult for non-players to use complex videogames for learning, e.g. (Brom et al., 2010). Here one should also note the similarity to Mayer's pre-training principle (Mayer, 2001).

More surprising is the finding that mathematical abilities also predict learning outcomes, even though the topic of our simulation was unrelated to mathematics. This could have happened a)

because the simulation experience involved reading graphs and histograms, b) because people with greater mathematical abilities can generally acquire new mental models quicker, or – most likely – c) due to a combination of these possibilities. A similar finding was reported in (Ackerman et al., 1995), but there the learners' task was to acquire a complex skill (i.e. to control air traffic in a simulation) rather than a mental model. To investigate this issue further, it would be enlightening to conduct a study employing an intervention with the following characteristics: a) it teaches a complex mental model, b) the learning experience does not involve operating with numbers/graphs etc. so that mathematical skills are not explicitly invoked during learning, and c) the topic is unfamiliar to the learners so that a low a priori knowledge can be expected.

## *6.4 Delayed Effects of Personalization*

The impact of the personalization can possibly be found after a longer period of time despite no immediate effects are found. To our knowledge, only McLarren's team also used (one-week) delayed post-tests and their outcome mirrored the outcome of immediate tests (McLarren et al., 2011a). Therefore, our new contribution also is reporting on the long-term effects of the intervention. We found that participants receiving personalized as well as non-personalized simulation were able to reconstruct a lot of information a month after the main experiment; however, there were no between-group differences.

Our results do not fit well with outcomes from the field of digital game-based learning where delayed effects are sometimes found despite no immediate differences (see Brom et al., 2011). However, those works tended to compare a game-based learning intervention to "conventional" classroom teaching, which is something different than comparing two similar interventions differing in one particular aspect. Anyway, we think it is useful to administer

delayed tests if possible; CATLM postulates that mental models are eventually "stored" in the long-term memory and thus it makes sense to assess their quality after a longer period of time.

# 7. Conclusion

The primary goal of this work was to investigate an important boundary condition of the personalization principle: the complexity of the phenomenon being modeled and longer time exposure. We failed to demonstrate the personalization effect under these conditions: our personalized and non-personalized treatment, two versions of the same interactive simulation, resulted, on average, in the same final learning outcome; despite the fact that *a priori* preference for the personalized version was demonstrated and despite the fact that the learners using the personalized simulation spent 20% more time on it. We also directly showed that some learners feared that personalization would distract them.

The same learning outcome could have been achieved by several ways. In our opinion, the most probable interpretation, given current data, is as follows. While finding the application similarly likable *post hoc*, learners using the personalized version probably proceeded more carefully than those working with the non-personalized version but – at the same time – could have spent more time processing information related to personalization, including the background narrative, and thus not the learning content *per se*. In other words, the positive effect of voluntarily longer exposure to the personalized treatment and the negative effect of distraction by personalization could have canceled each other out.

These results are important because they cast doubts on the robustness of the personalization principle and may have practical implications for the developers of educational simulations and serious games. To further explore boundary conditions of the principle becomes an important future work.

Because information about what type of educational simulations/games work for which learners is limited, another key finding of the present study is that learners with higher mathematical abilities and also frequent computer game players outperform the other learners, no matter the treatment type. Future research should attempt at finding what features a simulation/game should possess so that these particular results are *not* replicated. Developers of educational simulations and games probably do not want to create their products only for frequent game players and experts in mathematics.

# References

Ackerman, P. L., Kanfer, R, & Goff, M. (1995) 'Cognitive and noncognitive determinants and consequences of complex skill acquisition', *Journal of Experimental Psychology: Applied*, 1(4), 270-304.

Aiken, S., West, S. G. 'Multiple Regression: Testing and Interpreting Interactions', Sage, 1991.

Beck, I. L., McKeown, M. G., Sandora, C., Kucan, L., & Worthy, J. (1996). 'Questioning the author: A yearlong classroom implementation to engage students with text', *The Elementary School Journal*, 96, 385 - 414.

Betz, N.E., Hacket G. (1983) 'The relationship of mathematics self-efficacy expectations to the selection of science-based college majors', *Journal of Vocational Behavior*, 23, 329 - 345.

Brom, C., Sisler, V. and Slavík, R. (2010) 'Implementing Digital Game-Based Learning in Schools: Augmented Learning Environment of 'Europe 2045' ', *Multimedia Systems,* 16(1), 23-41.

Brom, C., Preuss, M., Klement, D. (2011) 'Are Educational Computer Micro-Games Engaging And Effective For Knowledge Acquisition At High-Schools? A Quasi-Experimental Study', *Computers & Education,* 57, 1971-1988.

Brown, R., Gilman, A. (1960) 'The Pronouns of Power and Solidarity', *American Anthropologist*, 4(6), 24–39.

Clark, R. C., Mayer, R. E. (2011) *E-learning and the science of instruction* (3rd ed.). Pfeiffer, A Willey Imprint.

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

Comenius, J. A (1627 - 1632) 'Didactica to jest Umění umělého vyučování' (in Czech), first known print from 1849. Latin version ['Didactica magna universale omnes omnia docendi artificium exhibens'] first printed in Comenius, J. A., ed. *Opera didactica omnia* (Part I.2), Amsterdam, in 1657. English version available as *The Great Didactic*, New York: Russel & Russel, 1967, URL: http://core.roehampton.ac.uk/digital/froarc/comgre/

Cooper, G., Tindall-Ford, S., Chandler, P., & Sweller, J. (2001) 'Learning by imagining', *Journal of Experimental Psychology: Applied*, 7, 68–82.

Cox, D. R., McCullagh, P. (1982). 'Some Aspects of Analysis of Covariance', *Biometrics*, 38(3), 541-561.

DeLeeuw, K. E., Mayer, R. E. (2008) 'A Comparison of Three Measures of Cognitive Load: Evidence for Separable Measures of Intrinsic, Extraneous, and Germane Load', *Journal of Educational Psychology*, 100(1), 223–234.

Doolittle, P. (2010) 'The Effects of Segmentation and Personalization on Superficial and Comprehensive Strategy Instruction in Multimedia Learning Environments', *Journal of Educational Multimedia and Hypermedia*, 19(2), 159 - 175.

Dowling, D.M. (1978) 'The development of a mathematics confidence scale and its applications in the study of confidence in women college students' Unpublished doctoral dissertation, Ohio State University.

Engeser, S., Rheinberg, F. (2008) Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion*, 32(3), 158 - 172.

Esslinger, H. M., ed. (2009) *Handbook of Brewing: Processes, Technology, Markets*. Wiley.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). 'Effect size estimates: current use, calculations, and interpretation'. *Journal of Experimental Psychology: General*, 141(1), 2.

Gentner, D., Stevens, A. L., eds. (1983) *Mental models.* Lawrence Erlbaum Associates.

Günizi, K. (2010) 'Does language matter in multimedia learning? Personalization principle revisited.', *Journal of Educational Psychology*, 102(3), 615-624.

Hackett, G., Betz, G. (1989) 'An exploration of the mathematics self-efficacy/mathematics performance correspondence', *Journal for Research in Mathematics Education*, 20 (3), 261-273.

Johnson-Laird, P. N. (1983) *Mental models.* Harvard University Press.

Judd, C. M., Smith, E. R. and Kidder, L. H. (1991) *Research Methods in Social Relations.* Fort Worth: Hartcourt Brace.

Leahy, W., Sweller, J. (2004) 'Cognitive Load and the Imagination Effect', *Applied Cognitive Psychology*, 18, 857 - 875.

Lester, J. C., Stone, B. A., & Stelling, G. (1999) 'Lifelike pedagogical agents for mixexi-initiative problem solving in constructivist learning environments', *User Modeling and User-Adapted Interaction*, 9, 1-44.

Mayer, R. E. (2001) *Multimedia learning,* New York: Cambridge University Press.

Mayer, R. E. (2011) 'Multimedia Learning and Games' in Tobias, S. and Fletcher J. D., eds. *Computer Games and Instruction* (pp. 281 - 306), Information Age Publishing.

Mayer, R. E., Fennell, S., Farmer, L., & Campbell, J. (2004) 'A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style'*, Journal of Educational Psychology*, 96, 389– 395.

Mayer, R. E., Johnson, W. L., Shaw, E., Sandhu, S. (2006) 'Constructing computer-based tutors that are socially sensitive: Politeness in educational software', *International Journal of Human-Computer Studies*, 64(1), 36-42.

McKenize, 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.

McLaren, B. M., Lim, S., Gagnon, F., Yaron, D., Koedinger, K. R. (2006) 'Studying the Effects of Personalized Language and Worked Examples in the Context of a Web-Based Intelligent Tutor', *Lecture Notes in Computer Sciences (Proc. 8th International Conference on Intelligent Tutoring Systems)*, 4053, 318 - 328.

McLaren, B. M., DeLeeuw, K. E., Mayer, R. E. (2011a). 'A politeness effect in learning with web-based intelligent tutors', *International Journal of Human Computer Studies*, 69,

70 - 79.

McLaren, B. M., DeLeeuw, K. E., Mayer, R. E. (2011b) 'Polite web-based intelligent tutors: Can they improve learning in classrooms?' *Computers & Education*, 56, 574 - 584.

Moreno, R. (2005) 'Instructional technology: Promise and pitfalls' In L. Pytlik Zillig, M. Bodvarsson and R. Bruning, eds. *Technology-based education: Bringing researchers and practitioners together,* Greenwich, CT: Information Age Publishing, 1 - 19.

Moreno, R., & Mayer, R. E. ( 2000). 'Engaging students in active learning: The case for personalized multimedia messages', *Journal of Educational Psychology*, 92, 727– 733.

Moreno, R., Mayer, R. (2004) 'Personalized Messages That Promote Science Learning in Virtual Environments', *Journal of Educational Psychology*, 96(1), 165 - 173.

Moreno, R., Mayer, R. (2007) Special Issue on Interactive Learning Environments: Contemporary Issues and Trends. Interactive Multimodal Learning Environments. Springer Science + Business Media.

Papert, S. A. (1993) *Mindstorms: Children, Computers, And Powerful Ideas*. Basic Books.

Peters, M. E. (2013) 'Examining the relationships among classroom climate, self-efficacy, and achievement in undergraduate mathematics: a multi-level analysis', *International Journal of Science and Mathematics Education*, 11(2), 459 - 480.

Pierfy, D. A. (1997) 'Comparative simulation game research: Stumbling blocks and stepping stones', *Simulation and Games,* 8(2), 255-268.

R Core Team (2013). 'R: A language and environment for statistical computing', R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/ [accessed: 16th June, 2013]

Randel, J. M., Morris, B. A., Wetzel, C. D. and Whitehill, B. V. (1992) 'The effectiveness of games for educational purposes: A review of recent research', *Simulation & Gaming,* 23(3), 261-276.

Rheinberg, F. (2004). 'Motivationsdiagnostik' [Motivation diagnosis]. Gottingen: Hogrefe. (in German)

Rheinberg, F., Vollmeyer, R., & Engeser, S. (2003). 'Die Erfassungdes Flow-Erlebens' [The assessment of flow experience]. In J. Stiensmeier-Pelster & F. Rheinberg, eds., *Diagnostik von Selbstkonzept, Lernmotivation und Selbstregulation* [*Diagnosis of motivation and self-concept*] (pp. 261–279). Gottingen: Hogrefe. (in German)

Tobias, S., Fletcher, J. D., Dai, D. Y., Wind, A. P. (2011) 'Review of Research on Computer Games' in Tobias, S. and Fletcher J. D., eds. *Computer Games and Instruction* (pp. 127 - 222). Information Age Publishing.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008) 'The politeness effect: pedagogical agents and learning outcomes', *International Journal of Human Computer Studies*, 66, 96–112.

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.

Welch, B. L. (1947) 'The generalization of Student's problem when several different population variances are involved', *Biometrika*, 34, 28–35.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013, February 4). 'A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games', *Journal of Educational Psychology*. Advance online publication. doi: 10.1037/a0031311

Wilensky, U. (1999) NetLogo, Center for Connected Learning and Computer-Based Modeling, Northwestern University, available: http://ccl.northwestern.edu/netlogo/ [accessed 16.6.2013].

# Appendix A: Simulation Details

This appendix details what the simulation experience consists of and how the user can interact with the simulation.

1. The first part is a tutorial demonstrating how to control the simulation while explaining the first phase of beer brewing: mashing. This part usually takes 10 - 20 minutes to complete.

2. The second part, a so-called *linear part*, demonstrates in a linear fashion how to brew beer from beginning to end, when every step is done correctly. The learner repeats the mashing phase and then proceeds to the boiling, fermenting and conditioning phases. This part takes 30 - 50 minutes to complete.

3. The third part, a so-called *error part*, demonstrates the consequences of making errors or not following the standard procedure as previously described. The learner repeats all four phases of beer brewing in a linear fashion. This part takes 35 - 60 minutes to complete.

4. In the final part, so-called *tasks*, the learner can use the simulation to brew his/her beer of a specific type. We use four different tasks described in detail in Section 5. It usually takes the learner about 10 - 20 minutes to complete one task.

The tutorial features 10 instructions (i.e. 10 process instructions and 10 tutorial instructions); each shown on an individual screen. The linear part features 24 instructions and the error part 33 instructions. The learner can move freely to any previous instruction; however, he/she can move forward only after correctly performing the step described in the tutorial instructions.

Concerning the final part, tasks, the 24 process instructions (but not the tutorial instructions) of the linear part are depicted and the learner can freely move among them.

The learner can adjust the temperature in the fermentation vessel at any moment with immediate effect (i.e. no warming up or gradual cooling). He/she can also carry out (possibly repeatedly) one type of operation in each phase: add malt in the mashing phase, add hops during the boiling phase, add yeast during the fermenting phase and add sugar in the conditioning phase. Each phase is controlled by three buttons (start the phase, carry out the phase's operation, end the phase). All 12 buttons are shown at any given moment, but only the three that correspond to the actual phase are active. (It would actually be better to show only the three buttons corresponding to the current phase; however, this is not possible in Netlogo.)

The learner can press the "assessment" button anytime. That shows him/her the assessment of the product quality; including the actual proportions of product ingredients with respect to the current phase. The learner can also speed up or slow down the simulation at any time. In case things go wrong, the learner can restart the current phase (i.e. the learner does not need to restart the whole process from the beginning).

The following ingredients are typically visibly present in the product: 1) enzymes, bacteria, sugar, starch and enzymes from malt during the mashing phase; 2) sugar, residual starch and hops in the boiling phase; 3) sugar, residual starch, yeast, alcohol, $CO_2$ and fusel alcohol in the fermenting and conditioning phases. When things go wrong, bacteria and acetone can appear during any phase. The user can monitor the amount of the ingredients through the graphs, histograms and numerical panels. When the simulation is running, the graphs, histograms and numerical panels are constantly updated, and the content of the fermentation vessel is animated.

# Appendix B: P Version of the Simulation

This appendix details changes made to the N Version to make it into the P Version.

1. The imperatives were changed from V-forms to T-forms, such as: "*Click$_V$ the button*" ("*Stiskněte tlačítko*" in Czech) (N Version) was changed to "*Click$_T$ the button*" ("*Stiskni tlačítko*" in Czech) (P Version). There were hundreds of changes of this type in the instructions.

2. The information presented in the active plural form in the N Version such as "*For each degree of beer we add to the tank roughly enough malt to equal 1.5% of the volume of the water.*" was changed to the imperative in the T-form: "*For each degree of beer you$_T$ add to the tank roughly...*". Some information presented in the active plural form in the N Version was changed to the first person singular as if the grandpa were expressing his knowledge/opinion. For instance: "*From this moment on we [I] will call what's in the tank the PRODUCT.*" ("*we*" in the N Version was converted to "*I*" in the P Version). There were dozens of changes of this type in the instructions.

3. Pronouns referring to the brewery ownership were added, such as: "*Because the [this] brewing tank holds 1000 liters of water...*" ("*this*" was used in the P Version) or "*Beer can be brewed in different ways. The simplest is [In our brewery we brew it using] the INFUSION METHOD.*" ("*In our brewery we brew it using*" was used in the P Version). There were 16 changes of this kind in the whole set of instructions.

4. When a learner did something correctly, a laudatory comment was added from time to time; for instance: "*Excellent! Because this brewing tank holds 1000 liters of water...*" ("*Excellent!*" was added in the P Version). There were 7 modifications of this kind.

5. Although rarely, sometimes a polite formulation was used in the P Version instead of a direct imperative, such as "*Now try$_T$ to click on the „>>“ button.*" (P Version) instead of "*Click$_V$ on the „>>“ button.*" (N Version). Investigation into the usage of polite vs. direct language used for instructions is a research endeavor that runs parallel to the personalized vs. non-personalized messages issue (e.g., Mayer et al., 2006; Wang et al., 2008; McLaren et al., 2011a; McLaren et al., 2011b). Thus, for future analysis, it is important to know to what extent our P Version was also polite. In fact, polite formulations, except in Point 2 above, were rare in our P Version. There are usually up to 10 imperatives in every tutorial instruction of our simulation. However, in terms of (Mayer et al., 2006), only three were polite in the P Version of the whole tutorial, five in the whole linear part and eight in the whole error part. One formulation is also polite in both versions of the tutorial, and two in both versions of the error part.

6. Filler conversational formulations were added in the P Version to stress that it is the grandpa who is talking "through" the instructions. An example is: "*[But watch$_T$ out!] The malt can also contain BACTERIA that we [you$_T$] will have to get rid of later.*" (the first sentence was added in the P Version and, according to Point 2, "*we*" was changed to "*you$_T$*".) or "*Click$_V$ „>>“*" was changed to "*Click$_T$ „>>“ and you$_T$ will find out what happens next.*". Thirty-eight of these filler formulations were added in the P Version.

# Appendix C: Questionnaires

*Pre-questionnaire.*

The Pre-questionnaire solicited information about participants' gender, age and field of study. We also asked the participants the following questions concerning their gaming experience:

- "How often do you play computer games?" with the scale "*1) less than one hour a week; 2) 1 - 5 hours a week; 3) 6 - 10 hours a week; 4) more than 10 hours*";
- "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.)?" with the scale: "*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*.";
- "How often have you played board games or card games during the past 10 years (e.g. Carcassonne, Contract Bridge...)?" with the scale: "*1) never or less than once a year; 2) approx. once or twice a year; 3) more than twice a year, but less than every month; 4) at least once a month*".

An additional question solicited answers on the frequency of computer use using the same scale as the question on frequency of playing computer games above.

In order to measure participants' *Self-assessed knowledge of mathematics*, we included the following question with a 6-point Likert scale: "Check one of the following to indicate your knowledge of mathematics." (*1 - very good; 6 - very weak*). To further investigate participants' *Self-perceived ability to acquire mental models of mechanisms and processes*, we included the following question with a 7-point Likert scale: "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 nearly hundreds of steamboats and five functionally-distinct types of propellers. What would you prefer to learn?" (*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*). Note that complex mathematics self-efficacy and mathematics performance tests exist, such as the Mathematics Self-Efficacy Scale (Betz & Hacket, 1983) or the Mathematics Confidence Scale (Dowling, 1978). However, we could not use them due to time constraints. Because the self-rating of mathematical abilities was demonstrated in the past to predict actual mathematical performance, e.g. (Hacket & Betz, 1989; Ackerman et al., 1995; Peters, 2013), we decided to include the two questions above and supplement them with a simple test of graphing skills, TOGS (McKenzie & Padilla, 1986), to investigate our Hypothesis 3.

We included the following questions to measure indirectly participants' knowledge of beer brewing and making alcohol:

- (Q1) "Check the items that are true in your case: 'My relatives (or I personally) brew beer,' 'I have taken part in an excursion to a brewery,' 'We learnt about beer brewing in school,' 'I know what *Saccharomyces cerevisiae* is,' 'I know how *Lactobacillus* can influence beer,' 'I know why malt is added to beer before yeast.'"
- (Q2) "Please write down whether you have ever tried to learn about the topic of beer brewing. If so, when and where?" This question was an open-ended one.
- (Q3) "Should you be asked to explain why and when alcohol is created during the beer brewing process, would you consider yourself to be:" using the scale "*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?" with a 6-point Likert scale (*1 - definitely yes; 6 - definitely no*).
- (Q5) "How often do you discuss the topic of beer brewing with your friends or family?" with a 6-point Likert scale (*1 - always; 6 - never*).
- (Q6 - 8) "Check to indicate your knowledge of beer brewing [Q6] / wine-making [Q7] / whiskey production [Q8]." with a 6-point Likert scale (*1 - very good; 6 - very weak*). These were three separate questions.

*Motivation questionnaire 1*

The wording of questions in Motivation questionnaire 1 was as follows:

- Two questions intended to assess learners' self-perceived knowledge of beer brewing: "Check to indicate your knowledge of beer brewing." (*Knowledge 1* question). Note: the same question was also present in the Pre-questionnaire (Q6). The second question was: "Check to indicate how much you have learnt today about beer brewing." (*Learnt 1* question).
- The following question was intended to assess learners' interest: "Check to indicate how much you liked today's lesson on the topic of beer brewing." (*Like 1* question).
- The following two questions were included to assess learners' perceptions of difficulty learning from the materials: "Check to indicate how difficult the simulation was for you" and "Check to indicate how much effort it took for you to learn about beer brewing using this simulation." (*Hard* questions).

*Motivation questionnaire 2*

The wording of questions in Motivation questionnaire 2 was as follows:

- Two questions were intended to assess learners' self-perceived knowledge of beer brewing: "Check to indicate your knowledge of beer brewing." (*Knowledge 2* question), and: "Check to indicate how much you learnt about beer brewing a month ago." (*Learnt 2* question).
- The following question was intended to assess learners' interest: "Check to indicate how much you liked last month's lesson on the topic of beer brewing." (*Like 2* question).
- The following question was intended to assess learners' self-perception of their learning motivation during the original session: "Check to indicate how hard you worked to learn something at the workshop a month ago." (*Motivation* question).
- Two questions were included to solicit information about the frequency of drinking beer and alcohol in general: "Do you drink beer?" and "Do you drink alcoholic beverages other than beer?" using the scale: "*1) never or less than once a year; 2) more than once a year, but less than every month; 3) more than once a month, but less than once a week; 4) more than once a week.*"
- To check whether participants talked/sought out information about beer brewing, the following two questions were included. "How often have you spoken to someone about the topic of beer brewing during the past month?" and "After the workshop ended (one month ago), did you try to look up additional information on beer brewing based on your own interest? " using the scale: "*1) never; 2) once or twice; 3) three or four times; 4) more than four times.*" The latter question was supplemented with the open-ended question: "If so, what information did you look for? ....."

# Appendix D: Beer Brewing A priori Knowledge Score

The *beer brewing a priori knowledge score* (BB-a-priori score) was calculated as follows: For each item checked in the list for Question Q1, two points were assigned (12 points is the maximum). Two additional points could be assigned for the answer to Q2, if it did not repeat an item from Q1. Zero to three points could be assigned for the answer on Q3 corresponding to the question's scale 1 .. 4 (Zero points for 1, three points for 4). Zero to 2.5 points could be assigned for the answers to Q4, Q5, Q7, and Q8 corresponding to their respective scales 1 .. 6 (reverse coded; zero points for 6, 2.5 points for 1). Zero to five points could be assigned for the answer to Q6 corresponding to its scale 1 .. 6 (reverse coded; zero points for 6, 5 points for 1). Taken together, the maximum BB-a-priori score could be 32. In our opinion, an expert could achieve around 25 points, while a moderately educated home-brewer could earn at least 15 points. No participant achieved that score.

# Appendix E: Additional Analysis

For exploratory purposes, we conducted a supplementary moderation analysis according to Baron and Kenney (1986). In particular, we tested whether affective variables, self-assessed mathematical knowledge and frequency of playing games moderate learning outcome as captured by *Transfer test* and *Retention test* variables.

## Data Analysis

As suggested by Baron and Kenney (1986), we used analysis of covariance (ANCOVA) to test for a possible moderation effect (between-subject factor: treatment type (P/N); covariate: an affective variable/Math score/Frequency of computer games playing). Effect size was captured by *partial* $\eta^2$ (Fritz et al., 2012) with the following classification: small (.01), medium (.06) and large (.14) (Cohen, 1988).

## Description of Results

Results of all ANCOVA analyses are reported in tabular form. Because they all share the same degrees of freedom (residual $df$ =66), the tables contain only $F$ and $p$ values for each factor. Each table has three main columns: *P/N effect*, *Covariate effect* and *Interaction effect*. The column *P/N effect* describes the main effect of the treatment type. Because we have already shown in Section 5.2.2 that the treatment type did not influence learning outcomes, we can expect that ANCOVA reveals no main effect of the treatment type as well. The column *Covariate effect* describes the effect of each covariate. This corresponds to the median-split technique used in Sections 5.2.3 and 5.2.4. The column *Interaction effect* describes interaction between the treatment type and a covariate; in other words, if the covariate moderates learning. Values in **bold** denote a significant effect, values in *italics* denote trends.

## Results: Does an Affective Variable Moderate the Learning Outcome?

The results are summarized in Tables E1 (Transfer test) and E2 (Retention test). As expected, no main effects of the treatment type were revealed.

Concerning the Transfer test variable, all affective variables influence learning outcomes significantly, except for PANAS– (non-significant) and Motivation (marginally significant). This corresponds to the results from Table 14. None of the interaction terms are significant, which means that none of the variables have a moderating effect on learning.

Concerning the Retention test variable, all affective variables influence significantly learning outcomes, except for PANAS– (non-significant). This corresponds to the results from Table 13. Again, none of the interaction terms are significant, which means that none of the variables have a moderating effect on learning.

--- Insert Table E1 about here ---

Table E1: Results of ANCOVA testing if affective variables moderate learning outcome measured by the Transfer test variable.

| | P/N effect | | | Covariate effect | | | Interaction effect | | |
|---|---|---|---|---|---|---|---|---|---|
| Transfer test | F | p | $\eta_p^2$ | F | p | $\eta_p^2$ | F | p | $\eta_p^2$ |
| **Like** | 1.01 | 0.318 | 0.02 | **7.40** | **0.008** | **0.10** | 0.01 | 0.942 | 0.00 |
| **Learnt** | 1.10 | 0.297 | 0.02 | **12.45** | **0.001** | **0.16** | 1.36 | 0.247 | 0.02 |
| *Motivation* | 0.96 | 0.330 | 0.01 | *3.51* | *0.066* | *0.05* | 0.18 | 0.671 | 0.00 |
| **Hard** | 1.15 | 0.288 | 0.02 | **16.73** | **<0.001** | **0.20** | 0.18 | 0.669 | 0.00 |
| **Flow** | 1.14 | 0.290 | 0.02 | **15.6** | **<0.001** | **0.19** | 0.76 | 0.386 | 0.01 |
| **PANAS+** | 1.11 | 0.296 | 0.02 | **4.36** | **0.041** | **0.06** | 0.37 | 0.546 | 0.01 |
| PANAS– | 1.07 | 0.305 | 0.02 | 0.74 | 0.392 | 0.01 | 1.39 | 0.243 | 0.02 |

--- Insert Table E2 about here ---

Table E2: Results of ANCOVA testing if affective variables moderate learning outcome measured by the Retention test variable.

| | P/N effect | | | Covariate effect | | | Interaction effect | | |
|---|---|---|---|---|---|---|---|---|---|
| Retention test | F | p | $\eta_p^2$ | F | p | $\eta_p^2$ | F | p | $\eta_p^2$ |
| **Like** | 0.17 | 0.684 | 0.00 | **10.62** | **0.002** | **0.14** | 0.00 | 0.958 | 0.00 |
| **Learnt** | 0.18 | 0.672 | 0.00 | **16.97** | **<0.001** | **0.2** | 0.00 | 0.956 | 0.00 |
| **Motivation** | 0.16 | 0.694 | 0.00 | **5.22** | **0.026** | **0.07** | 0.04 | 0.845 | 0.00 |
| **Hard** | 0.16 | 0.694 | 0.00 | **5.25** | **0.025** | **0.07** | 0.12 | 0.733 | 0.00 |
| **Flow** | 0.17 | 0.679 | 0.00 | **13.19** | **0.001** | **0.17** | 0.09 | 0.764 | 0.00 |
| **PANAS+** | 0.12 | 0.725 | 0.00 | **5.47** | **0.023** | **0.08** | 0.36 | 0.551 | 0.01 |
| PANAS– | 0.12 | 0.729 | 0.00 | 0.01 | 0.912 | 0.00 | *3.69* | *0.059* | *0.05* |

## Results: Does Math Score or Frequency of Games Playing Moderate the Learning Outcome?

The results are summarized in Tables E3 (Transfer test) and E4 (Retention test). As expected, no main effects of the treatment type were revealed.

Concerning covariates, Tables E3 and E4 reveal similar findings to those discovered using the median split technique (Tables 9 and 10). There were, however, two exceptions. First, concerning Transfer tests, ANCOVA showed the main effect of Math score while the median split technique showed only a trend. This is not very surprising: recall that the median split technique also revealed significant differences when the High and Low performing groups were formed using Immediate tests only as well as using Delayed tests only (and not an average of both, as described in Sec. 5.1.7). Second, concerning Retention tests, the main effect of Math score was found to be significant in the ANCOVA analysis ($p$=0.026), but in

the median split technique, the difference between the High and Low group was not significant ($p$=0.111). On the contrary, ANCOVA revealed no main effect of Frequency of playing computer games ($p$=0.133), but this variable was significant when using the median split technique ($p$=0.031). This also influenced significance for the Gamers score variable. While this difference may be of some theoretical interest, note that, first, differences in effect sizes are not very large (they change from medium to small or vice versa) and, second, the most important and the largest difference, which concerns the TOGS score, was revealed by both analyses. Thus the discrepancies between the two analyses do not undermine the subsequent discussion.

Concerning interaction terms, no interaction was revealed by ANCOVA. Only interaction between the Retention test variable and Gamers score was found to be marginally significant ($p$=0.079). Perhaps this might be an analogy to the expertise-reversal effect, but one should avoid drawing strong conclusions due to a) the exploratory nature of this analysis and b) the fact that no similar effect was found regarding transfer tests.

--- Insert Table E3 about here ---

Table E3: Results of ANCOVA testing if several variables related to participants' background moderate learning outcome measured by Transfer test variable.

| | P/N effect | | | Covariate effect | | | Interaction effect | | |
|---|---|---|---|---|---|---|---|---|---|
| Transfer test | F | p | $\eta_p^2$ | F | p | $\eta_p^2$ | F | p | $\eta_p^2$ |
| **Math score** | 0.99 | 0.324 | 0.01 | **10.63** | **0.002** | **0.14** | 0.56 | 0.455 | 0.01 |
| **Freq. comp. games pl.** | 1.08 | 0.303 | 0.02 | **11.81** | **0.001** | **0.15** | 0.01 | 0.904 | 0.00 |
| Freq. LARP playing | 0.94 | 0.336 | 0.01 | 1.79 | 0.185 | 0.03 | 0.28 | 0.596 | 0.00 |
| Freq. board games pl. | 0.96 | 0.330 | 0.01 | 1.42 | 0.237 | 0.02 | 2.30 | 0.134 | 0.03 |
| **TOGS score** | 1.22 | 0.273 | 0.02 | **22.1** | **<0.001** | **0.25** | 0.14 | 0.713 | 0.00 |
| **Mental models score** | 0.99 | 0.322 | 0.01 | **5.86** | **0.018** | **0.08** | 0.08 | 0.776 | 0.00 |
| **Gamers score** | 1.13 | 0.293 | 0.02 | **20.88** | **<0.001** | **0.24** | 0.97 | 0.329 | 0.01 |

--- Insert Table E4 about here ---

Table E4: Results of ANCOVA testing if several variables related to participants' background moderate learning outcome measured by Retention test variable.

| | P/N effect | | | Covariate effect | | | Interaction effect | | |
|---|---|---|---|---|---|---|---|---|---|
| Retention test | F | p | $\eta_p^2$ | F | p | $\eta_p^2$ | F | p | $\eta_p^2$ |
| **Math score** | 0.3 | 0.585 | 0.00 | **5.20** | **0.026** | **0.07** | 2.30 | 0.134 | 0.03 |
| Freq. comp. games pl. | 0.15 | 0.700 | 0.00 | 2.31 | 0.133 | 0.03 | 0.23 | 0.634 | 0.00 |
| Freq. LARP playing | 0.15 | 0.704 | 0.00 | 0.12 | 0.731 | 0.00 | 0.41 | 0.524 | 0.01 |
| Freq. board games pl. | 0.15 | 0.704 | 0.00 | 0.04 | 0.841 | 0.00 | 0.67 | 0.417 | 0.01 |
| **TOGS score** | 0.19 | 0.664 | 0.00 | **20.14** | **<0.001** | **0.23** | 0.82 | 0.368 | 0.01 |
| Mental models score | 0.15 | 0.699 | 0.00 | 1.50 | 0.225 | 0.02 | 1.72 | 0.194 | 0.03 |
| **Gamers score** | 0.3 | 0.584 | 0.00 | **4.73** | **0.033** | **0.07** | *3.19* | *0.079* | *0.05* |

In summary, the exploratory analysis did not reveal anything particularly interesting beyond the analyses presented in Section 5.2.