Citation

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

Title page

1) Full title

Flow, Social–Interaction Anxiety and Salivary Cortisol Responses in Serious Games: a Quasi-Experimental Study

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

Michaela Buchtová
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00, Prague, the Czech Republic
Faculty of Arts, Charles University in Prague
U kříže 8, 15800 Prague 5, Czech Republic
michaela.buchtova@ff.cuni.cz

Vít Šisler
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00, Prague, the Czech Republic
vsisler@gmail.com

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

Rupert Palme
Institute of Medical Biochemistry, Department of Biomedical Sciences, University of Veterinary Medicine Vienna, Austria
Veterinärplatz 1, A-1210 Vienna, Austria
rupert.palme@vetmeduni.ac.at

Lisa Maria Glenk
Comparative Medicine, Messerli Research Institute, University of Veterinary Medicine Vienna, Medical University Vienna, University of Vienna, Austria
Veterinärplatz 1, A-1210 Vienna, Austria
lisa.glenk@vetmeduni.ac.at

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 and Vít Šisler were 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. Lisa Maria Glenk was partially supported by a doctoral grant given by the University of Veterinary Medicine Vienna.

5) Acknowledgement

We thank all research assistants and lecturers who helped to conduct the experiment, namely: Ivo Šebek, Jakub Vlasák, Lukáš Kolek, Viktor Dobrovolný, Tereza Selmbacherová, Marie Grecká, Martina Denemarková, Jana Tupá, Ondřej Smíšek, Lucie Bauerová, Radka Římanová, Jiří Lacík, Miroslava Abrahámová, Jakub Fiala, Monika Hampacherová, Kamil Pavelka, Edita Bromová, Tereza Stárková, Edith Klobetz-Rassam. We thank all high-school teachers enabling us to conduct the experiment, and also ZDrSEM organizers (http://www.zdrsem.cz). We also thank the staff of ICV, s.r.o., (http://www.icv.cz) who provided places for part of the experiment.
The human data were collected with APA ethical principles in mind.

Flow, Social Interaction Anxiety and Salivary Cortisol Responses in Serious Games: a Quasi-Experimental Study

Abstract

Serious games are supposed to instigate engagement and, in turn, improve learning. High engagement is frequently connected with a positive affective state and a high flow state. However, the alleged link between a learner's affective state, his/her flow state and learning outcomes has not been investigated in detail in the context of serious games. Even less information is available on how serious games may influence markers of physiological arousal. To fill this gap, participants of this exploratory study (*N* = 171) played one of six different serious game-based treatments, while we measured their affect, flow, cortisol secretion and learning achievement. The treatments were supposed to generate different levels of engagement and cortisol responses, because some of them were designed for a single user, while others were team-based, featuring so-called social-evaluative threat (ST) components. Our results revealed that flow was positively related to positive affect and negatively to negative affect. While flow and positive affect were related to learning gains, almost no relationship between either of these three variables and cortisol levels was found. Negative affect and cortisol were elevated in social-interaction anxious males in team-based conditions. This study contributes to the limited body of research on the relationship between engagement and learning in serious games. We provide new perspectives on the relationships between flow, positive/negative affect and cortisol. Our findings highlight the fact that team-based serious games with ST components may have adverse effects on learners, particularly males, with high social-interaction anxiety.

Highlights

• We studied the link between affect, flow, cortisol secretion and learning achievement.

• Learners participated in one of the six different serious game-based treatments.

• Flow was positively related to positive affect and negatively to negative affect.

• Flow/positive affect was related to achievement, but not to cortisol levels.

• Cortisol was elevated in social interaction anxious males in team-based conditions.

Key words

Digital game-based learning; collaborative games; flow; PANAS; social interaction anxiety; cortisol; learning

1. Introduction

Digital game-based learning (DGBL) presents a new instructional technology with many alleged advantages in the context of a formal schooling system. Digital games for education, oftentimes called *serious games*, have been gradually coming into use by schools (Wastiau et al., 2009; Huizenga et al., 2013). The number of research studies investigating serious games' usage, learning effects and the attitudes of different stakeholders towards games' adoption in formal education is growing (e.g., Hays, 2005; Sitzmann, 2011; Tobias & Fletcher, 2011; Connolly et al., 2012; De Grove et al., 2012; Girard et al., 2013; Wouters et al., 2013).

One of the key alleged advantages of the DGBL approach is that games could motivate learners via play and this, in turn, could improve learning (*Motivation → learning* hypothesis). This idea has been articulated by many researchers (e.g., Garris et al., 2002; Hays, 2005; Wouters et al., 2013; see also Malone, 1981; Malone & Lepper, 1987; Habgood and Ainsworth, 2011). However, despite a large body of research on disentangling the link between emotions and cognition (e.g., Robinson et al., 2013; Eysenck and Keane, 2010, Ch. 15) and emotions and memory/learning (e.g., Anderson, 2009; Reisberg and Hertel, 2003; but see also Pekrun, 2005), the issue of mere establishing a clear link between games' motivational factors and students' learning gains has not been sufficiently addressed in the DGBL context. First, as suggested in an older review of educational game studies (Hays, 2005; p. 47), games may be inherently more engaging than conventional instruction methods but that may not necessarily result in better learning outcomes. A game's motivational factors, deemed to promote learning by increasing the learner's interest and making him/her invest more energy into learning, may also serve as distractors and thereby reduce learning gains; i.e., a trade-off (cf. Mayer, 2009; Moreno, 2005; van Dijk, 2010; Um et al., 2012). Second, DGBL studies only rarely report correlations between affective and knowledge measures. The most recent meta-analysis, and probably also the most rigorous so-far (Wouters et al., 2013), indicated that games are slightly better for learning, when compared to traditional types of instruction, as well as slightly more motivating, but the latter finding was only marginally significant[1]. In addition, the relation between the affective and cognitive dimensions was not elucidated. Only a handful of studies have directly investigated this relationship in the DGBL field (e.g., van Dijk, 2010; Ritterfeld et al., 2010; see also Habgood and Ainsworth, 2011) or in the field of multimedia learning (e.g., Um et al., 2012; der Meij, 2013; Plass et al., 2014). Finally, classical measures - mostly questionnaires with Likert-items, often self-constructed and administered after the intervention - were sometimes questioned due to low validity (e.g., Wang et al., 2008, p. 110; Wouters et al., 2013, p. 261).

Recent research has attempted to identify transient affective states experienced by learners during a learning task (e.g., Craig et al., 2004; Elliot and Pekrun, 2007; Hussain et al., 2011). These states often include anxiety, boredom, confusion, frustration, curiosity, delight and engaged concentration (Baker et al., 2010; Lester et al., 2013; D'Mello & Graesser, 2012). Engaged concentration, also called state engagement, has so far not been operationalized precisely, but it is tentatively linked to mild generalized positive affect and certain components of flow state; such as focused and intense attention[2]. Affect has a complex structure, but generalized positive and negative affect emerge as "two dominant and relatively independent dimensions" (Watson et al., 1988, p. 1063). Flow state is often conceptualized as: a) highly focused concentration on the activity; b) coherence of the activity; c) balance between one's skills and the activity's demands; d) deep sense of control; e) distorted temporal experience; and f) a feeling that the activity is innately rewarding (Csikszentmihalyi, 1975; cf. Keller & al., 2011; Engeser & Rheinberg, 2008). Even though the concept of engaged concentration originated in the field of tutoring systems (see Baker et al., 2010), it is also highly

[1] The second recent meta-analysis (Sitzmann, 2011) reported similar findings, as concerns the cognitive dimension, but noted that "the scarcity of [comparative] research ... precludes an empirical test of the effect of simulation games on post-training motivation, effort, and trainee reactions." (p. 495). In studies with random sampling, the positive effect of games on learning gains significantly diminishes in (Wouters et al., 2013) but not in (Sitzmann, 2011). These two meta-analyses have minimal overlap in primary literature.

[2] The relationship between positive affect and flow state, on the one hand, and engaged concentration, on the other hand, was pointed out to us by Sidney D'Mello [email correspondence from 9 March 2014].

relevant for the DGBL field, because it is arguably one of the most crucial affective states instigated by playing games. In this study, we will assess it indirectly by measuring generalized positive affect and flow. Notably, positive affect and flow are correlated when participants are engaged in interesting tasks (Rogatko, 2009; Brom et al., 2014). Both flow and positive (as well as negative) affect can be assessed by standardized research instruments, such as the Flow Short Scale (Rheinberg et al., 2003) and the Positive and Negative Affect Schedule (PANAS; Watson & al., 1988), respectively. Yet only few DGBL studies have investigated learning effects, flow and positive–negative affective states all at the same time.

Digital games frequently involve competitive or challenging tasks that strongly influence players' engaged concentration. It is often assumed that this influence is generally positive; however, from a psycho-physiological perspective, these tasks may be inherently stressful for some players. Both physical and psychological stress can activate the hypothalamus-pituitary-adrenal (HPA) axis, resulting in triggered secretion of the glucocorticoid hormone cortisol (Wingfield & Sapolsky, 2003). Mediating cascading levels of physiological arousal, the primary function of cortisol is to help an organism adapt to its environment. Increases in cortisol have been linked to stressful experiences that require an individual to cope with internal or external demands (Chrousos, 2009). Because challenging tasks in games are supposed to increase players' engaged concentration (i.e., positive affect and/or flow state) and certain challenging tasks are also connected to elevated cortisol (Dickerson & Kemeny, 2004), we can conjecture that engaged concentration may be connected to elevated cortisol too. Notably, it has also been suggested that cortisol levels vary with positive and negative outcomes on learning and memory (Roozendaal, 2002). Could cortisol play a role in linking engaged concentration and learning?

Over the past decades, analysis of salivary cortisol in response to a stressor has established itself as a state-of-the art method in psycho-physiological research (Hellhammer et al., 2009). In humans, cortisol secretion follows a typical circadian pattern, with increasing levels in the early morning hours and a peak at the time of waking. Afternoon is perhaps the best time for conducting laboratory research that includes cortisol sampling (Dickerson & Kemeny 2004). Nevertheless, if confronted with a powerful stimulus (i.e., stressor), cortisol levels sampled during any part of the day can even rise above those of the circadian peak (Kudielka & al., 2009). Saliva sampling is non-invasive and can be carried out easily under natural conditions outside of a laboratory (Inder et al., 2012), including during game playing.

In general, past research on (non-educational) digital games has previously incorporated cortisol measurements. For instance, the cortisol-modulating effects of built-on music during video game playing have been described by Hébert et al. (2005). Violent content in video games has been controversially linked to subsequent increases in salivary cortisol (Hossini et al., 2011; Ivarsson et al., 2009, Oxford et al., 2010). Stressful video games can decrease reaction time in the accomplishment of attentional tasks in absence of a concomitant increase in salivary cortisol levels (Skosnik et al., 2000). However, to the best of our knowledge, salivary cortisol has not yet been measured in the DGBL context. As argued above, cortisol levels could be particularly interesting to complement information about the participants' subjectively perceived affective state and flow when investigating the influence of the affective state/flow on learning gains. Thus, we designed a study with the following goals:

(1) To explore the link between affective state/flow of a learner, i.e., constructs related to engaged concentration, and his/her immediate learning gains in a DGBL intervention. The affective state is measured primarily by PANAS (Watson & al., 1988) and flow by Flow Short Scale (Rheinberg et al.,

2003), which are brief and therefore relatively non-invasive when administered in situ in the DGBL context. This part of the study is exploratory; we put forward no specific hypothesis.

(2) To elucidate several hypotheses on the link between flow state and cortisol levels. These hypotheses are introduced in Section 1.1.

(3) To explore the link between a learner's positive–negative affective state, measured by PANAS, and cortisol levels. This part of the study is exploratory. This goal is detailed in Section 1.2.

(4) To explore the link between immediate learning gains and cortisol changes. This part of the study is also exploratory. This goal is detailed in Section 1.3.

(5) The methodological goal of this study is to evaluate salivary cortisol assessment in the field of DGBL and to discuss its methodological possibilities and limitations based on our experience gained during this study. This goal is described in detail in Section 1.4.

To accomplish these goals, we use intentionally five *different* educational interventions in this study. The first three are the multi-player computer game, *Europe 2045*, and two derivatives of this game. *Europe 2045* and its derivatives feature both collaborative and competitive aspects. The fourth is an interactive computer simulation in which learners learn how to brew beer. Purposefully, this simulation is single-"player" rather than multi-"player." The fifth intervention is a short, experiential, non-computer simulation, embedded in a several-days-long, first aid training course. This simulation is team-based and collaborative. All of these interventions are expected to elevate the level of learners' arousal, but differ substantially regarding their conceptual design (team vs. single, different type of participant interaction, different situation). To enable generalization across different delivery media, we purposefully use computer-based and non-computer-based interventions (cf. Clark, 2012; Ross and Morrison, 1989).

1.1 Goal 2: Cortisol Levels and Flow

We are aware of only four studies investigating relationship between flow and salivary cortisol responses (Keller et al., 2011, Exp. 2; Peifer, 2012; Peifer et al., 2014). Based on her results, Peifer (2012) put forward an *Inverted-U* hypothesis, positing that the relationship between flow and physiological arousal is reflected by an inverted-U relationship; that is, the flow is high for a medium physiological arousal (measured, for instance, by salivary cortisol levels) and low for both low as well as high arousal. A different, but also plausible, hypothesis was supported by the study of Keller et al. (2011; Exp. 2), and also by one of Peifer's studies (2012); the *Perceived-fit* hypothesis. According to the classical conceptualization of flow (Csikszentmihalyi, 1975), the flow tends to be high when a participant's skills required to accomplish a given task match the task's demands. This condition is often called skills-demand-compatibility (but see also Engeser and Rheinberg, 2008; Løvoll and Vittersø, 2014). The Perceived-fit hypothesis posits that cortisol levels would be highest when participants are in the skills-demand-compatibility condition; in other words, when they perceive a fit between their skills and the task's demands and thus are in flow. The cortisol levels would be low both when skills are higher than demands (a so-called boredom condition) as well as when the demands exceed skills (a so-called anxiety/stress condition). In both of these conditions, the flow is also expected to be low. Thus, both the Keller et al.'s study (Exp. 2) and one Peifer's study indicated that high flow state is connected with elevated cortisol. As a consequence, Keller et al. (p. 852) also questioned, from the psycho-physiological perspective, beneficial effects of flow state when it is experienced over a prolonged period.

It is, however, also possible that the general relationship between experiencing flow and cortisol levels is rather minimal and the flow-cortisol relationship is heavily moderated by the type of intervention, personal characteristics, and interaction of both. We call this idea *Treatment-specificity-and-personal-characteristics* hypothesis (TSPC). For instance, specific tasks featuring social-evaluative threat (ST) components, for example the "Trier Social Stress Test", that requires a person to give an impromptu speech to an evaluation panel, can significantly increase cortisol levels from pre- to after exposure (Kirschbaum et al., 1993; Schommer et al., 2004). A large meta-analysis of cortisol-based laboratory studies reported that tasks containing both ST and uncontrollability elements were associated with the largest cortisol increases (Dickerson & Kemeny, 2004). Multi-player game-based activities may feature these two characteristics when a learner has to speak to an audience comprised of his/her peer learners and/or has to participate in discussions with them while the speech/discussion have an impact on the game's progress and/or can be judged by the player's peers. We can thus expect that such DGBL activities will be associated with greater cortisol increases than single-player serious games/simulations whereas the intensity of flow will depend on different aspects of the interventions. In other words, we may be able to dissociate high flow state and elevated cortisol. The TSPC hypothesis does not predict that the relationship between flow and cortisol levels can never be observed, but it postulates that the existence of this relationship is largely dependent on the treatment type and participants' characteristics.

Our Goal 2 is to elucidate the tension between the three hypotheses above by dissociating high flow and elevated cortisol.

In addition, gender is a potentially confounding factor for studying stress hormones in psychosocial research (Kudielka et al., 2004). Stress experience by means of cortisol release is related to gender differences; generally reporting amplified responses to acute stress in men compared to women (Kajantie & Phillips, 2006; Kudielka & Kirschbaum, 2005). However, it has been suggested that the observed effects of gender may be stressor-specific. In a study by Stroud et al. (2002), men exhibited higher cortisol levels after challenging cognitive tasks, while women responded significantly stronger to a challenge involving social rejection. Some studies failed to identify gender-related differences in stress responsiveness (Wang et al., 2007). Hence, it would be particularly interesting to find out whether a team-based game with ST components could be linked to differences in the hormonal responsiveness of men and women.

1.2 Goal 3: Cortisol Levels and PANAS

Regarding PANAS, increased levels of cortisol have been reported to alter the perception of a stimulus from objectively neutral to more arousing; however, it seems that these changes are not related to the self-reported negative affect state (Abercrombie et al., 2005). In at-risk adolescent males undergoing a psychological challenge task, cortisol showed no correlation with a positive affect and only partial correlations with a negative affect (McBurnett et al., 2005). Het et al. (2012) showed inverse relationship between cortisol and negative affect for participants undergoing the Trier Social Stress Test. Thus, our next goal is exploratory; we aim to investigate the relationship between both dimensions of PANAS and cortisol increases without putting forward any particular hypothesis regarding this relationship.

1.3 Goal 4: Cortisol Levels, Learning and Testing Conditions

In the context of learning, the release of glucocorticoids has been controversially linked to memory formation (Roozendaal, 2002). During a moderate increase, cortisol can improve memory consolidation and retrieval with a stronger outcome for emotionally aroused individuals (Abercrombie et al., 2006). Furthermore, in a written exam situation, individual achievements (i.e., the number of correct answers) were linked to cortisol increases (Flegr & Priplatova, 2010). Interestingly, women using oral contraceptives seem to be less susceptible to cortisol-modulating effects on memory retrieval (Kuhlmann & Wolf, 2005).

There is also evidence that the very testing conditions in an experimental setting can modulate the effects of cortisol on memory (Kuhlmann & Wolf, 2006). An early study by Tennes & Kreye (1985) proposed the suitability of studying cortisol stress responses in children at school; reporting higher cortisol levels on days when students were taking a test as compared to regular school days. Testing conditions can be inherently stressful also for adolescent students. However, the severity of perceived stress, when measured by means of salivary cortisol, varies based on the kind of cognitive tasks demanded (Minkley & Kirchner, 2012). Knowledge reproduction tests were associated with the strongest responses in cortisol; the effect was however more pronounced in males than in females (Minkley & Kirchner, 2012).

Our Goal 4 is to explore the link between learning outcomes and cortisol increases; and between post-hoc testing and cortisol increases. Based on the findings mentioned above, we put forward no specific hypothesis regarding the former relationship, but we do hypothesize that cortisol levels would be elevated in measurements taken after the post-intervention testing (i.e., compared to previous measurements).

1.4 Goal 5: Methodological Issues of Cortisol-based Research

DGBL research raises many questions that can be addressed by various types of experiments (see, e.g., Mayer & Johnson, 2010 for an example of classification of DGBL research). It is known that integrating a serious game within the formal schooling system is difficult, and expectations gained in a laboratory may not materialize in the real world (e.g., Egenfeldt-Nielsen, 2005; Klopfer, 2008). Thus, a substantial portion of studies investigating actual usage of serious games, motivation of learners and learning gains have to be conducted in schools as field studies or in a laboratory, in which a regular school day is plausibly modeled (see Brom et al., 2012 for the rationale). However, cortisol research produces the most reliable data when conducted during the afternoon in a laboratory with carefully controlled timing of the assessment (Dickerson & Kemeny, 2004). Doing this is nearly impossible in real schools. Even if the school day is modeled in a laboratory, the experiment has to be conducted before or around noon. Moreover, cortisol cannot be measured very often nor at a precise time, since that would disrupt the course of activities. When conducting a field or semi-field study, it is hard or impossible to obtain some variables and standardize certain conditions. For instance, the phase of menstrual cycle can influence salivary cortisol responses (Kudielka & al., 2009) and it is common to ask female participants on the day of their menstrual cycle in a cortisol study; this is, however, problematic in case of a teenage girl in a (semi-)field experiment (first, there are ethical concerns; second, students would start to behave differently). Similarly, nicotine, caffeine or glucose intake can influence salivary cortisol responses (Kudielka & al., 2009) and it is nearly impossible to standardize participants' nutritional state prior the experiment by instructing them, the previous day, to avoid certain food or drinks the morning before the experiment (in the case of a (semi-)field study, many participants would not comply or they would refuse to participate). Consequently, high noise in the cortisol data can be expected, and thus, it should be questioned whether salivary cortisol is a method applicable in the context of the DGBL research when conducted in the field or when the real-world context is modeled in a laboratory. Hence, our final goal is to report on these methodological issues so that other DGBL researchers considering inclusion of cortisol measurement techniques in their research can benefit from our experience.

1.5 Structure of the Paper

The paper proceeds as follows. In Section 2, we first describe all five interventions used in this study. Then we will continue with the classical structure: Methods (Section 3) – Results (Section 4) – Discussion (Section 5).

2. Interventions Used

To accomplish our goals, we used five different interventions in this study. Crucially, these interventions differed in the amount of uncontrollability and social-evaluative threat aspects they featured. The first three were derivatives of a team-based serious game, *Europe 2045*, which was successfully integrated into the formal schooling environment at dozens of Czech high schools (Brom et al., 2010). All three treatments lasted about five hours and featured uncontrollability and ST components because learners were required to verbally present certain topics to their peers. The fourth intervention was an interactive computer simulation in which learners learned how to brew beer. This simulation had been developed solely for research purposes. The intervention lasted 2-3 hours and it was for a single "player." Therefore it did not feature an ST component and it featured very little uncontrollability. The fifth intervention was a 15-minute-long, team-based, experiential, non-computer simulation of a car accident. The simulation was embedded in a several-day-long, first aid training course. The simulation had been used successfully over several years in various contexts in the Czech Republic. It featured some aspects of uncontrollability and ST, but less than *Europe 2045*.[3]

Note that while the duration of the treatments differs, we measured cortisol levels, in all treatments, of only certain parts of the treatments. These parts were of similar length: they lasted approximately 15 - 30 minutes each. This is also the precision achievable using salivary cortisol sampling.

2.1 Europe 2045

The game *Europe 2045* was developed for educational purposes in 2008. The game, as used in schools, is fully described in (Brom et al., 2010), where it is also showed that the game is motivating for the learners. Here, we describe only the three game derivatives used for this study. These derivatives were: 1) a full-fledged computer game (EU-comp groups); 2) a classical frontal teaching approach capitalizing on the teaching method used in *Europe 2045* (EU-class groups); and 3) a non-computer game very similar to *Europe 2045* (EU-no-comp groups). The ST and uncontrollability components were present in all the three treatments to the very similar extent.

The treatments descriptions are relatively detailed, because it is important to describe the content of activities for which we measured cortisol response, i.e., the stressors. However, to make the description more concise, several details were included in Appendix A.

[3] Two of the present study's authors (V. Š. and C. B.) are co-authors of *Europe 2045*. One of this paper's authors (C. B.) is a co-author of the beer-brewing simulation. None of the authors of the present study were involved in organizing/developing these first aid training courses.

2.1.1 The EU-comp treatment

This condition featured *Europe 2045* serious game. *Europe 2045* combines the principles of two game genres: multi-player on-line videogames and social role-playing games. The latter is not only played on computers, but also in the classroom. Both games are interconnected.

Europe 2045 is played in student teams, while the teacher assumes the role of coach/tutor. Each student represents a member state of the European Union. At the beginning, the game situation closely resembles the real state of affairs in Europe as of today. The game proceeds in rounds with each round representing one year.

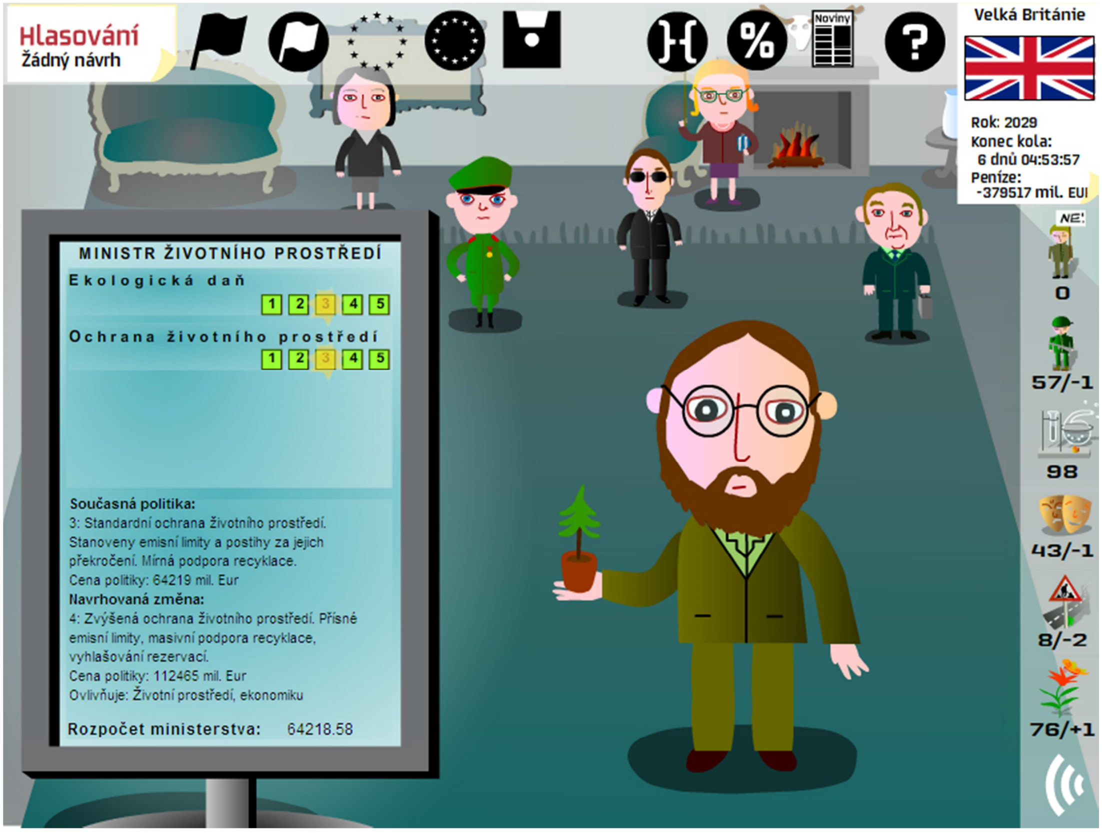
In schools, one possible way how the game can be played is within a “project-day.” In this study, we modeled such a “project-day” in a controlled laboratory environment. The game was played for about five hours. For the study’s purpose, the game was “standardized” (and therefore constrained) as follows. Each time, the game was played by exactly eight players in six rounds. Previously instructed teachers, all members of the research team, participated in our standardized setting (see Sec. 3.4.1). A time schedule with to-the-minute precision was constructed and the teachers did their best to follow it as closely as possible.

Two of three layers of *Europe 2045’s* game play were played: the economic layer and the diplomatic layer. In the economic layer, each student defines the domestic policy of his/her state, such as tax levels and the level of environmental protection (Fig. 1).[4]

In the diplomatic layer, which is most important for the purpose of this study, the player has an opportunity to present drafts for policy changes to the EU (for issues such as common immigration policy, stem-cell research or agricultural quotas). A teacher moderates discussions about these changes and these simulate negotiations at a wide array of EU institutions.

--- Insert Fig. 1 around here ---

[4] The knowledge that can be acquired by this level was not tested by knowledge tests.



Velká Británie

Rok: 2029
Konec kola:
6 dnů 04:53:57
Peníze:
-379517 mil. EUI

MINISTR ŽIVOTNÍHO PROSTŘEDÍ

Ekologická daň

1 2 3 4 5

Ochrana životního prostředí

1 2 3 4 5

Současná politika:
3: Standardní ochrana životního prostředí. Stanoveny emisní limity a postihy za jejich překročení. Mírná podpora recyklace.
Cena politiky: 64219 mil. Eur
Navrhovaná změna:
4: Zvýšená ochrana životního prostředí. Přísné emisní limity, masivní podpora recyklace, vyhlašování rezervací.
Cena politiky: 112465 mil. Eur
Ovlivňuje: Životní prostředí, ekonomiku

Rozpočet ministerstva: 64218.58

Fig. 1 A screenshot from the *Europe 2045* game. The economic layer: GUI of domestic politics settings.

More specifically, each player represents a different project to try to push through at the European level. A project is connected with a number of policies that should be put in place/suspended (e.g., the Green Europe project supports environmental protection and investment into alternative energy resources). A player can always find a teammate to support his/her intended particular policy change. Thus, the game features both collaborative and competitive aspects at the same time.

In the first two “tutorial” rounds, the players were familiarized with the game and they could choose their project and familiarize with it by reading expository texts. Each project was described at about two A4 paper sheets and each of its policies at about 1-2 A4 paper sheets.

In each of the subsequent four rounds (the 3rd to the 6th), the following happened. Players were able to briefly control their states (i.e., play the economic layer). Afterwards exactly four players (randomly selected by a computer) proposed a draft for a policy change chosen of their free will. Each student presented a draft exactly twice during the game. After a period of eight minutes for preparing for the draft’s presentations, students introduced their drafts to their fellow players (4 x 1.5 minutes). Opponents or other proponents could then react/ask questions during a discussion moderated by the teacher (4 x 2-3 minutes). When all the four proposals were presented, the negotiation for or against support of the proposed policy changes started (5 minutes). The negotiation was *not* moderated by the teacher. Finally, students voted on each draft presented. The results were presented at the beginning of the next round; including the current game ranking of the players.

Concerning learning new knowledge, we concentrated only on knowledge that can be acquired by means of the diplomatic layer. We thus operationalized “learning effectiveness” by means of the amount of knowledge about a) the player’s own project; b) all other projects; c) policies each player presented himself/herself; and d) the process of negotiations on policy changes.

Note that there is an inherent ST aspect in the presentations and discussions and these presentations can be consequently quite stressful for some students (as also confirmed by our qualitative data). Moreover, the condition is also uncontrollable by the presenting student to some extent. For instance, the student must have used up the whole 1.5 minutes and present the pros of the proposed draft policy change, despite whether he/she was personally against this particular change (this happened in about 20-50% of cases despite the fact that students could choose their own proposal).

Cortisol was measured when the game was most heated, i.e., after the 4th and the 5th round.

More detailed description of the treatment can be found in Appendix A.

2.1.2 The EU-class treatment

This condition modeled, in a laboratory, a “typical” project day on the topic of European Union, as it would be implemented within a school, without *Europe 2045*. We strove to design the project day so that the learning benefits were maximal for learners (i.e., “the best possible” replacement) and so that the learning experiences in the EU-class and EU-comp conditions were as similar as possible. All expository texts were the same as in the EU-comp groups.

Concerning main differences between the treatments, the introduction to the game was replaced by an unrelated 40-minutes-long frontal lecture on the EU using PowerPoint slides and by an unrelated 20-minute-long, pen-and-paper “heat up” mini-game on the topic of the EU and EU law. Afterwards, we carefully avoided the word “game” (and any competition). Each EU-class learner was paired with an EU-comp (or EU-no-comp) learner and was assigned the peer’s project. Thus the EU-class learners could not choose their projects. They were also instructed “to study a project” rather than “to play a project role.” As concerns policies, each EU-class learner was assigned a policy to study and to present based on what his/her peer had chosen in the EU-comp group (i.e., the possibility of choice was absent). The time for studying and introducing policies was the same as in the EU-comp groups. Similarly to the EU-comp groups, after each presentation, a brief discussion about the proposed policy started. Negotiation was replaced by a longer discussion about all presented policies together. There was no voting and as concerns its replacement, we added an unrelated short film about an EU topic at the very end of the workshop (around 20 minutes long). There was no economic layer.

There are two technical issues worth commenting. First, due to reasons detailed in Appendix A, we had to replace four rounds of the EU-comp treatment with two “rounds” in the EU-class treatment. In both of these “rounds,” each participant prepared him/herself for the presentations that directly followed. In other words, all students in the EU-class treatment presented their drafts for policy changes twice: once in each “round.”

Second, due to the process of assigning participants to subgroups (see Appendix C), there could have been 6-10 students in each EU-class group (and not exactly eight as in the EU-comp groups). Actually, the treatment’s format permitted this easily, since we could either omit a project (6, 7 students) or assign a project twice while assigning other policies not presented in the EU-comp group (9, 10 students).

From the perspective of this study, it was important that the amount of ST remained more or less unchanged between the EU-comp and EU-class treatments. What we did change was the level of “engagement” (expected to be higher in game groups, see Fig. 2). The level of uncontrollability was also probably slightly higher, because the EU-class participants could not choose their project and policy proposals.

More detailed description of the treatment can be found in Appendix A.

--- Insert Fig. 2 around here ---

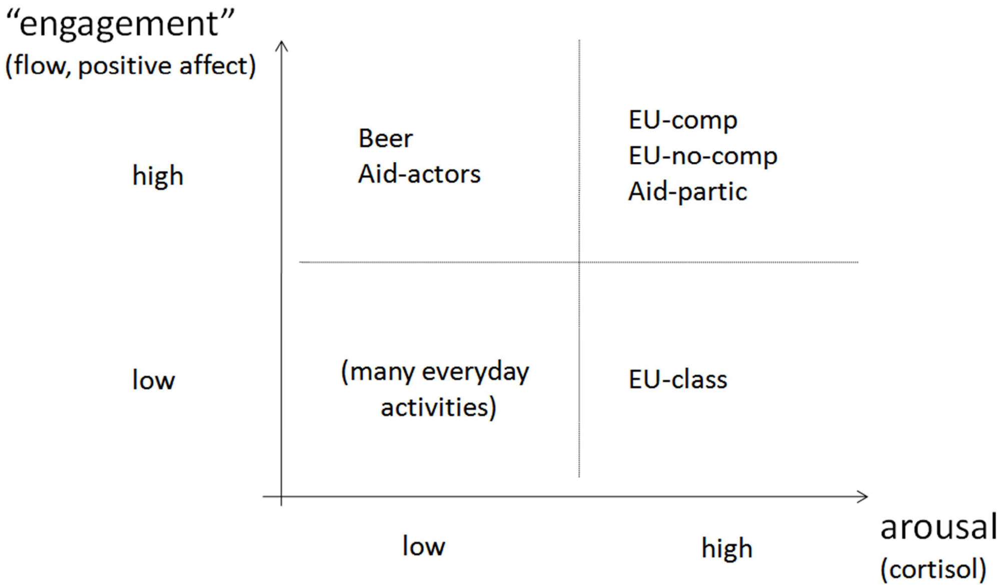


Fig. 2 Hypothesized differences between the six groups in terms of state engagement, operationalized as flow and positive affect, and arousal, measured by cortisol levels. The state engagement is predicted to be high in simulation/game-based intervention, no matter the delivery media. The arousal is predicted to be high in treatments with ST/uncontrollability components.

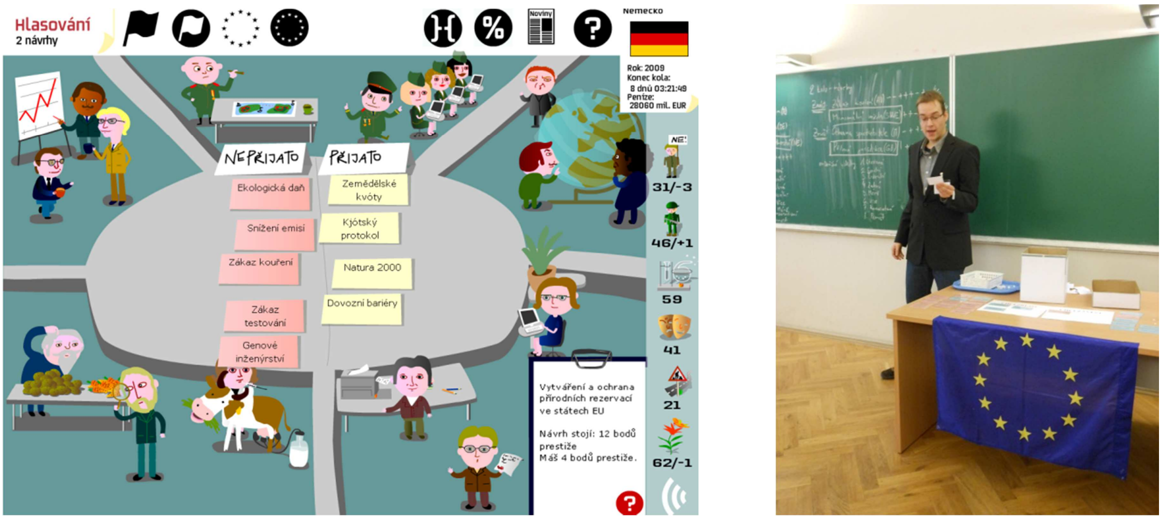
2.1.3 The EU-no-comp treatment

This condition featured *Europe 2045*'s diplomatic layer played without computers. The voting system was implemented in the classroom using a ballot box. It was impossible to replace so easily the game's economic layer, thus it was absent in this condition.[5]

The introduction, the assignment of projects, the discussions and the negotiations were organized as in the EU-comp groups. The schedule and content, including expository texts, were exactly the same as in the EU-comp treatment with the following exceptions: the players did not have the opportunity to play the economic layer. The roughly 15 minutes usually spent by the EU-comp players controlling their state were filled by an extra break and a longer voting process (the votes had to be counted manually in the EU-no-comp treatment). Due to the procedure of an assignment to groups (Appendix C), this treatment was for 6-8 players, each playing a different project.[6]

Remarkably, the amount of ST and uncontrollability remained unchanged.

--- Insert Fig. 3 around here ---



Vytváření a ochrana přírodních rezervací ve státech EU

Návrh stojí: 12 bodů prestiže
Máš 4 bodů prestiže.

(a)

(b)

[5] The economic and diplomatic layers are interconnected, but this connection is relatively weak and can be undone.

[6] We also remark that the EU-no-comp participants actually had slightly more time for studying the expository texts than the EU-comp and EU-class participants. This is because during the "8 minutes for reading", the economic layer was absent. Thus, the four students not preparing themselves for presentation could either do nothing or study materials about policies associated with their own projects or about policies proposed by the other four players. They actually did the latter relatively often; being motivated by the game. This point is relevant in this study as far as possible motivational benefits of the removed economic layer are concerned but not regarding learning outcomes, which will be presented in full detail elsewhere.

Fig. 3 A) Voting interface in *Europe 2045*, i.e., in the EU-comp treatment. Nine ballots for nine drafts of policy changes are depicted. B) A teacher standing next to the ballot box announcing results in the EU-no-comp treatment. There are the most recent proposals written on the board behind him/her (top), as well as the player's latest ranking (bottom). There are ballots for individual drafts of policy changes at the table around the ballot box.

2.2 Beer Brewing Interactive Simulation

While *Europe 2045* is played in a team, the beer brewing interactive simulation is intended for a single learner (referred to as the Beer treatment throughout). During the interaction with the simulation, the learner never presents anything to an audience; therefore, the interaction does not feature an ST component. Also the uncontrollability aspect is minimal. However, at the same time, the simulation is highly engaging; as is also demonstrated later in this paper. Therefore, it is a useful treatment for our experiment, since it is reasonable to speculate that it can increase engaged concentration, but not cortisol levels.

The simulation, including textual instructions, was developed by the research team solely for research purposes (not only for the purpose of the present study). The learning process is relatively long, about 2-3 hours, to enable at least partial comparison to the three *Europe 2045* treatments. The phenomenon being modeled – the process of beer brewing – is motivating enough for participants to stay with the simulation over the whole period (Brom et al., 2014). Because of the complexity of the process, the target audience was, in general, older than in the case of *Europe 2045*; i.e., they were college students.

The simulation is detailed in (Brom et al., 2014). For present purposes, we now provide its brief overview. The simulation was developed using the Netlogo toolkit (Wilensky, 1999). Its graphical interface (Figure 4) consists of the several elements, including: textual instructions, an animation panel showing the content of the fermentation vessel, a supplementary explanation panel relaying the meaning of graphical elements, panels with graphs and histograms showing the amount of ingredients in the product, an adjustable thermometer and buttons for controlling the processes. The application was designed to allow learners to proceed at their own pace. The whole simulation has four parts. These parts include a tutorial, a linear part, an error part and tasks.

1. The tutorial demonstrates how to control the simulation (approx. 10-20 minutes).

2. The linear part demonstrates in a linear fashion how to brew beer from beginning to end, when every step is done correctly (approx. 30-50 minutes).

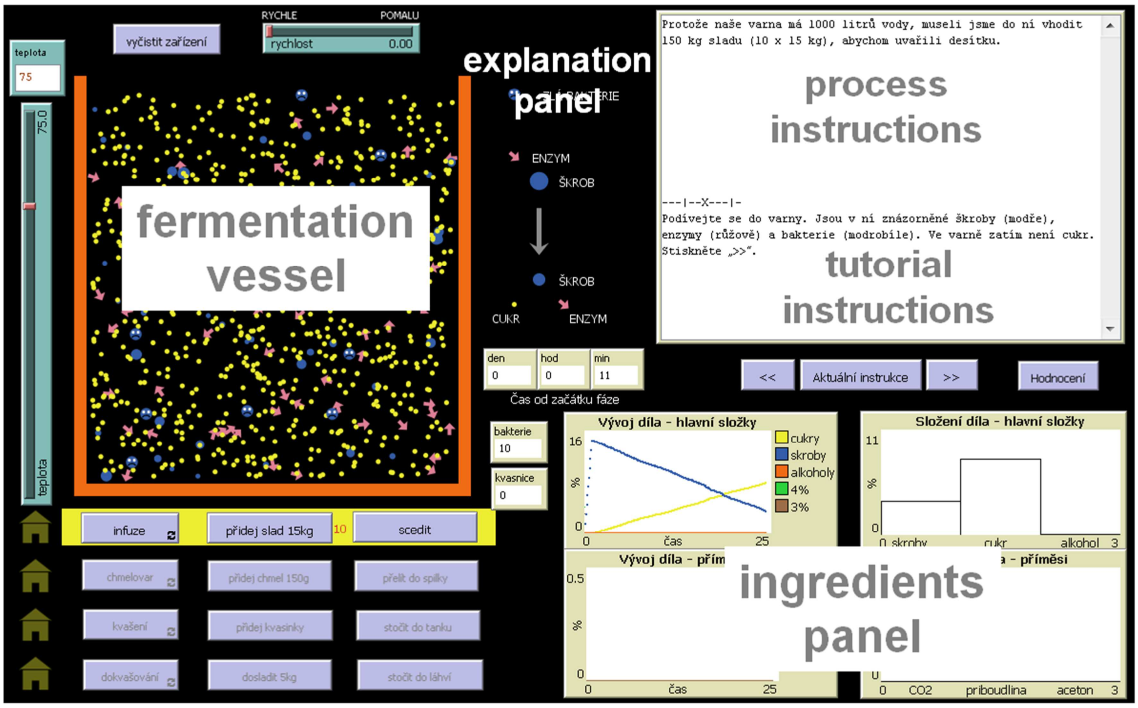
3. The error part demonstrates the consequences of making errors or of not following the standard procedure as previously described (approx. 35-60 minutes).

4. In the final part, so-called tasks, the learner uses the simulation to brew his/her several beers of a specific type (approx. 10-20 minutes to complete one task).

The learner proceeds by reading instructions and conducting suggested steps (except for the final part). The learner has several means of controlling/influencing the simulation (such as adding ingredients, starting/stopping the process, adjusting temperature).

Several key ingredients are animated in the fermentation vessel, such as enzymes, starch or yeast. When things go wrong, bacteria and acetone can appear as well. The learner can monitor the amount of the ingredients through the graphs, histograms and numerical panels. The simulation provides feedback via an “assessment” button.

--- Insert Figure 4 around here ---



Protože naše varna má 1000 litrů vody, museli jsme do ní vhodit 150 kg sladu (10 x 15 kg), abychom uvařili desítku.

---|--X---|-

Podívejte se do varny. Jsou v ní znázorněné škroby (modře), enzymy (růžově) a bakterie (modrobíle). Ve varně zatím není cukr. Stiskněte „>>“.

Fig. 4 The simulation screenshot. The main elements of the graphical interface are described.

To make the simulation more engaging, we included the following background story. 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." (see Table 1)

As a consequence, on-screen instructions are given in a conversational style rather than a formal style; it is the grandpa who is speaking to the learner. The instructions have, in total, 6,750 words.

--- Insert Table 1 around here ---

Table 1 Examples of two instructions. Note every instruction has two parts: a process instruction, explaining the process, and a corresponding tutorial instruction, explaining what to do next. Some words were highlighted using capital letters (in the original).

	Instruction #6	**Instruction #7**
Process instruction	*Excellent! Because the brewing tank holds 1000 liters of water, you had to add 150 kg of malt (10 x 15 kg) to it in order to brew 10-degree beer.*	*Now you 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 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 „>>" and you will find out what happens next.*	*Set the right temperature. Then look at the „Infusion" button. The button can be clicked on or off: in doing so you either start or stop the infusion process. Now try to use the button several times to either start or stop the simulation. Also notice that the TIME INDICATOR, below the image panel, shows the time that has elapsed SINCE THE PHASE BEGAN. Let the infusion run for 5 to 10 minutes and then stop the simulation and click „>>".*

2.3 Car Accident Experiential Simulation

The last treatment complemented the previous four. It was a multi-"player" treatment, yet the amount of uncontrollability and social-evaluative threat was somewhat lower than in the *Europe* 2045 treatments, at least for one type of the last treatment's participants, as detailed below.

A team of physicians and paramedics, called ZDrSEM, organizes first aid courses from 1996.[7] The courses usually last more than one day and their key characteristic is that they include multiple experiential simulations for the participants. The major simulation is often a simulation of a car accident. The simulation happens in the morning; i.e., in the time comparable to the time of the other four treatments.

The course is usually visited by 10-20 different learners. The car accident simulation is for about five learners; thus, it is replayed several times during the morning. One simulation lasts about 15 minutes and involves five or six external actors. The setting involves a car full of passengers hitting a pedestrian or a cyclist. Minor to fatal injuries are simulated, including both minor scratches and disorientation as well as open fractures, arterial bleeding, inner injuries, cardiac arrest and comas. The state of some "wounded" persons progresses spontaneously during the simulation.

Each simulation starts when organizers bring a group of (about five) participants to the area near the simulated accident, after the crash has just happened. The participants' goals are to orient themselves in the situation, prioritize, provide the necessary first aid, e.g., by limiting arterial bleeding or by cardiopulmonary resuscitation, and call an ambulance. The simulation ends with the ambulance's arrival. Generally, from the perspective of learners, the situation is quite messy, complicated and challenging. However, the actors have the course of events, to a large extent, under their control. Participants must cooperate intensely to manage the situation.

Before the simulation starts, the actors put on their make-up while the learners receive a lecture or take part in a discussion. The same actors act in each simulation during a particular day; there is only a few-minutes-long break between consecutive simulations. After the last simulation ends, about an hour-long debriefing starts. Apart from contextualization of the situation by a lecturer, the learners have to express their opinion regarding what they did correctly and what went wrong. The actors contribute by describing the situation from their perspective. The debriefing generally has (intentionally) a positive tone. It usually ends about four hours after the morning lecture started.

Because the experiential simulation is highly immersive and demanding, both for the learners as well as the actors, a high state engagement can be expected in both groups (denoted as Aid-actors and Aid-partic). However, while the evolution of the experiential simulation is only partially controllable by the participants, it is highly controllable by the actors. In general, the simulation experience is also more stressful for the participants than the actors. Social-evaluative threat is not imminent during the simulation, but it could be, to some extent, present during the debriefing for the participants (but less so for the actors). Thus, some cortisol level differences between the participants and actors can be expected.

We point out here that the simulation's target audience is adults; both the actors and the participants are usually 20-40 years old. Saliva was collected in the field during a real course, not in a laboratory like in the previous four treatments.

[7] http://www.zdrsem.cz. The group stems from the Vocational School Lipnice (VSL), a civic organization promoting experiential pedagogy in the Czech Republic for more than two decades now (http://www.psl.cz/). The VSL is member of the Outward Bound organization (http://www.outwardbound.net/).

3. Method

3.1 Experimental Design

The study used between-subject design with six different groups and five different treatments. The treatments were as follows: three derivatives of the multi-player *Europe 2045* educational game described in Sec. 2.1 (EU-comp, EU-class, EU-no-comp; or EU* all together); the single-“player” educational simulation on the topic of beer brewing described in Sec. 2.2 (Beer); and the multi-“player” car accident experiential simulation described in Sec. 2.3 embedded in a first aid course, which featured two groups: actors (Aid-actors) and participants (Aid-partic; or Aid* both together).

The study compared, between the six groups, pre-post and post-delayed differences in cortisol levels and the positive/negative affective states and flow state of the learners measured either during the intervention or right after it. These were the study’s main dependent variables. The study also investigated relationships between these variables and, for the three EU* treatments, the following auxiliary variables: overall learning effects, enjoyment of competition, contentiousness, social interaction anxiety and subjectively assessed likability of the whole treatment.

The EU* treatments lasted, including the introduction and questionnaire administration, around 7 hours. Several tests and inventories were also administered to these participants a month after the intervention. The EU* treatments were investigated in a laboratory, in which a school project day was modeled according to (Brom et al., 2012). The Beer treatment lasted, including the introduction and questionnaire administration, around 4 hours. This treatment was also investigated in a laboratory setting. The Aid* treatments were investigated in the field and the whole data collection lasted around 4 hours.

All treatments started in the morning and ended around noon or in the early afternoon, with respect to cortisol circadian periodicity. Details of the timing are described in Sec. 3.4.

3.2 Participants

3.2.1 EU* Groups

Our aim was to have a relatively heterogeneous sample of adolescent and young adult participants (to recruit people with different personal characteristics; primarily social interaction anxiety and enjoyment of competition). For the EU* treatments, we recruited seven groups of 15 to 26 participants (127 in total). Two groups were formed from college participants (mainly students of computer science or psychology), who participated for a course credit or 400 CZK (around 20 USD) (Mean age = 22.19; *SD* = 2.21). Four groups were formed from older high school students (Mean age = 16.37; *SD* = 0.64). One group was formed from younger high school students (Mean age = 13.25; *SD* = 0.58). On an experimental day, the whole group arrived at once. Each high school group consisted of one class and arrived with the class’ regular teacher. In the high school groups, participants knew each other well. In one college group, most participants also knew each other, because they all studied computer science at the same faculty (though in a different year of study). In the remaining college group, most participants did not know each other in advance.

--- Insert Table 2 around here ---

Table 2 Breakdown of participants.

Intervention	Background of participants	Nr. of subgroups	Nr. of participants *included* (*recruited*)
EU-comp	College	2	males: 10 (10)
			females: 6 (6)
	Lower high school	1	males: 5 (5)
			females: 3 (3)
	Higher high school	2	males: 5 (5)
			females: 9 (11)
EU-class	College	2	males: 11 (11)
			females: 4 (4)
	Lower high school	1	males: 5 (6)
			females: 3 (3)
	Higher high school	4	males: 18 (19)
			females: 12 (13)
EU-no-comp	College	0	
	Lower high school	0	
	Higher high school	4	males: 12 (15)
			females: 14 (16)
Beer	College	N.A.	males: 9 (9)
			females: 7 (7)
Aid-partic	Adults	5	males: 15 (15)
			females: 11 (11)
Aid-actors	Adults	2	males: 4 (4)
			females: 8 (8)

After filling in of a pre-questionnaire and after an introductory lecture, each group was divided into two or three subgroups according to the criteria described in Appendix C. One of the subgroups always received the EU-class treatment. The other received either the EU-comp or the EU-no-comp treatment; or both in the case of three subgroups. Not all participants agreed with taking part in the saliva sampling; those who did not were excluded for the purpose of this study (Tab. 2). A few others were excluded from specific tests due to partly missing data. The high school classes were recruited by a convenience sampling (we used classes whose teachers were willing to participate, and making sure to include diverse classes: both in terms of their quality as well as their subject specialization).[8]

3.2.2 Beer Group

We recruited 16 participants from Charles University in Prague for participation in the Beer treatment. These persons participated in exchange for course credit (mainly students of computer science or psychology) (Mean age = 23.87; *SD* = 4.73). These participants were tested in groups of between 2 to 8 persons, each sitting at separate computer.

3.2.3 Aid* Groups

For participation in the Aid* treatments, we recruited 26 participants and 12 actors from two different first aid courses. In the first course, the car accident simulation was replayed twice; in the second course, it was replayed three times. Each simulation session involved six actors and five or six participants. Participants paid for the course, but they received 200 CZK (around 10 USD) as compensation for participating in the saliva sampling and for filling in of questionnaires. Actors received 200 CZK for acting in the simulation from the course organizers and additional 200 CZK for participating in the experiment from us. Because this was a field study, we did not collect background data on the participants, except for participants' gender. The actors as well as participants were 20-40 years old, with two or three exceptions.

3.3 Materials

3.3.1 Interventions

Interventions are described in Sec. 2. Their usage is detailed in Sec. 3.4.

3.3.2 Pen-and-paper materials

EU* treatments

At the beginning of the experimental day, participants filled in a *pre-questionnaire*. Its purpose was, first, to solicit information about participants' gender, age and time of waking up that day; and second, to provide information about the amount of prior participants' knowledge about the EU. To assess prior knowledge, we used five self-assessment questions as well as four knowledge questions (see Appendix B). Each question was assigned 1-4 or 1-5 or 0/4 points, giving us a possible score in the range of 5-38 (*Mean* = 20.38; *SD* = 5.22).

[8] The EU* part of the present study was conducted as part of a larger experiment involving seven additional older high school classes. However, saliva was not sampled in these additional classes.

To measure the participants' experience of flow during the treatment, we administered a *Flow Short Scale* (FSS; 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. Examples of questions are: "I do not notice time passing," "I am totally absorbed in the discussion" or "I feel I have everything under control." Flow questionnaires were analyzed through T-norms provided with the standardized Flow Short Scale (Rheinberg, 2004). The possible score transformed via T-norms ranges from 21-74.

To obtain information about participants' affective state during the treatment, we administered a *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. We used a state variant of the inventory asking the participants: "The following words describe different feelings and emotions. Read each item and mark to what extent did you experience these feelings during the last discussion." The list of feelings include: "Interested" or "Strong" (positive), and "Distressed" or "Ashamed" (negative). Each scale consists of ten 5-point Likert items with a possible score from 10-50.

FSS and PANAS were always administered in sequence during the game/workshop. The Cronbach alpha was 0.87 for the positive scale and 0.81 for the negative scale of PANAS, and 0.85 for the FSS, across all three EU* treatments. The respective variables will be denoted as *Flow* for the FSS score, *PANAS+* for the score of the positive PANAS subscale and *PANAS–* for the negative PANAS subscale.

After the game/workshop ended, participants filled in a *post-questionnaire*, from which only one question is relevant for the purpose of this study: "How did you like today's workshop compared to a regular school lesson?" The question used a 6-point Likert scale (1 – *much more*; 6 – *much less*) and it will be denoted as *Like* question.

Participants also filled in four different knowledge tests. Together, these tests required participants to select, from a list, words related to the project they played/were assigned to study; draw a mental map of their project; select, from a list, all the drafts of policy changes that were discussed that day; answer two short answer, one open-ended and two multiple-choice questions related to the policy they presented in the second round of discussions in the EU-class groups or in the second or the third round of discussions in the EU-(no-)comp groups (each student presented exactly one draft for policy change in these rounds); and answer two open-ended questions on the process of political negotiation (see Brom et al., in prep.[9] for details). Open-ended questions were scored by two scorers. For present purposes, only the overall *Test score* of these immediate knowledge tests is relevant. This score will be expressed as percentages of the maximum possible score.

Finally, participants also filled in a short version of a *SIAS*, social interaction anxiety inventory, consisting of ten 5-point Likert questions (Kupper & Denollet, 2012). We used a shortened version due to severe time constraints. Question examples are: "I have difficulty making eye-contact with others" or "I worry about expressing myself in case I appear awkward." The resulting score ranges from 0 to 40. The Cronbach alpha was 0.88.

[9] Brom, C., Šisler, V., Buchtová, M., Selmbacherová, T., & Zdeněk Hlávka (under review, 17 July 2014). Positive Affect and Learning in Repeated Academic Controversies: Effects of Social Role-play Gaming.

A month after the intervention, participants filled in a second battery of knowledge tests and several additional inventories, such as a short version of the Big Five Inventory (Rammstedt & John, 2007). Only one of these is relevant for present purposes, the *RCI*, Revised Competitiveness Index (Harris & Houston, 2010). This instrument features 14 items with a 5-point Likert scale that can be divided into two subscales; enjoyment of competition (nine items) and contentiousness (five items). These subscales will be denoted as *RCI.comp* and *RCI.cont*, respectively. The questions include: “I like competition” or “I often try to outperform others” (RCI.comp), and “I try to avoid arguments” or “I often remain quiet rather than risk hurting another person” (RCI.cont). The inventory was administered in the delayed testing session due to the time constraints of the original session. This inventory seems to assess competitiveness as a stable trait (Harris & Houston, 2010). The Cronbach alpha was 0.93 for the RCI.comp and 0.79 for the RCI.cont.

Beer treatment

At the beginning of the experimental day, participants filled in a *pre-questionnaire*. Its main purpose was, first, to solicit information about participants’ gender and age; second, to provide information about the amount of participants’ prior knowledge about the topic of beer brewing. In the context of this study, the only relevant outcome regarding the second point is that all participants were very low-prior knowledge learners.

During their interaction with the simulation, participants filled in the FSS and the PANAS; the same tests as used in the EU* groups. The tests were again administered in immediate succession. The tests were administered twice during the simulation, as detailed in Sec. 3.4.2. As a resulting score, we use the average of the results from the two tests. The Cronbach alpha was 0.79 for the first PANAS+, 0.89 for the second PANAS+, 0.75 for the first PANAS– and 0.67 for the second PANAS–.The Cronbach alpha for the first FSS was 0.87 and it was 0.89 for the second FSS.

After completing the simulation, participants filled in retention tests and transfer tests on the topic of beer brewing. For the purpose of this study, these tests are irrelevant, since it does not make any sense to compare their outcome to the outcome of the EU* knowledge tests.[10]

Aid* treatments

After the car accident simulation, the participants filled in the FSS and the PANAS, the same tests as used in the EU* groups. Actors filled in the same tests in between consecutive simulations. The researchers made notes on the participants’ and actors’ gender.

[10] From a broader perspective, we note that the purpose of administering these knowledge tests was to finalize test questions for a consecutive study with between-subject design that compared the learning effects of two different versions of the beer brewing simulation; one with personalized instructions and the other with formal instructions (see Brom et al., 2014 for details). This consecutive study (n = 75) did not use cortisol sampling, but it did use the two *PANAS* and the two *FSS* instruments in the same way as the present study. For completeness, we will also report here the results from the consecutive study concerning the PANAS and the FSS. In the present study, participants received the personalized version of the simulation, i.e., with the grandpa addressing the learner in a conversational style (see Mayer, 2009 for more on the personalization principle).

3.3.3 Saliva sampling

Sample collection

Commercial sampling devices (Salivette®, Sarstedt) without any saliva-stimulating additives were used to obtain human saliva. Participants were thoroughly instructed how to collect their saliva by putting a cotton roll in the cheek pouch, letting it soak with saliva for approximately 60-80 seconds and re-placing the cotton roll in the device container. The collected material was stored in an ice box, so that the salivary devices were immediately cooled before final storage at -20°C. Prior to analysis, samples were thawed and centrifuged at room temperature at 3000g for 15 minutes to obtain the clear saliva.

Sample analysis

A portion (10 µl of a 1:10 dilution) of the clear saliva was used for the analysis. Analyses were carried out at the Institute for Biochemistry at the University of Veterinary Medicine in Vienna with a highly sensitive enzyme immunoassay kit for salivary cortisol. Samples were assayed in duplicates and cortisol concentrations were assessed by double-antibody biotin-linked enzyme immunoassay (Palme & Möstl, 1997). Duplicate samples with a coefficient of variance > 10% were replicated and considered in the analysis when a coefficient of variance < 10% was achieved. If the sample volume fell below the limit needed to run duplicates or ran out before reaching a coefficient of variance < 10%, the sample was dismissed from the analysis. The average intra- and inter-assay coefficients of variance were less than 10% and 15%, respectively.

3.4 Procedure

3.4.1 EU* Groups

We organized seven different experimental days with the EU* treatments. The course of every day evolved according to a fixed “optimal” schedule (Fig. 5) and the research team followed the schedule as precisely as possible. However, some discrepancies were unavoidable. The beginning of a day differed by +/– 15 minutes and the actual schedule differed from the optimal one by +/– 20 minutes due to accumulations of prolonged/shorter parts of the play or breaks (due to cortisol circadian periodicity, we strove to collect the samples at the same time, and therefore describe these timing details as precisely as possible).

After the introduction, the participants were instructed to avoid eating and drinking anything except still/carbonated water for the next 30 minutes because of the saliva sampling, and to avoid caffeine products during the whole day[11] so as to yield reliable results for the salivary cortisol analysis. The participants were then assigned numbers to keep their data anonymous and they filled in pre-questionnaires. Then the participants received an introductory lecture about the EU (approx. 20 minutes, PowerPoint slides). Participants were seated as if in a regular class. The lecturer was a member of the experimental team.

[11] In an ideal laboratory setting, the nutritional state of each participant would have been standardized for the experiment with pre-assigned eating and drinking schedules from the early morning until the afternoon. Our experiments took place under more natural conditions on a regular school day for all high school participants, requiring us to deal with their daily routines. Students participated voluntarily and, according to our experience, most likely would have refused to continue with the experiment or might have been less motivated under more stringent conditions.

After the lecture ended, the *first saliva sample* was taken. Although there are inter-individual differences in responses, cortisol secretion generally peaks around 20-40 minutes after the onset of a stressor (Dickerson & Kemeny, 2004). Therefore, this sample was related to participants' cortisol responses during the filling in of pre-questionnaires and/or the listening to the first part of the lecture. This sample was considered a pre-exposure condition.

After the sample collection, the class was divided into two or three subgroups; each of which was assigned one of the following treatments: EU-comp, EU-no-comp, EU-class (as described in Sec. 2.1). Participants were matched based on their pre-test score. For every group, we also took care to achieve similar boys/girls ratio in its subgroups (see Appendix C for details).

Each subgroup moved into a different room. The participants were instructed to avoid any interaction with other subgroups' participants until the experiment ended and the research team did its best to prevent such interaction. Each participant was provided a pen and blank paper sheets. In the EU-comp subgroup, each participant was seated at a separate computer.

Each subgroup had its own teacher, who was a member of the experimental team. We used a pool of six teachers: all males younger than 35 years of age, with similar clothing style, short hair and similar speech and teaching styles. These teachers rotated in their positions. Each teacher had an assistant, who administered the questionnaires and helped with technical issues.[12]

After the splitting into subgroups, each subgroup continued as described in Sec. 2.1. until the treatment interaction ended soon after noon. During the introduction to the game, the EU-comp and the EU-no-comp participants were provided badges with flags of their states and flag stands. The same expository texts were provided in all subgroups (students had a personal copy of a material related to their own project and policies, but they could also access materials related to any project).

In the EU-comp and EU-no-comp subgroups, the *second* and the *third saliva samples* were collected after the fourth and fifth round, respectively. The third sample was considered as reflecting the main treatment effect. In the fifth round, the game was usually the most "heated". The sampling was scheduled at around 30-40 minutes after the discussion started and at least 10-15 minutes after the negotiations ended, so that the samples were related to participants' cortisol responses to the discussions plus negotiations. Keeping in mind the intra- and inter-individual nature of cortisol responsiveness in humans, in our setting with +/– 5 minutes in the different subgroups, it is difficult to separate cortisol responses from the discussion and the negotiation parts. About half an hour before every collection, we instructed participants to avoid smoking or ingesting any food and drink except for still/carbonated water, which was provided throughout the day (see Fig. 5). Cookies were also provided throughout the day, except during the "no food" periods, and a light meal (a baguette) was provided during the lunch break.

In the EU-class subgroups, the second and the third samples were collected after the first and second round of discussions, respectively (see Fig. 5). As already said, we strove to collect the samples at the same time; but +/– 20 minutes differences were unavoidable due to treatment properties and naturally

[12] Each subgroup also had one independent research observer, who coded students' verbal and non-verbal behavior during the discussions. These data are irrelevant for present purposes, but we want to emphasize the presence of another person in the room.

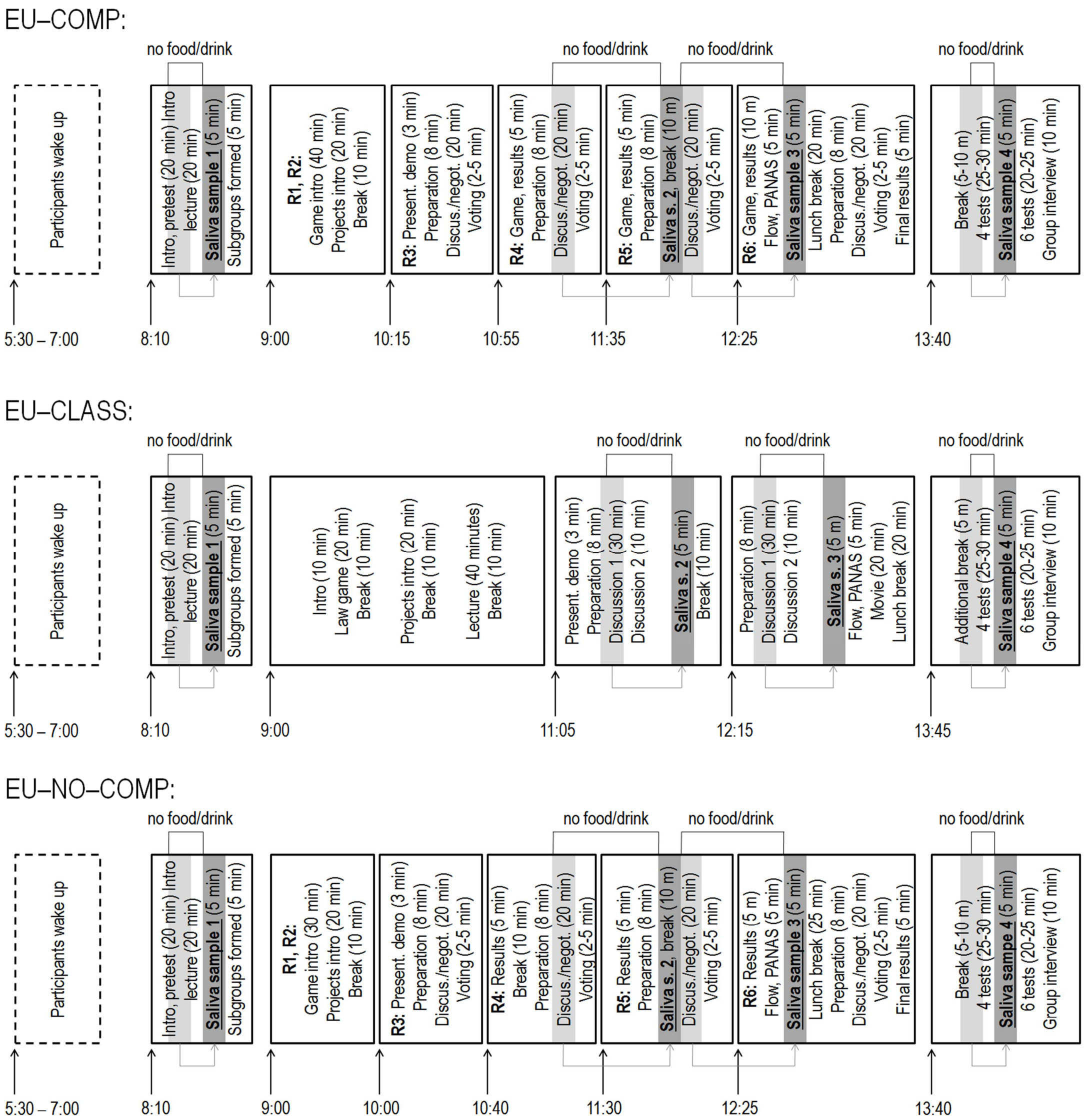
occurring delays or accelerations in different subgroups. Especially in the EU-class subgroups, we usually collected the second and the third samples slightly later than in the other subgroups.[13]

The FSS and the PANAS were administered prior to the saliva sampling (Fig. 5).

After the treatment and after a short break, post-questionnaires and the battery of knowledge tests, including the SIAS, were administered. Each subgroup was tested in its own room.

--- Insert Fig. 5 around here ---

[13] Due to the pairing of EU-class participants with either EU-comp or EU-co-comp participants, the discussions could not start in the EU-class subgroups until proposals from the respective EU-comp or EU-no-comp class were known.



EU–COMP:

no food/drink
no food/drink
no food/drink
no food/drink
Participants wake up
Intro, pretest (20 min) Intro lecture (20 min) Saliva sample 1 (5 min) Subgroups formed (5 min)
R1, R2: Game intro (40 min) Projects intro (20 min) Break (10 min)
R3: Present. demo (3 min) Preparation (8 min) Discus./negot. (20 min) Voting (2-5 min)
R4: Game, results (5 min) Preparation (8 min) Discus./negot. (20 min) Voting (2-5 min)
R5: Game, results (5 min) Preparation (8 min) Saliva s. 2, break (10 m) Discus./negot. (20 min) Voting (2-5 min)
R6: Game, results (10 m) Flow, PANAS (5 min) Saliva sample 3 (5 min) Lunch break (20 min) Preparation (8 min) Discus./negot. (20 min) Voting (2-5 min) Final results (5 min)
Break (5-10 m) 4 tests (25-30 min) Saliva sample 4 (5 min) 6 tests (20-25 min) Group interview (10 min)
5:30 – 7:00
8:10
9:00
10:15
10:55
11:35
12:25
13:40

EU–CLASS:

no food/drink
no food/drink
no food/drink
no food/drink
Participants wake up
Intro, pretest (20 min) Intro lecture (20 min) Saliva sample 1 (5 min) Subgroups formed (5 min)
Intro (10 min) Law game (20 min) Break (10 min) Projects intro (20 min) Break (10 min) Lecture (40 minutes) Break (10 min)
Present. demo (3 min) Preparation (8 min) Discussion 1 (30 min) Discussion 2 (10 min) Saliva s. 2 (5 min) Break (10 min)
Preparation (8 min) Discussion 1 (30 min) Discussion 2 (10 min) Saliva s. 3 (5 m) Flow, PANAS (5 min) Movie (20 min) Lunch break (20 min)
Additional break (5 m) 4 tests (25-30 min) Saliva sample 4 (5 min) 6 tests (20-25 min) Group interview (10 min)
5:30 – 7:00
8:10
9:00
11:05
12:15
13:45

EU–NO–COMP:

no food/drink
no food/drink
no food/drink
no food/drink
Participants wake up
Intro, pretest (20 min) Intro lecture (20 min) Saliva sample 1 (5 min) Subgroups formed (5 min)
R1, R2: Game intro (30 min) Projects intro (20 min) Break (10 min)
R3: Present. demo (3 min) Preparation (8 min) Discus./negot. (20 min) Voting (2-5 min)
R4: Results (5 min) Break (10 min) Preparation (8 min) Discus./negot. (20 min) Voting (2-5 min)
R5: Results (5 min) Preparation (8 min) Saliva s. 2, break (10 m) Discus./negot. (20 min) Voting (2-5 min)
R6: Results (5 m) Flow, PANAS (5 min) Saliva sample 3 (5 min) Lunch break (25 min) Preparation (8 min) Discus./negot. (20 min) Voting (2-5 min) Final results (5 min)
Break (5-10 m) 4 tests (25-30 min) Saliva sampe 4 (5 min) 6 tests (20-25 min) Group interview (10 min)
5:30 – 7:00
8:10
9:00
10:00
10:40
11:30
12:25
13:40

Fig. 5 Time schedules for the EU* conditions. The dark gray color denotes sampling periods, while the light gray color denotes the periods to which the individual samples relate. “R” denotes the round.

--- Insert Fig. 6 around here ---

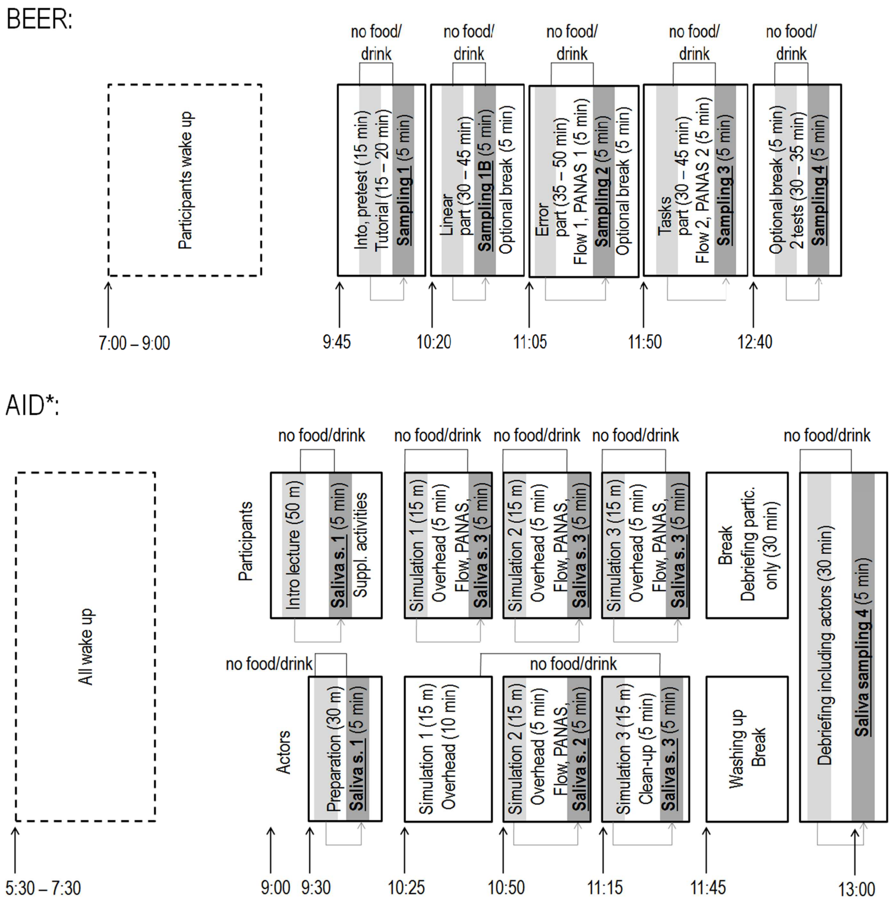

BEER:
no food/ drink
no food/ drink
no food/ drink
no food/ drink
no food/ drink
Participants wake up
Into, pretest (15 min) Tutorial (15 – 20 min) Sampling 1 (5 min)
Linear part (30 – 45 min) Sampling 1B (5 min) Optional break (5 min)
Error part (35 – 50 min) Flow 1, PANAS 1 (5 min) Sampling 2 (5 min) Optional break (5 min)
Tasks part (30 – 45 min) Flow 2, PANAS 2 (5 min) Sampling 3 (5 min)
Optional break (5 min) 2 tests (30 – 35 min) Sampling 4 (5 min)
7:00 – 9:00
9:45
10:20
11:05
11:50
12:40
AID*:
no food/drink
no food/drink
no food/drink
no food/drink
no food/drink
Participants
All wake up
Intro lecture (50 m) Saliva s. 1 (5 min) Suppl. activities
Simulation 1 (15 m) Overhead (5 min) Flow, PANAS, Saliva s. 3 (5 min)
Simulation 2 (15 m) Overhead (5 min) Flow, PANAS, Saliva s. 3 (5 min)
Simulation 3 (15 m) Overhead (5 min) Flow, PANAS, Saliva s. 3 (5 min)
Break Debriefing partic. only (30 min)
Debriefing including actors (30 min) Saliva sampling 4 (5 min)
no food/drink
no food/drink
Actors
Preparation (30 m) Saliva s. 1 (5 min)
Simulation 1 (15 m) Overhead (10 min)
Simulation 2 (15 m) Overhead (5 min) Flow, PANAS, Saliva s. 2 (5 min)
Simulation 3 (15 m) Clean-up (5 min) Saliva s. 3 (5 min)
Washing up Break
5:30 – 7:30
9:00
9:30
10:25
10:50
11:15
11:45
13:00

Fig. 6 Time schedules for the Beer and Aid* conditions. The dark gray color denotes sampling periods, while the light gray color denotes the periods to which the individual samples relate.

The *fourth saliva* sample was collected 30-40 minutes after the treatment interaction ended, so that the cortisol response related to the break and the first part of testing. This sample was considered a post-exposure condition.

Around a month later, we entered the school to administer subsequent knowledge tests and a few inventories, including the RCI. Students were not informed in advance. The testing period lasted 90 minutes (i.e., two 45-minute school lessons). We followed recommendations of Brom et al. (2012) for posttests administration, including dividing the class into 5-6 small groups, each with its own administrator. Usually, only two groups were tested in one room.

3.4.2 Beer Group

Each participant was seated at a separate computer. After the introduction, participants filled in the pre-questionnaire. Afterwards, they were introduced to the simulation by the experimenter. Then they interacted with the tutorial part of the simulation (Fig. 6).

The *first saliva sample* was collected after the tutorial interaction ended. This sample represented participants' cortisol responses while filling in the pre-questionnaires and the start of their interfacing with the tutorial. These cortisol data were considered a pre-exposure condition and matched with the first sample from other treatments. The slight time shift compared to the other groups was caused by the fact that the participants from the Beer group woke up later.

After the sampling, participants interacted with the linear part of the simulation. Then another sample was taken, related to the cortisol response on the linear part interaction. This sample will be denoted as *1B* because it was not matched with other treatments' samples.

Then the participants interfaced with the error part of the simulation and finally were assigned tasks to solve within the simulation (see Sec. 2.2). They were assigned several tasks (usually two or three) so that they solved the tasks over a period of at least 30 minutes.

The *saliva sample no. 2* was taken after the error part and the *sample no. 3* was taken after solving the tasks. Each of them was related to the cortisol response to the learners' interaction with the simulation. Note that technically these were the third and fourth samples, respectively, but we mark them as "2nd" and "3rd", because they will be matched with the samples with the same numbers from other treatments. The sample no. 3 was considered as reflecting the main treatment effect, because solving the tasks was most interesting for the learners.

Finally, participants filled in two knowledge tests and the last saliva sample, *sample no. 4*, was collected. This sample is related to the cortisol response to filling in of the tests and it was considered a post-exposure condition.

A break was offered, but not insisted on, after the linear, error and tasks parts.

During the introduction, the participants were instructed to avoid caffeine products and smoking during the whole experiment. About half an hour before every collection, we instructed the participants to avoid any food and drink except for still/carbonated water.

3.4.3 Aid* Groups

We were invited to conduct our experiment in two different first-aid courses on two different days. On the first experimental day (D1), participants formed two subgroups. On the second one (D2), they formed three subgroups. Thus, in D1, the simulation was replayed twice; in D2, it was replayed three times. Each participant was assigned to one subgroup. Apart from the simulation participants and actors, around four ZDrSEM members and two members of our research team (one responsible for the actors and the other for participants) were present each day.

Before the actual car accident simulation, all participants listened to an introductory lecture given by a member of the ZDrSEM team. During the lecture, they were also informed that they could participate in our experiment if they so wished and were introduced to our research staff. Actors were informed about the experiment in advance.

The actors prepared themselves for the simulation such that the participants did not know about them. Participants were informed about the simulation at the end of the lecture, but the actual depth of the simulation came as surprise to them.

The *saliva sample no. 1* was collected immediately after the lecture (for the participants) and after the preparatory phase (for the actors), so that this pre-exposure sample was related to cortisol response to listening to the lecture (participants) or to preparing for the simulation (actors). This sample will be matched with the first sample from other treatments.

Subgroups proceeded one after another to the simulation. The simulation started when the participants spotted the car. One simulation session lasted around 15 minutes and the preparation of actors for the consecutive simulation session lasted around 10 minutes.

After the simulation ended, the participants were brought back to the lecture room by ZDrSEM organizers where they were administered the FSS and the PANAS by our experimenter. The experimenter also collected *saliva sample no. 3* (note there is no sample no. 2 in the case of participants). The sampling occurred about 10 minutes after the stressor ended, i.e., around 25 minutes after its onset. Because the time schedule was tight, it was not possible to wait longer, even though 30-40 minutes would have perhaps been better than 25 minutes.

Meanwhile, the actors prepared themselves for the consecutive simulation. In D1, *saliva sample no. 2* was collected after the first simulation and *saliva sample no. 3* after the second simulation. In D2, *saliva sample no. 2* was collected after the second simulation and *saliva sample no. 3* after the third simulation. Sampling always took place at the end of the preparation phase, so that the difference between the stressor onset (i.e., a simulation start) was maximum. Usually, this interval was 20-25 minutes. Sample no. 3 will be considered the main one, as in the case of other treatments. In both D1 and D2, the FSS and the PANAS were filled in after the second simulation.

A subgroup of participants waiting for the simulation or having just completed saliva sampling was engaged by ZDrSEM organizers in supplementary non-stressful activities.

At the end, participants were debriefed. Actors took off their make-up, took a short break and were then engaged in the debriefing. Eventually, *sample no. 4* was collected from both actors and participants. This sample will be considered a post-exposure sample.

About half an hour before every sample collection, both participants and actors were reminded to avoid eating and drinking anything except still/carbonated water for the next half an hour.

In general, sample collection occurred earlier in the Aid* groups than in the Beer or EU* groups.

3.5 Data Analysis

All analyses were conducted in statistical program R 3.0.0 (R Core Team, 2013). Cortisol samples were checked for outliers. We discarded all values higher than 80 ng/ml (with the exception of the first sample, where we discarded values higher than 100 ng/ml), because cortisol levels decline during the day and such high values may have been caused by blood contaminations. We transformed cortisol values using natural logarithm because the primary values were not normally distributed. After the transformation, we tested normality using Shapiro-Wilk test (Shapiro & Wilk, 1965) and the results suggested that the data were normally distributed ($W = 1.00$; $p = 0.461$). In the all following sections, we will use only log-transformed values.

Correlations were expressed by Pearson correlation coefficient. Effect sizes for correlation were classified according to Cohen (1988) into small (r = 0.1), medium (r = 0.3) and large (r = 0.5). Differences between two treatments were tested using two-sample t-test, or using t-test with Welch correction in case of unequal variance (Welch, 1947). Differences between three or more treatments were tested using one-way ANOVA followed by *post hoc* analysis using Tukey HSD test, which corrects p values for multiple comparisons (Miller, 1981). In Section 4.2 we used Tukey-Kramer method for post hoc testing differences between treatments (due to unequal sample sizes, see Jaccard et al., 1984). Differences between genders while taking other variables into account were tested using two-way ANOVA. Prior to one-way or two-way ANOVA analysis, we tested homogeneity of variance using Levene test (Levene, 1960). All measured groups had equal variance.

Effect sizes for t-tests, Tukey HSD test and Tukey Kramer method were expressed by Cohen's d with classification into small (Cohen's d = 0.2), medium (Cohen's d = 0.5) and large (Cohen's d = 0.8). Effect sizes for one-way ANOVA were expressed by *partial* η^2 (Fritz et al., 2012) with classification into small (η^2 = 0.01), medium (η^2 = 0.06) and large (η^2 = 0.14).

Because we were interested in cortisol level differences between the pre-exposure sample (nr. 1) and the main treatment sample (nr. 3), and also between the main treatment sample (nr. 3) and the post-exposure sample (nr. 4), we used the following variables in the cortisol data analysis. Variable *3–1* denoted the difference between sample nr. 3 and sample nr. 1 while variable *4–3* denoted the difference between sample nr. 4 and sample nr. 3 (after the individual samples' log transformation). To express general cortisol trend during the day, we used also *(3–1)–(4–3)* variable, which is the difference between *3–1* and *4–3* variables. This variable is similar, though not equivalent, to the variable often called "the area under the curve with respect to increase," when distances between the 1st and 3rd, and between the 3rd and 4th samples are defined as unitary (Pruessner et al., 2003).

4. Results

4.1 Do Learners' Self-reported Affective State and Learning Effects Differ Between Treatments?

As can be seen in Table 3, one-way ANOVA revealed significant differences between treatments for Flow, PANAS+ and Like question. Differences between treatments approached significance for PANAS–. No difference was found concerning Test score for EU* treatments.[14] Results from post hoc comparisons using Tukey HSD test are summarized in Tables 4 – 6. In addition, for Like question, the post hoc test found significant differences between the EU-class and the EU-comp treatments (p < 0.05; d = 0.59). As concerns Flow and PANAS+, a general pattern is that participants in all treatments, except for EU-no-comp, scored significantly higher in Flow and PANAS+ than EU-class participants. In addition, Aid-actors participants scored significantly higher in Flow than EU-comp participants, and Beer and Aid* participants scored significantly higher in Flow than EU-no-comp participants. No differences were revealed for PANAS–.

[14] Here, we report the learning outcome results only for participants that underwent cortisol sampling. We point out that learning outcomes were actually investigated on a larger sample (see Sec. 3.2.1) and that the key variables of interest were not related to immediate learning outcomes, but rather to learning achievements measured a month after the treatment. These results will be reported in detail elsewhere.

--- Insert Table 3 around here ---

Table 3 Means and SDs for the investigated affective variables and the Test score variable for every treatment, and also *F* value with the corresponding *p* value of between-treatment comparison (one-way ANOVA).

	EU-comp		**EU-no-comp**		**EU-class**		**Beer**		**Aid-partic**		**Aid-actors**			
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*F*	*p*
Flow	50.97	6.14	47.00	8.85	45.57	7.85	55.38	6.67	53.04	7.53	58.58	7.39	10.35	<0.001
PANAS+	30.20	6.44	28.50	8.52	24.57	6.39	32.97	5.76	33.92	6.30	34.08	8.74	9.42	<0.001
PANAS-	17.24	5.40	18.13	6.51	17.86	6.44	14.59	2.33	19.25	4.93	14.25	4.96	2.21	0.056
Like	5.26	1.00	5.12	0.86	4.63	1.19	-	-	-	-	-	-	4.39	0.015
Test Score	0.63	0.12	0.56	0.10	0.55	0.16	-	-	-	-	-	-	2.10	0.130

--- Insert Table 4 around here ---

Table 4: Results of post hoc comparisons for Flow (Tukey HSD test).

Flow	EU-comp	EU-class	EU-no-comp	Beer	Aid-partic	Aid-actors
EU-comp	-					
EU-class	-0.75*	-				
EU-no-comp	-0.54	0.18	-			
Beer	0.70	1.29***	1.03**	-		
Aid-partic	0.31	0.96***	0.73*	-0.32	-	
Aid-actors	1.18*	1.68***	1.37***	0.46	0.74	-

Note: Each cell contains effect size of the difference between conditions. Minus sign in before the effect size indicates that the treatment in the column is larger than the treatment in the row.

**p* < .05 ***p* < .01 ****p* < .001

--- Insert Table 5 around here ---

Table 5 Results of post hoc comparisons for PANAS+ (Tukey HSD test).

PANAS+	EU-comp	EU-class	EU-no-comp	Beer	Aid-partic	Aid-actors
EU-comp	-					
EU-class	-0.88**	-				
EU-no-comp	-0.23	0.55	-			
Beer	0.44	1.34***	0.59	-		
Aid-partic	0.58	1.47***	0.72.	0.16	-	
Aid-actors	0.55	1.39***	0.65	0.16	0.02	-

Note: Each cell contains effect size of the difference between conditions. Minus sign in before the effect size indicates that the treatment in the column is larger than the treatment in the row.

·*p* < .1 ***p* < .01 ****p* < .001

--- Insert Table 6 around here ---

Table 6 Results of post hoc comparisons for PANAS– (Tukey HSD test).

PANAS-	EU-comp	EU-class	EU-no-comp	Beer	Aid-partic	Aid-actors
EU-comp	-					
EU-class	0.10	-				
EU-no-comp	0.15	0.04	-			
Beer	-0.56	-0.57	-0.66	-		
Aid-partic	0.39	0.23	0.19	1.13	-	
Aid-actors	-0.56	-0.58	-0.64	-0.09	-1.01	-

Note: Each cell contains effect size of the difference between conditions. Minus sign before the effect size indicates that the treatment in the column is larger than the treatment in the row. Note the differences are not significant despite large effect sizes due to correction for multiple testing (e.g., *p* = 0.12 for the difference between the Beer and Aid-partic treatments).

Exploratory correlation analysis (Tab. 7) revealed positive significant correlations between Flow and PANAS+ (medium to very large effect size) and negative significant correlations between Flow and PANAS– (medium to large effect size). Noteworthy, the results are similar to results of our different experiment that used the Beer treatment (*n* = 75; *r* (Flow, PANAS+) = 0.57; *r* (Flow, PANAS–) = –0.49) (Brom et al., 2014; see also Footnote (10)). Moreover, Like question strongly correlates with Flow and PANAS+ for EU* treatments. Except for Flow, PANAS– does not correlate substantially with any other variable. We remark that PANAS– is largely orthogonal to PANAS+ (Watson et al., 1988), therefore, it is not surprising we found only small correlation between PANAS– and PANAS+.

Concerning EU* treatments, we found small to medium positive link between Test score and Flow, and between Test score and PANAS+. A similar relationship, both for Flow and PANAS+, was also found in the second experiment using the Beer treatment mentioned above (Brom et al., 2014).

There were no gender differences in affective variables and Test score with the exception of RCI.comp where males scored significantly higher than females ($t(99) = 4.86$; $p < 0.001$; $d = 0.98$). This subscale measures enjoyment of competition, thus this result is not very surprising and it is also consistent with past results (Houston et al., 2005).

To conclude, the data indicated that there were between-treatment differences in learners' self-reported positive affective state (and flow state) between about half of the treatments. The result is more complex than predicted on Fig. 2, but the EU-class treatment elicited the lowest "engagement", as predicted, both in terms of Flow and PANAS+.

--- Insert Table 7 around here ---

Table 7 Correlation matrices of affective variables and Test score.

	All					EU*				
	Flow	PANAS+	PANAS-	Like	Test sc.	Flow	PANAS+	PANAS-	Like	Test sc.
Flow	-					-				
PANAS+	0.66***	-				0.68***	-			
PANAS-	-0.43***	-0.22**	-			-0.41***	-0.19*	-		
Like	0.51***	0.52***	-0.09	-		0.51***	0.52***	-0.09	-	
Test score	0.29*	0.32**	0.10	0.17	-	0.29*	0.32**	0.10	0.17	-

	Beer			Aid-actors			Aid-partic		
	Flow	PANAS+	PANAS-	Flow	PANAS+	PANAS-	Flow	PANAS+	PANAS-
Flow	-			-			-		
PANAS+	0.43	-		0.07	-		0.63***	-	
PANAS-	-0.43.	-0.09	-	-0.07	-0.28	-	-0.63***	-0.38.	-

Note: Like and Test score variables relate only to the EU* treatments.

·p < .1 *p < .05 **p < .01 ***p < .001

4.2 Does the Cortisol Response Differ Between Treatments?

At first, we tested if cortisol levels differ between genders. Two-way ANOVA (gender x sample) revealed no differences between males and females ($F(1, 636) = 0.45$; $p > 0.1$; $\eta_p^2 = 0.00$)), nor the interaction with sample ($F(3, 636) = 1.73$; $p > 0.1$; $\eta_p^2 = 0.01$). However, when we tested for gender differences in *3–1* and *4–3* variables directly, we found that there is significantly higher decrease in cortisol in males compared to females, as measured by *3–1* variable ($t(154) = –2.42$; $p < 0.05$; $d = 0.39$), and significantly lower decrease in cortisol in males compared to females, as measured by *4-3* variable ($t(159) = 2.65$; $p < 0.01$; $d = 0.42$). This indicates that gender may play a moderating role. Therefore, this and the following section will first examine data when both genders are combined and afterwards inspect the issue of gender differences.

Changes in cortisol levels, with males and females combined, can be seen on Figure 7. One-way ANOVA found marginally significant between-treatment differences in *3–1* variable ($F(5, 150) = 2.14$; $p < 0.1$; $\eta_p^2 = 0.07$) (see Fig. 8). The Tukey-Kramer post hoc tests revealed significant differences between Beer and EU-comp ($p < 0.01$; $d = 0.93$), Aid-actors and EU-comp ($p < 0.05$; $d = 0.88$), and Aid-partic and Beer treatments ($p < 0.05$; $d = 0.86$).

Because of our Goal 2, we aimed at investigating differences between treatments featuring a higher level of ST/uncontrollability and the Beer treatment with the lowest level of these elements. Therefore, we also run directly t-test (with Welch approximation for unequal variances) between EU* and Beer, and between Aid-partic and Beer treatments. The results were strongly significant for the former comparison ($t(31.09) = 3.27$; $p < 0.01$; $d = 0.58$) as well as for the latter comparison ($t(34.12) = 2.86$; $p < 0.01$; $d = 0.86$), which indicates that there is a difference in cortisol response between treatments with high and low levels of ST/uncontrollability.[15] The differences can be inspected visually on Fig. 8, on which also the low *3–1* values for the second group with relatively low levels of ST/uncontrollability, Aid-actors, are apparent. Considered together, these outcomes tend to agree with our prediction schematized on Fig. 2 that high ST/uncontrollability treatments would elicit higher cortisol values.

One-way ANOVA found significant between-treatment differences in *4–3* variable ($F(5, 157) = 4.25$; $p < 0.01$; $\eta_p^2 = 0.12$). Post hoc comparison showed significant differences between EU-class and EU-comp ($p < 0.05$; $d = 0.72$), EU-comp and Beer ($p < 0.01$; $d = 1.01$), and EU-comp and Aid-actors treatments ($p < 0.01$; $d = 1.08$). Similarly as in the case of *3–1* variable, we also used t-test with Welch approximation, which found significant difference between EU* and Beer treatments ($t(25.14) = –2.46$; $p < 0.05$; $d = 0.47$). In general, it seems that the cortisol levels tended to increase during the period of filling in of tests/debriefing for participants engaged in one of the following three treatments: Beer, Aid-actors and partly EU-class. These are treatments, except for the EU-class, in which the level of ST/uncontrollability was relatively low and during which cortisol levels, as measured by *3–1* variable, tended to decrease the most, creating a V-shape cortisol function (see Fig. 7).

[15] The t-test results seem to be "stronger" than the results of the Tukey-Kramer post hoc test in that the Tukey-Kramer test revealed significant differences neither between EU-no-comp and Beer, nor between EU-class and Beer, while the t-tests returned quite low *p* values. However, we stress that, contrary to the Tukey-Kramer post hoc test, the t-tests are uncorrected for multiple comparisons, so these results should be treated with caution.

--- Insert Figure 7 around here ---

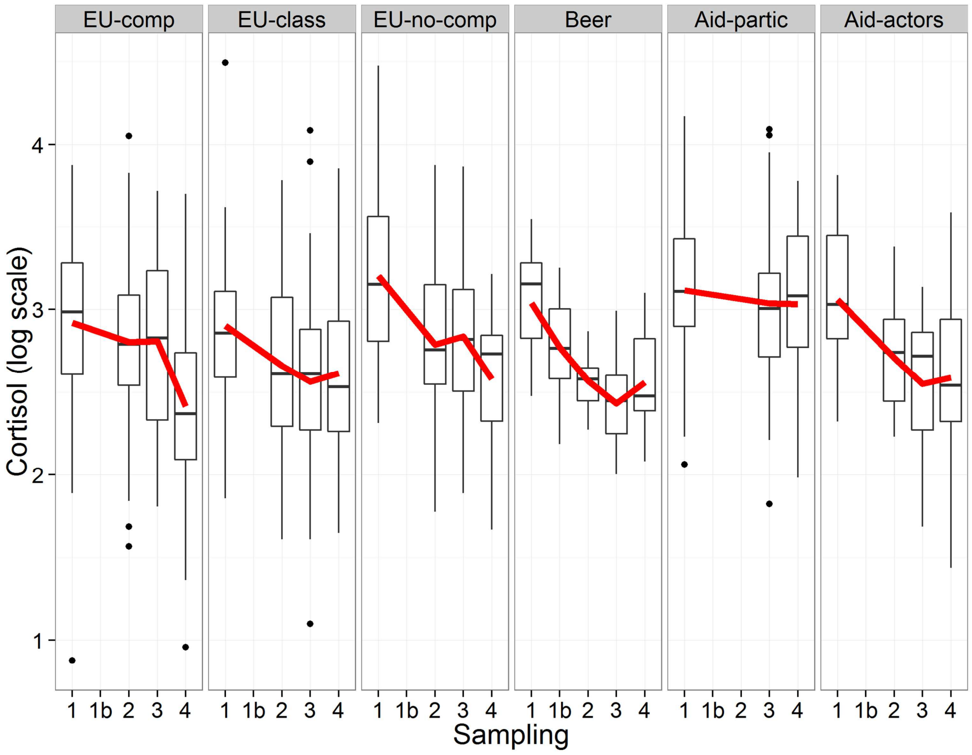


Fig. 7 Cortisol levels for each sampling and treatment. Each boxplot shows the 1st and 3rd quartile (the upper and the lower "hinge"), the bold lines show median. The thick red line connects means. The upper whiskers show values 1.5*IQR from the upper hinge and the lower whisker 1.5*IQR from the lower hinge. More extreme values are shown specifically as dots.

--- Insert Figure 8 around here ---

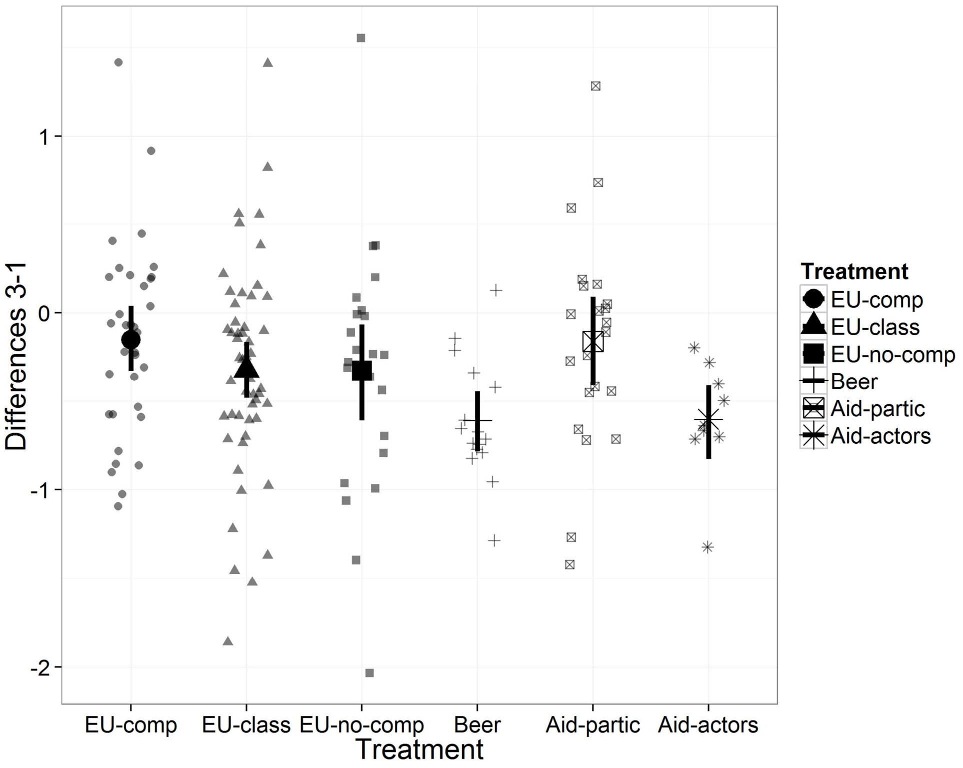


Fig. 8 Cortisol *3–1* variable for each treatment. Each mean value is depicted with a larger symbol and lines denote standard errors of the mean.

As concerns interactions between gender and treatment, these were not significant.

For confirmatory purposes, we also correlated cortisol differences *3–1* and *4–3* with time when participants woken up. These correlations were non-significant for both cases.

4.3 Does the Cortisol Response Relate to Affective Variables?

Because of Goals 2 – 4, we inspected correlations between *3–1* variable and Flow, PANAS+, PANAS–, SIAS, RCI.comp, RCI.cont, Like variable and Test score. We also inspected correlations between these variables and *4–3* variable. Correlations for Flow, PANAS+ and PANAS– were computed for the EU*, Beer and Aid* treatments separately, correlations for the remaining variables were computed only for the EU* treatments. We started with both genders combined. These results are summarized in Table 8.

SIAS and PANAS– significantly correlated with *3–1* and *4–3* variables. Both RCI variables, enjoyment of competition and contentiousness, also correlated with *3–1* and *4–3* variables (although RCI.comp only marginally), but in the opposite direction than SIAS/PANAS–. Thus, the general pattern seems to be that for participants with higher social interaction anxiety and higher negative affective state (PANAS–), the cortisol levels tended to increase during the treatment exposure and tended to decrease during the period of filling in of tests; while this pattern is reversed for participants who liked competition. The existence of this pattern can be also explored by correlating the variables in question with *(3–1)–(4–3)* variable, which describes cortisol trends for individual participants across the whole experiment. The results, also in Tab. 8, indeed support the existence of this pattern. The existence of this pattern would be even more supported if SIAS and PANAS– are positively related whereas RCI variables and SIAS/PANAS– variables are negatively related. We indeed found large negative correlations between RCI.cont and SIAS is (*r* = –0.50; *p* < 0.001) and between RCI.comp and SIAS (*r* = –0.50; *p* < 0.001), and moderate negative correlations between RCI.cont and PANAS– (*r* = –0.22; *p* < 0.05), and positive between SIAS and PANAS– (*r* = 0.38; *p* < 0.001). RCI.comp and PANAS– do not correlate (|*r*| < 0.1; *p* > 0.1). In addition to revealing the mentioned pattern, this also indicates that participants enjoying competition tended *not* to be social-interaction anxious whereas participants with higher social-interaction anxiety tended to report higher negative affect. These are meaningful outcomes.

--- Insert Table 8 around here ---

Table 8 Correlations of affective variables with cortisol *3–1*, *4–3* and *(3–1)–(4–3)* variables.

Variable	3-1	4-3	(3-1)-(4-3)	Treatment
SIAS	0.31***(109)	-0.29**(113)	0.34***(109)	EU*
RCI.comp	-0.18.(93)	0.19.(97)	-0.19.(93)	EU*
RCI.cont	-0.28**(93)	0.33***(97)	-0.34***(93)	EU*
Like	0.04 (109)	-0.04 (113)	0.05 (109)	EU*
Flow	-0.11 (108)	0.03 (112)	-0.09 (108)	EU*
PANAS+	0.05 (108)	-0.11 (112)	0.09 (108)	EU*
PANAS-	0.30**(108)	-0.24*(112)	0.30**(108)	EU*
Test score	-0.05 (66)	-0.12 (69)	0.02 (66)	EU*
Flow	0.10 (16)	0.16 (15)	0.03 (15)	Beer
PANAS+	-0.15 (16)	0.40 (15)	-0.31 (15)	Beer
PANAS-	0.02 (16)	0.20 (15)	-0.18 (15)	Beer
Flow	0.00 (31)	0.03 (34)	0.01 (30)	Aid*
PANAS+	0.02 (29)	-0.20 (33)	0.14 (29)	Aid*
PANAS-	0.12 (29)	-0.08 (33)	0.15 (29)	Aid*

Note: Numbers in brackets denote the number of participants with valid data used for calculations (especially some cortisol samples contained so low saliva volume that they had to be dismissed from the analysis and one class did not filled in one knowledge test, which means Test score could not be calculated).

·p < .1 *p < .05 **p < .01 ***p < .001

As concerns between-gender differences, we computed correlations for each gender separately, but only for EU* treatments because other treatments had a small sample size. There were noticeable between-gender differences concerning cortisol correlations with SIAS, PANAS– and, to a lesser extent, with Flow and RCI.comp. These differences are reported in Table 9. Correlations were notable only for males. Thus, the pattern mentioned in the previous paragraph seems to be caused by males rather than females, as also illustrated on Fig. 9, 10. To conclude: The higher the negative affective state of males, the lower the Flow. At the same time, the higher the negative affective state of males, the higher the relative cortisol level in the main condition (sample nr. 3) and lower the relative cortisol level in the post-exposure condition (sample nr. 4). Note, however, that this part of the study is exploratory and the results should be interpreted with caution. Especially, we do not put much stock into the findings related to RCI.comp, because all the respective correlations are weak and may be easily caused by chance.

We also explored correlations separately for treatments EU-comp, EU-no-comp and EU-class, but due to the small sample sizes in EU-comp and EU-no-comp, results were not robust and correlations could easily change direction due to the presence of outliers. Therefore, we included this supplementary analysis into Appendix D.

--- Insert Table 9 around here ---

Table 9 Correlations of affective variables with the cortisol-related variables according to gender (across all EU* treatments).

	3-1		4-3		(3-1)-(4-3)	
Variable	males	females	males	females	males	females
SIAS	0.40**(62)	0.14 (47)	-0.34**(63)	-0.21 (49)	0.43***(62)	0.19 (47)
Flow	-0.24.(62)	0.09 (46)	0.23.(63)	-0.21 (48)	-0.28*(62)	0.17 (46)
PANAS-	0.52***(62)	0.11 (46)	-0.50***(63)	-0.06 (48)	0.59***(62)	0.09 (46)
RCI.comp	-0.23 (54)	0.08 (39)	0.12 (55)	0.13 (41)	-0.20 (54)	-0.00 (39)

Note: Only affective variables for which the between-gender differences were notable are depicted. Numbers in brackets denote the number of participants with valid data used for calculations (one participant did not fill in gender).

·*p* < .1 **p* < .05 ***p* < .01 ****p* < .001

--- Insert Figure 9 around here ---

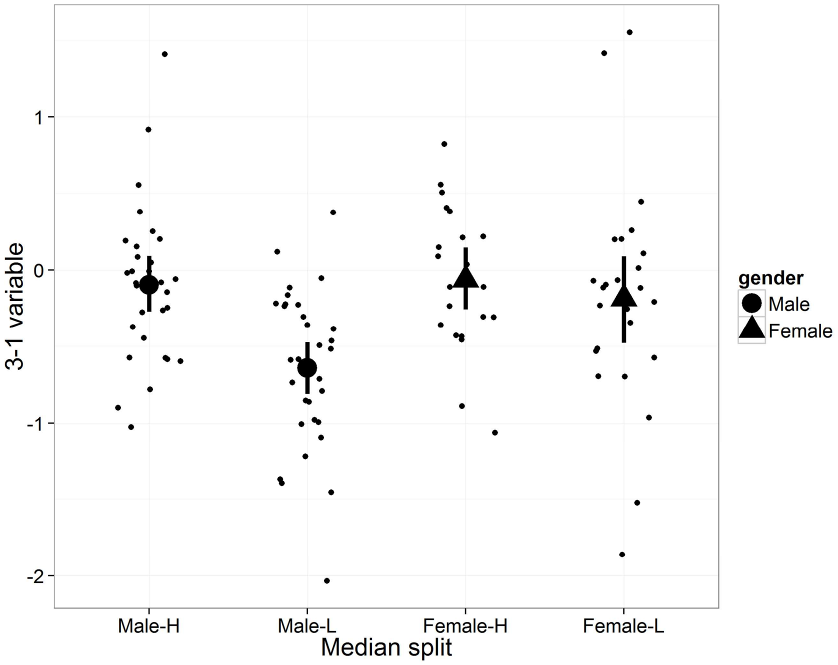


Fig. 9 Between-gender differences in *3–1* variable. “H” denotes the high PANAS– group while “L” denotes the low PANAS– group (according to median-split).

--- Insert Figure 10 around here ---

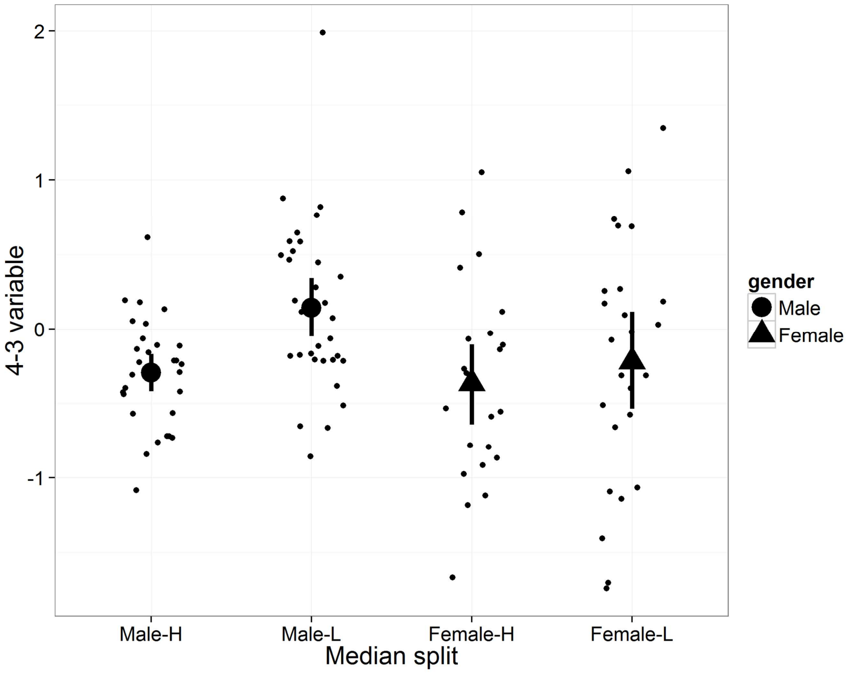


Fig. 10 Between-gender differences in *4–3* variable. “H” denotes the high PANAS– group while “L” denotes the low PANAS– group (according to median-split).

5. Discussion

The primary purpose of this study was to explore the link between learners’ positive–negative affective states, their flow states, salivary cortisol levels and immediate learning gains. Its secondary purpose was to bring the technique of salivary cortisol sampling to the field of digital game-based learning (DGBL). Despite the innate appeal of the idea that higher engagement leads to better knowledge acquisition, this link had only rarely been explored directly in the past; especially in connection with salivary cortisol levels.

Our primary purpose also included investigation of the tension between three different hypotheses predicting different flow-cortisol relationships. This aspect of the study is also new. To this end, we intentionally used four treatments that tended to feature inherent social-evaluative threats and some aspects of uncontrollability, which are known to lead to higher cortisol responses (Dickerson & Kemeny, 2004), and two treatments in which these features were absent or reduced. At the same time, we predicted that these treatments would elicit different levels of flow, as shown on Fig. 2.

5.1 Goal 1: Do the Positive–Negative Affective and Flow States Relate to Each Other and to Learning Gains?

We demonstrated that our treatments generated different levels of flow and positive/negative affect among learners; a fact that can be considered a successful manipulation check (Tab. 3-6). Generally, the following relationship among the groups was observed:

Flow: Aid-actors ≥ {Beer, Aid-partic} ≥ EU-comp ≥ {EU-no-comp, EU-class}

PANAS+: {Beer, Aid* } ≥ {EU-comp, EU-no-comp} ≥ EU-class

PANAS–: {Aid-partic, EU*} ≥ {Aid-actors, Beer}[16]

The ordering is consistent across all the three measures, with only one exception. Generally, both Aid* groups and the Beer treatment yielded higher flow and positive affect and lower negative affect, compared to EU* treatments. Among the EU* treatments, the EU-class treatment, which was based on a “traditional” discussion-based model, scored the worst. The EU-no-comp treatment scored slightly worse than the EU-comp treatment, possibly due to removed economic layer of the game. The exception to this pattern is that the highest negative affect was found in the Aid-partic group. However, this is not surprising, given the “bloody” situation participants in the car accident simulation faced (see Sec. 2.3). In fact, we see an interesting dissociation of the two Aid* groups: both have the same PANAS+ but different PANAS– (Tab. 3).

The relationship among these three variables was also confirmed in a correlation analysis (Tab. 7); it is especially noteworthy that participants in flow tended to report higher PANAS+ and lower PANAS– than those who were not in flow. It thus seems that flow can be instigated when participants experience high positive affect and low negative affect. This pattern is very similar to the pattern that we observed on a larger sample undergoing two variations of the Beer treatment in a different research project (Brom et al., 2014; see also Footnote (10)) and also to results reported in the field of positive psychology (e.g., Smolej-Fritz and Avsec, 2007; Rogatko, 2009). Given the set of feelings PANAS+ is composed of (such as active, attentive, determined), the relationship between PANAS+ and flow is not that surprising. Nevertheless, this relationship helps to justify the tentative concept of engaged concentration (introduced in Sec. 1; cf. Baker et al., 2010). While generalized positive affect and flow are clearly different constructs, they probably share a common denominator. It would be useful to pin this denominator down in future because it is probably highly relevant for learning. Appendix D shows that this result generalizes across two different media, a computer-based and non-computer-based (Tab. 11 - 13). It thus seems that this finding has some robustness. Moreover, Flow and PANAS+ are highly correlated to the Like variable. Negligible correlation between PANAS+ and PANAS– is not surprising, given that these scales are supposed to be orthogonal (Watson et al., 1988). The minor caveat to this view is the lack of a relationship between Flow and PANAS+/– variables for the Aid-actors group (Tab. 7). However, this is most likely caused by the Aid-Actors group’s small size (n = 12) and approaching the maximum value for the Flow variable in this group, which is 74 (after the FSS T-norm transformation). Removing a single outlier with the lowest Flow value (39) increases the Flow x PANAS+ correlation to r = 0.34 while the Flow x PANAS– decreases r to –0.25 in this group.

[16] Note that a higher number for PANAS– means a higher *negative* affect.

The existence of a direct link between immediate learning outcome, as measured by the Test score variable, and a positive affective state/flow was supported, but the relationship is only modest (Tab. 7). Appendix D again shows that this result tends to hold when the instruction is delivered either via a computer-based or a non-computer based intervention (Tab 11 – 13). Notably, a similarly modest relationship was found in our second experiment with the Beer treatment (Brom et al., 2014). In the field of multimedia learning, small to modest positive correlations between positive affect, as measured by PANAS+, and learning outcomes were found by Plass et al. (2014) and Um et al. (2012) and almost no relationship was found between positive affective variables, different from PANAS+, and learning outcomes by Plass et al. (2013). A modest correlation between learning gains and flow, as judged by independent observers, was reported in (Craig et al., 2004), and strong correlations between flow, measured by an adapted FSS, and learning outcomes were found by van der Meij (2013). In the flow research, Vollmeyer and Rheinberg (2006) reported modest relationship between motivation/flow and learning achievements. In the DGBL field, Ritterfeld et al. (2010) reported small, mostly non-significant positive correlations between gained interest and knowledge gain variables; van Dijk (2010) reported mainly strong correlations between learners' motivation and their learning outcomes in the guided discovery condition featuring a game, but no correlation in the worked example (i.e., control) condition featuring PowerPoint slides; and Giannakos (2013), using median-split technique, reported that learners' enjoyment was moderately related to their performance.[17] At the same time, we observed no relationship between PANAS– and immediate learning outcome in the present study. In general, it seems that the idea that positive affect and high flow are related to higher knowledge gain has some support in the data, but the alleged link may be weaker than some DGBL proponents may intuitively assume. More research is needed to elucidate this topic. In particular, it would be useful to investigate how treatments' characteristics, such as various kinds of extraneous but engaging details, influence this relationship (cf. Mayer, 2009; Moreno, 2005; Um et al., 2012; Plass et al., 2014) and how specific emotions and their dynamic changes during learning contribute to learning gains (cf. D'Mello et al., 2012; D'Mello et al., 2013; Craig et al., 2004).

Finally, there were no notable gender differences in affective variables, indicating that females and males can be equally engaged by game-based interventions, if they are specifically tailored to educational purposes.

5.2 Goal 2: Is Salivary Cortisol a Physiological Correlate of Flow?

One of our goals was to investigate the tension between the three hypotheses predicting different relationships between flow state and cortisol levels: the Inverted-U, Perceived-fit and Treatment-specificity-and-personal-characteristics (TSPC) hypotheses.

We found only a small, marginally significant negative relationship between cortisol levels and flow, and only for males (Tab. 9): the higher the flow, the higher the cortisol *decrease*. This is exactly the opposite of what Keller et al. (2011) found in their Exp. 2 on a male-only sample: participants playing a single-player game, Tetris, under a condition supposed to yield a higher flow, compared to two other low-flow conditions, had higher cortisol levels (adjusted for baseline). Thus, our data failed to support the Perceived-fit hypothesis put forward by Keller et al.

Our data does not support Peifer's (2012) Inverted-U hypothesis either. Figure 11 illustrates the postulated relationship and our findings. Except for the EU-no-comp group, the outcome actually

[17] The audience of these studies ranged from 10 years old kids to college students.

accords with the prediction depicted on Fig. 2. Figures 12 and 13 give a full scatter plot for males and females, respectively.

--- Insert Fig. 11 about here ---

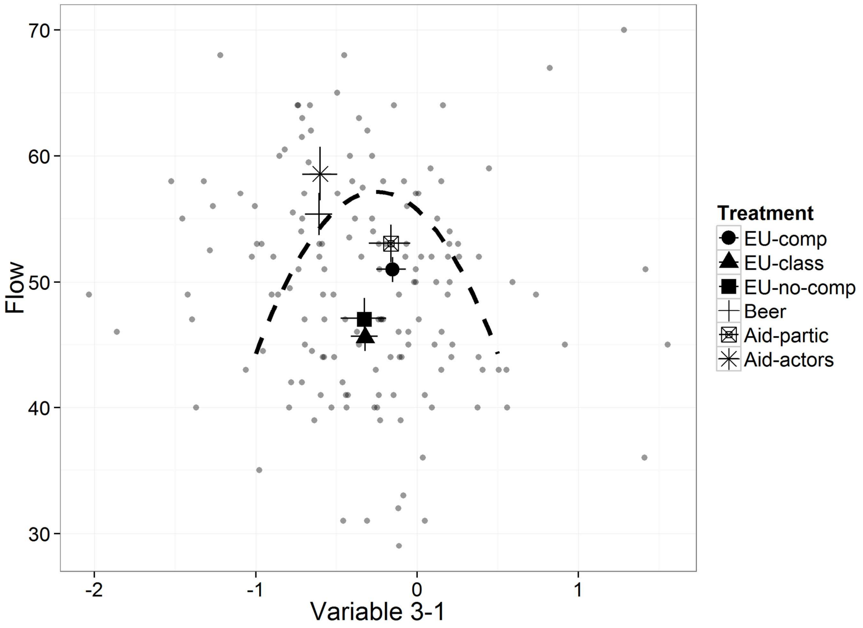


Fig. 11 Means for the Flow x *3–1* variable for the six groups. Lines denote standard errors of the mean. The hypothetical “inverted-U” relationship is also depicted.

--- Insert Fig. 12 about here ---

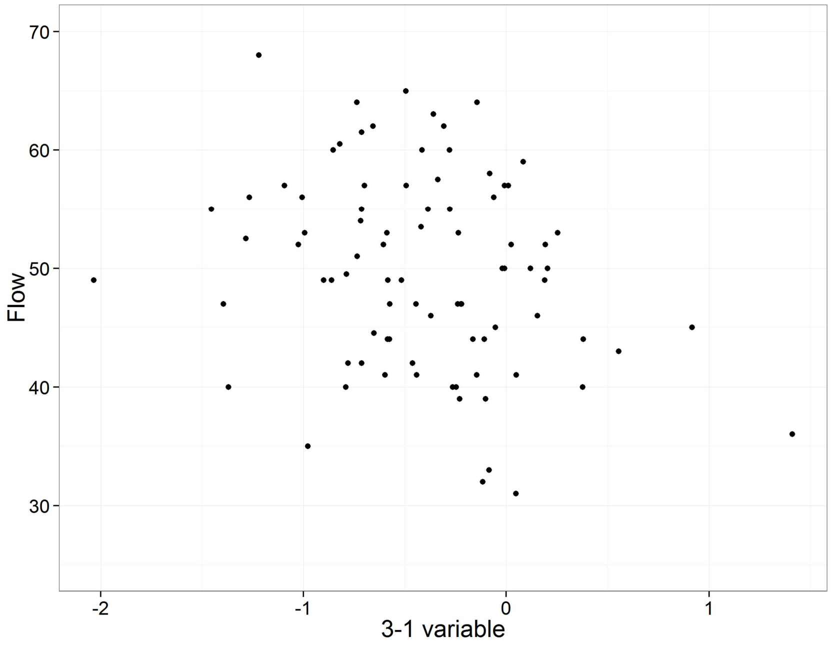


Fig. 12 Scatter plot for the Flow x *3–1* variable for males.

--- Insert Fig. 13 about here ---

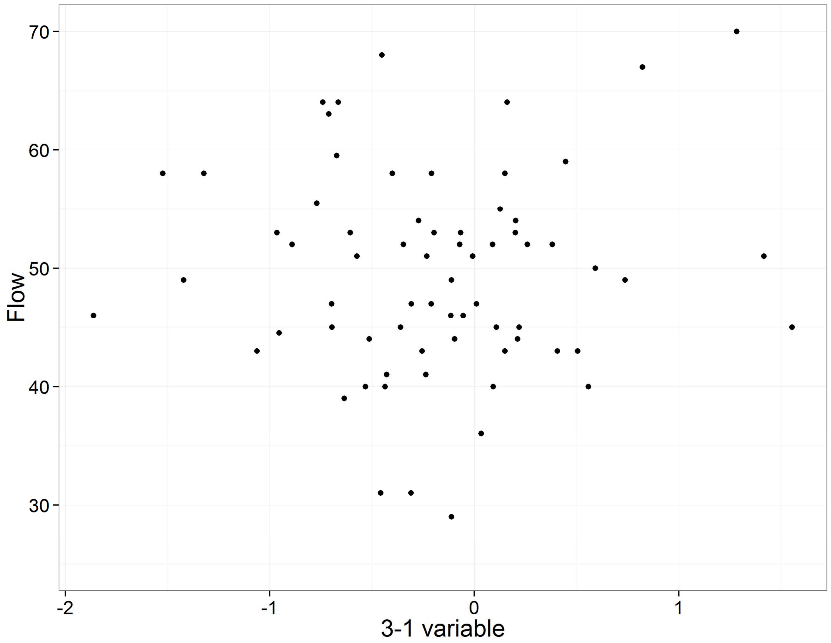


Fig. 13 Scatter plot for the Flow x *3–1* variable for females.

The results above, and the fact that they do not fit well with the previous findings, actually do support the TSPC hypothesis: each treatment will generate its own specific cortisol–flow relationship, depending on the treatment's and participants' characteristics. Notably, this hypothesis is further supported by the following findings of the present study:

a) moderate correlation between the SIAS and *3–1/4–3* variables for males rather than females (Tab. 9); the higher the social-interaction anxiety, the higher the cortisol value for the main condition, compared to the pre-exposure and post-exposure condition (across EU* treatments);

b) small to moderate correlations between the RCI.cont and *3–1/4–3* variables no matter the gender (Tab. 8); the higher the contentiousness (i.e., one subscale of the Revised Competitiveness Index), the lower the cortisol value for the main condition, compared to the pre-exposure and post-exposure condition (across EU* treatments);

c) high correlation between the PANAS– and *3–1/4–3* variables for males (Tab. 9); the higher the negative affect during the treatment interaction, the higher the cortisol value for the main condition, compared to the pre-exposure and post-exposure condition (across all treatments);

d) differences in *3–1* variable between the Beer treatment, which does not feature a social evaluative-threat and is highly controllable, and the EU*/Aid-partic treatments, which feature social evaluative-threats and are less controllable; note however that these differences are at the edge of significance (Sec. 4.2).

While each individual finding above may not be compelling enough, all of them together argue in favor of the TSPC hypothesis. Cortisol values are influenced by gender, in combination with some of the learners' personal characteristics (in our case: social-interaction anxiety and contentiousness) and treatment characteristics (social-evaluative threat, controllability), rather than by the learners' flow levels.

Several points deserve commenting. First, our study has a higher ecological validity than Peifer (2012), Peifer et al. (2014) and Keller et al. (2011); however, this comes at the cost of higher noise (see also Sec. 5.5). It is thus possible that under more controlled conditions, more nuanced patterns may emerge. It is also possible that for treatments sharing some specific characteristics, or for a specific group of people (recall that we intentionally used heterogeneous sample), one of the remaining two hypotheses could be still valid. Our findings should thus be interpreted only as a warning that predictions of these hypotheses should not be over-generalized. In addition, Peifer's research methods differ from ours in that her treatments could be more stressful than ours and/or generate higher cortisol values. Unfortunately, our raw cortisol values are not directly comparable to hers because we conducted the experiment in the morning rather than in the afternoon. Finally, Keller et al. (2011) used a different measure of flow, which also complicates direct comparison of their findings to ours; though the primary message would unlikely change were the same flow questionnaires used. Interestingly, Peifer used the FSS as we did, but in one of her experiments (Peifer et al., 2014), she found the inverted-U relationship only between the raw post-exposure cortisol values and one of the FSS' subscales, "absorption", which consists of 4 out of 10 questions. We run a supplementary analysis and found the inverted-U relation in our data neither between the absorption subscale and the *3–1* variable nor between the absorption subscale and the raw values of sample no. 3, our main condition sample.

Second, cortisol sample no. 3 was taken earlier in the Aid-partic group than in the Beer group. Because cortisol levels naturally decrease during morning hours, the lower *3–1* values for the Beer treatment compared to the Aid-partic group could have been caused by the later timing of the sample collection. However, the Beer treatment participants woke up later than the Aid-partic participants, and the difference still remains marginally significant when the earlier Beer sample (no. 2) is considered; i.e., when the Beer *2–1* variable is compared to the Aid-partic *3–1* variable ($t(32.04) = 1.92$; $p = .063$; $d = 0.57$; t-test with Welch approximation).

Finally, our findings indicate that special attention should be paid to team-based learning activities, including educational games, with low controllability and where the learners' performance could be negatively judged by others: especially in case of male learners. Moreover, team-based learning activities with competitive aspects, such as *Europe 2045*, may lead to higher physiological stress for learners who do not like competition. These learners may also tend be, in general, more social-interactive anxious (Sec. 4.3).

5.3 Goal 3: How Does a Learner's Positive–Negative Affective State Relate to Cortisol Levels?

We have already mentioned a high positive correlation between the PANAS– and *3–1/4–3* variables for males (Tab. 9). This is in line with our prediction from Sec. 1.2, but past results were less unequivocal than ours (Abercrombie et al., 2005; McBurnett et al., 2005; see also our Appendix D, Tab. 14) and some were the opposite (Het et al., 2012). However, not all experimental protocols are directly comparable. For instance, we used the instruction "mark to what extent did you experience the following feelings *during the last discussion*." whereas Het et al. (2012) used instruction "*now*," i.e., after participants finished the Trier Social Stress Test.[18] It is therefore possible that we measured affect *during* the event while Het et al. (2012) *after* the event. Het et al. (2012) indeed pointed out that cortisol may positively correlate with PANAS– during a stressful event, such as TSST, but negatively after it (pp. 29-30).

At the same time, there does not seem to be a general relationship between the PANAS+ and *3–1/4–3* variables (Tab. 8). This agrees with the results of McBurnett et al. (2005). In fact, as concerns the link between PANAS+ and cortisol, our results agree well with the "no-relationship" prediction put forward on Fig. 1. The situation is similar to the case of flow–cortisol link, described in Sec. 5.2.

To conclude, more research is needed to clarify the relationship between both the PANAS' dimensions and cortisol levels, considering different measurement times and the moderating effects of different treatment types and especially of ST treatments.

5.4 Goal 4: Are Cortisol Levels Related to Learning Outcomes and Testing/Debriefing Conditions?

Although associations between individual learning outcomes and cortisol levels have been previously described (e.g., Kuhlmann & Wolf, 2006; Flegr & Priplatova, 2010), our data investigating the EU*

[18] Email correspondence from 12th June 2014.

groups failed to demonstrate such a relationship. It is not yet fully understood what aspects of serious games contribute most to learning (Tobias et al., 2011, p. 194), which makes it difficult to put our results into context with existing data. More research into the differences between standard teaching methods and the DGBL is definitely needed to study cortisol-modulating effects on learning outcomes.

The post-exposure cortisol sampling was related to filling in of tests in the EU* and Beer groups, while it was related to participating in a debriefing in the Aid* groups. The cortisol values for the post-exposure condition adjusted to the main condition were higher in the Beer group compared to the EU-comp group (Sec. 4.2, Fig. 7); indicating that the period of filling in of tests could indeed increase the physiological stress in the former group. Also in Aid* treatments, cortisol levels remained approximately steady in the post-exposure condition; i.e., relatively high during and after the debriefing. However, the results were mixed when the three EU* treatments were considered, with a significant difference between the EU-comp and EU-class groups (Sec. 4.2, Fig. 7). The possible effect of filling in of questionnaires on cortisol levels, measured by the *4–3* variable, was surely confounded by relatively high cortisol values for the main condition in the EU* treatments compared to the Beer treatment and possibly somewhat lower cortisol values for the main condition in the EU-class treatment compared to the other two EU* treatments. Thus, our data partially support the findings of Minkley & Kirchner (2012), stating that filling in of tests can increase salivary cortisol levels compared to a baseline. However, a standardized research design that evaluates the pre-testing phase with regard to actual testing would be needed to draw any broader and more robust conclusion.

Notably, in the EU* treatments, cortisol levels in males with the low negative affect (PANAS–) during the intervention increased the most in the post-exposure condition (Fig. 9, 10). One of the possible interpretations of this finding is that the data were not confounded by relatively high values of the previous sample in this subgroup of participants.

5.5 Goal 5: Is Salivary Cortisol Sampling a Useful Research Tool in the DGBL Field?

To the best of our knowledge, this study is the first that has used salivary cortisol sampling on a larger scale in the context of DGBL research. Including a pilot run, we analyzed around 900 cortisol samples (~750 samples in the main study). The methodological question is whether we consider this technique usable in other DGBL studies or not, given the costs associated with sample analysis. If a DGBL study were conducted in a carefully controlled laboratory environment, it would have the same pros and cons as any other study using cortisol sampling. Thus, we will now focus on DGBL studies with a higher ecological validity; conducted in a field setting or modeling a school day in a laboratory. We now list the main limitations of cortisol sampling in this context:

1. The experiment usually has to be conducted during the morning; however, cortisol levels are high in the morning and decrease gradually during the day. This pattern tends to disguise possible cortisol increases caused by exposure to a stressor (cf. Dickerson & Kemeny, 2004). This issue materialized in our study, as we saw no robust increases in cortisol levels in the main conditions. Instead, we mainly found marginally smaller and larger decreases. However, the magnitude of these decreases can still be compared between groups, providing significant group differences.

2. Cortisol cannot be measured too often, as that would disrupt the course of activities. Salivary cortisol sampling is generally considered as non-invasive; however, some of our high school

participants actually found chewing a cotton roll disgusting. Therefore, it is hard to imagine that they would have agreed to repeated measurements (for example eight or more samples during the experiment) without being negatively influenced by it. Depending on the magnitude of a stressor, cortisol varies quickly over time with a peak at around 20-40 minutes after the onset of a stressor (Dickerson & Kemeny, 2004). Thus, the actual sampling procedure, which only took around 60 seconds, should not have affected our experimental data; especially with the between-sampling differences longer than 40 minutes for the EU* treatments, which employed high-school participants (see Fig. 5). Similarly, because of inter-individual differences in cortisol levels, it would be useful to obtain cortisol samples from different days, which would enable generating participant-specific baselines. In reality, it would not be easy to collect these extra samples from high-school participants.

3. The nutritional state of participants at the beginning of the experiment is hard to standardize. With high school participants, forbidding products with caffeine or glucose (even for a short period) and sometimes nicotine may negatively impact their attitude towards the experiment. For instance, several our participants were unhappy when coffee was temporarily forbidden. In a similar vein, it is hard to obtain some information (from high school participants) needed for a proper analysis; such as the timing of women’s menstrual cycles or if they use contraceptives.

4. Even if the research team follows a schedule as precisely as possible, some discrepancies are unavoidable. Given the diurnal rhythm of cortisol release in healthy subjects, it is particularly troubling when samples to be compared are collected at different times during the day. Thus, study protocols with multiple sampling points have to consider the circadian pattern of salivary cortisol secretion. Especially in the morning, variations in sampling should be avoided.

Consequently, a higher noise in the cortisol data can be expected in a relatively ecologically valid DGBL study, compared to a lab study conducted in the afternoon. This necessitates a larger sample. Note, for instance, that even with a sample of our size and with random assignment, the means of the pre-exposure cortisol levels were somewhat different between our groups (cf. EU-class and EU-no-comp; Fig. 7).

However, even though additional patterns may emerge in the data if the study is conducted in a more controlled manner, our results bring several significant findings, have internal consistency and, to a large extent, agree with what is already known about the impact of various interventions on cortisol levels in males and females.

Considering all these points together, in our opinion, the cortisol assessment can be used in the context of DGBL research. However, if the study is conducted in the field, a large samples size would be desirable and cortisol data should be ideally supported by other biomarkers that are known to vary with psychological arousal; for example, IgA and/or Alpha Amylase (Tsujita & Marimoto, 1999; Rohleder et al., 2006; Kang, 2010). Measurements of the autonomic nervous system can also add valuable insights into psycho-physiological reactions to stress (Sharpley et al., 2000; Donzella et al., 2000; see also Anolli et al., 2010)

5.6 General Limitations

First, as already discussed in Sec. 5.2., our study is not directly comparable to the studies of Peifer (2012), Peifer et al. (2014) and Keller et al. (2011). However, this argument does not discredit the significance of our main finding on the cortisol–flow relationship. Second, the EU-class groups had different dynamics than the EU-comp/EU-no-comp groups, because the EU-class intervention (Sec. 2.1.2) featured two longer discussion sessions rather than four shorter sessions. However, we argue that the setting with two discussions (for the EU-class treatment) is more ecologically valid than the setting with four discussions (for this particular treatment): our pilots showed that the latter arrangement did not work well and that teachers would unlikely choose it in a real school environment. Third, EU-class groups could have had more than eight participants, which means that one or two couples would have been assigned the same project and thus the whole class could have learned more about this (these) particular project(s). (This does not hold for policy proposals, because two students having the same project could be assigned different proposals.) This limitation has no practical impact on this study: (a) this issue arose only twice (see Appendix C); (b) between-group comparison as concerns learning gains is out of our present scope. Fourth, our sample would have ideally been larger to enable us to also investigate cortisol correlations within different treatment groups (cf. Appendix D). Still, the sample was large enough to draw the main conclusions concerning the relationship between the cortisol levels, flow and both the PANAS dimensions, as well as to elucidate gender differences. Fifth, it would be useful to have more data on Aid* participants and Beer participants (such as their RCI.cont score), but we were not able to administer more questionnaires in these groups due to time constraints. Finally, future research should also consider pinning down the concept of engaged concentration and also focusing on other specific emotions, because it is possible that more nuanced patterns would emerge as concerns the relationship between learning gains and specific affective states, such as confusion (cf. D'Mello et al., 2013; Craig et al., 2004).

6. Conclusions

This study investigated the relationship between the positive–negative affective state of learners interacting with a serious game, their flow, immediate learning gains and physiological arousal, measured by levels of the hormone cortisol. It worked with the tentative concept of engaged concentration (also called state engagement), provisionally linked to flow and positive affect. It was mainly exploratory, because it was one of the first studies investigating such a relationship directly. To our knowledge, it was the first one probing application of salivary cortisol sampling on a larger scale in the context of DGBL research.

The study provided several key results. First, it indicated that there is a high positive relationship between flow and positive affect and a moderate to high negative relationship between flow and negative affect. Second, its findings failed to support two hypotheses concerning the flow–cortisol relationship, but they supported the idea that each treatment would generate its own specific cortisol–flow relationship, depending on the treatment's and learners' characteristics (notably gender, social interaction anxiety, attitude towards competition and the presence of social-evaluative threat in the treatment and the degree to which it is controllable by the learner). Third, it indicated that team-based learning interventions with competitive elements and a higher social-evaluative threat/uncontrollability elicit higher physiological stress in male learners with a higher social-interaction anxiety and in learners, no matter the gender, with lower contentiousness score. Fourth, it directly demonstrated a link between affective state/flow and immediate learning outcomes, but this link was only modest, which seems to agree with related findings reported in the literature but which may come as a surprise to some DGBL proponents. In addition, it failed to demonstrate a relationship between immediate learning gains and changes in cortisol levels.

In general, this study provides new results for DGBL researchers and for those interested in flow and the link between learning and affect. Because of the relationship between positive affect and flow, it makes a step towards justification of the concept of engaged concentration. It also demonstrates that cortisol measurements are promising tool in the DGBL field, but it is not without pitfalls and should be always supplemented with other research methods.

References

Abercrombie, H. C., Kalin, N. H., & Davidson, R. J. (2005). Acute cortisol elevations cause heightened arousal ratings of objectively non-arousing stimuli. *Emotion,* 5(3), 354—359.

Abercrombie, H. C., Speck, N. S., & Monticelli, R. M. (2006). Endogenous cortisol elevations are related to memory facilitation only in individuals who are emotionally aroused. *Psychoneuroendocrinology,* 31, 187–196.

Anderson, M. C. (2009) Motivated forgetting. In Baddeley A., Eysenck. M. W., Anderson M. C. (Eds.) *Memory* (pp. 217–244). Psychology Press.

Anolli, L., Mantovani, F., Confalonieri, L., Ascolese, A., & Peveri, L. (2010). Emotions in Serious Games: From Experience to Assessment. *International Journal of Emerging Technologies in Learning*. 5(3), 7–15.

Badrick, E., Kirschbaum, C., & Kumari, M. (2007). The relationship between smoking status and cortisol secretion. *Journal of Clinical Endocrinology and Metabolism*, 92, 819–824.

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223-241.

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

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

Brom, C., Šisler, V., Buchtová, M., Klement, D., & Levčík, D. (2012). Turning High-Schools into Laboratories? Lessons Learnt from Studies of Instructional Effectiveness of Digital Games in the Curricular Schooling System. In *E-Learning and Games for Training, Education, Health and Sports 7th International Conference, Edutainment 2012 and 3rd International Conference, GameDays 2012, LNCS Vol. 7516*, Springer, 41–53.

Chrousos, G. P. (2009). Stress and disorders of the stress system. *Nature Reviews Endocrinology*, 5(7), 374-81.

Clark, R., (Ed.) (2012). *Learning from Media*. 2nd ed. Information Age Publishing.

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

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29, 241–250.

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety*. San Francisco: Jossey-Bass Publishers.

De Grove, F., Bourgonjon, J., & Van Looy, J. (2012). Digital games in the classroom? A contextual approach to teachers’ adoption intention of digital games in formal education. *Computers in Human Behavior*, 28, 2023-2033.

Dickerson, S. S., & Kemeny, M. E. (2004). Acute Stressors and Cortisol Responses: a Theoretical Integration and Synthesis of Laboratory Research. *Psychological Bulletin*, 130, 355–391.

D’Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145–157.

D’Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2013). Confusion can be beneficial for learning. *Learning and Instruction*, 29, 153–170.

Donzella, B., Gunnar, M.R., Krueger, W.K., & Alwin, J. (2000). Cortisol and vagal tone responses to competitive challenge in preschoolers: associations with temperament. *Developmental Psychobiology*, 37, 209–20

Dunnett, C. W. (1980). Pairwise Multiple Comparisons in the Homogeneous Variance, Unequal Sample Size Case. *Journal of the American Statistical Association*, 75(372), 789-795.

Egenfeldt-Nielsen, S. (2005) *Beyond Edutainment: Exploring the Educational Potential of Computer Games*, PhD thesis, University of Copenhagen.

Elliot, A. J., & Pekrun, R. (2007). Emotion in the hierarchical model of approach-avoidance achievement motivation. In P. A. Schutz (Ed.), *Emotion in education* (pp. 57-73): Elsevier Academic Press.

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

Eysenck, M., & Keane, M. T. (2010). *Cognitive Psychology: A Student's Handbook*, 6th Ed. Psychology Press.

Flegr, J., & Priplatova, L. (2010) Testosterone and cortisol levels in university students reflect actual rather than estimated number of wrong answers on written exam. *Neuroendocrinology Letters*, 31(4), 577–581.

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.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, Motivation, and Learning: A Research and Practice Model. *Simulation & Gaming*, 33, 441 - 467.

Giannakos, M. N. (2013) Enjoy and learn with educational games: Examining factors affecting learning performance. *Computers & Education*, 68, 429 - 439.

Girard, C., Ecalle, J., & Magnan, A. (2012). Serious games as new educational tools: how effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning*, 29(3), 207–219.

Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: exploring the value of intrinsic integration in educational games. *Journal of the Learning Sciences*, 20 (2), 169-206.

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

Hays, R. T. (2005) *The Effectiveness of Instructional Games: A Literature Review and Discussion,* Technical Report 2005-004, Orlando: Naval Air Warfare Center Training Systems Division.

Hébert, S., Béland, R., Dionne-Fournelle, O., Crete, M., Lupien, S. (2005) Physiological stress response to video-game playing: the contribution of built-in music. *Life Sciences*, 76, 2371–2380.

Hellhammer, D. H., Wust, S., & Kudielka, B. M. (2009). Salivary cortisol as a biomarker in stress research. *Psychoneuroendocrinology*, 34, 163–171.

Het, S., Schoofs, D., Rohleder, N., & Wolf, O. T. (2012). Stress-induced cortisol level elevations are associated with reduced negative affect after stress: indications for a mood-buffering cortisol effect. *Psychosomatic medicine*, 74(1), 23-32.

Hossini, F., Rezaeeshrazi, R., Salehian, M. H., & Dana, A. (2011). The Effect of Violent and Non-Violent Computer Games on Changes in Salivary Cortisol Concentration in Male Adolescents. *Annals of Biological Research*, 2(6), 175-178.

Houston, J.M., Harris P. B., Moore, R., Brummett, R. A., & Kametani, H. (2005). Competitiveness in Japanese, Chinese, and American undergraduates. *Psychological Reports*, 97, 205–212.

Huizenga, J., Admiraal, W., Ten Dam, G. (2013). Teaching with games in secondary education in the Netherlands. Presented at EARLI 2013. http://www.earli2013.org/programme/proposal-view/?abstractid=3426. Accessed 25th Dec 2013.

Hussain, M. S., AlZoubi, O., Calvo, R. A., & D'Mello, S. K. (2011) Affect Detection from Multichannel Physiology during Learning Sessions with AutoTutor. *Lecture Notes in Computer Science: Vol. 6738. Artificial Intelligence in Education* (pp. 131-138): Springer.

Inder, W. J., Dimeski, G., & Russell, A. (2012). Measurement of salivary cortisol in 2012 – laboratory techniques and clinical indications. *Clinical Endocrinology*, 77, 645–651.

Ivarsson, M., Anderson, M., Akerstedt, T., & Lindblad, F. (2009). Playing a violent television game does not affect saliva cortisol. *Acta Paediatrica*, 98(6), 1052-3.

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin*, 96(3), 589–596.

Kajantie, E., & Phillips, D. I. (2006). The effects of sex and hormonal status on the physiological response to acute psychosocial stress. *Psychoneuroendocrinology*, 31, 151–178.

Kang, Y. (2010). Psychological Stress-Induced Changes in Salivary α-amylase and Adrenergic Activity. *Nursing & Health Sciences*, 12, 477-484

Keller, J., Bless, H., Blomann, F., & Kleinböhl, D. (2011). Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology*, 47, 849 – 852.

Kirschbaum, C., Pirke, K., & Hellhammer, D. H. (1993). The “Trier Social Stress Test” — a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28, 76–81.

Kirschbaum, C., Wust, S., Faig, H.G., & Hellhammer, D.H. (1992). Heritability of cortisol responses to human corticotropin-releasing hormone, ergometry, and psychological stress in humans. *Journal of Clinical Endocrinology and Metabolism* 75 (6), 1526—1530.

Klopfer, E. (2008) *Augmented Learning*. The MIT Press.

Kudielka, B. M., Buske-Kirschbaum, A., Hellhammer, D. H., & Kirschbaum, C. (2004). HPA axis responses to laboratory psychosocial stress in healthy elderly adults, younger adults, and children: impact of age and gender. *Psychoneuroendocrinology*, 29, 83—98.

Kudielka, B. M., Hellhammer, D. H., & Würst S. (2009). Why do we respond so differently? Reviewing determinants of human salivary cortisol responses to challenge. *Psychoendocrynology*, 34, 2 - 18.

Kudielka, B. M., & Kirschbaum, C. (2005). Sex differences in HPA axis responses to stress: a review. *Biological Psychology*, 69(1), 113–32.

Kuhlmann, S., & Wolf, O. T. (2006). A non-arousing test situation abolishes the impairing effects of cortisol on delayed memory retrieval in healthy women. *Neuroscience Letters*, 399(3), 268–272.

Kuhlmann, S., & Wolf, O. T. (2005). Cortisol and memory retrieval in women: influence of menstrual cycle and oral contraceptives. *Psychopharmacology*, 183, 65–71.

Kupper, N., & Denollet, J. (2012). Social anxiety in the general population: introducing abbreviated versions of SIAS and SPS. *Journal of affective disorders*, 136(1), 90-98.

Lester, J. C., Ha, E. Y., Lee, S. Y., Mott, B. W., Rowe, J. P., & Sabourin, J. L. (2013). Serious Games Get Smart: Intelligent Game-Based Learning Environments. *AI Magazine*, *34*(4), 31-45.

Levene, H. (1960). Robust tests for equality of variances. In Olkin, I., Ghurye, S. G., Hoeffding, W., Madow, W. G., & Mann, H. B. (Eds.) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278–292). CA: Stanford University Press.

Løvoll, H. S., Vittersø, J. (2014) Can Balance be Boring? A Critique of the “Challenges Should Match Skills” Hypotheses in Flow Theory. *Social Indicators Research*, 115 (1), 117–136.

Malone, T. W. (1981) Toward a theory of intrinsically motivating instruction, *Cognitive Science*, 5(4), 333-369.

Malone, T. & Lepper (1987). Making Learning Fun: A Taxonomy of Intrinsic Motivations for Learning. In Snow, R. & Farr, M. J. (Ed) *Aptitude, Learning, and Instruction*, Volume 3: Conative and Affective Process Analyses. Hillsdale, NJ

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

Mayer, R. E., & Johnson, C. I. (2010). Adding Instructional Features that Promote Learning in a Game-like Environment. *Educational Computing Research*, 42(3), 241-265.

McBurnett, K., Raine, A., Stouthamer-Loeber, M., Loeber, R., Kumar, A. M., Kumar, M., & Lahey B. B. (2005). Mood and hormone responses to psychological challenge in adolescent males with conduct problems. *Biological Psychiatry*, 57, 1109–1116.

van der Meij, H. (2013) Motivating agents in software tutorials. *Computers in Human Behavior*, 29(3), 845–857.

Miller, R. G. (1981). *Simultaneous statistical inference*. (2nd. ed.) Springer.

Minkley, N., & Kirchner, W. H. (2012). Influence of test tasks with different cognitive demands on salivary cortisol concentrations in school students. *International Journal of Psychophysiology*, 86, 245–250.

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.

Oxford, J., Ponzi, D., & Geary, D. C. (2010). Hormonal responses differ when playing violent video games against an ingroup and outgroup. *Evolution and Human Behavior*, 31, 201-209.

Palme, R., & Möstl, E. (1997). Measurement of Cortisol Metabolites in Feces of Sheep as a Parameter of Cortisol Concentration in Blood. *International Journal of Mammal Biology*, 62, 192–197.

Peifer, C. (2012). Psychophysiological Correlates of Flow-Experience. In S. Engeser (Ed.), *Advances in Flow Research* (pp. 139-165). New York: Springer.

Peifer, C., Schulz, A., Schächinger, H., Baumannd, N., Antonia C. H. (2014) The Relation of Flow-Experience and Physiological Arousal Under Stress – Can U Shape It? *Journal of Experimental Social Psychology*, in press, http://dx.doi.org/10.1016/j.jesp.2014.01.009

Pekrun, R. (2005). Progress and open problems in educational emotion research. *Learning and Instruction*, 15(5), 497–506.

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

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

Pruessner, J. C., Kirschbaum, C., Meinlschmid, G., & Hellhammer, D. H. (2003). Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change. *Psychoneuroendocrinology,* 28(7), 916-931.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41, 203–212.

Reisberg, D., & Hertel, P. (Eds.) (2003) *Memory and Emotion*. Oxford University Press.

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)

Ritterfeld, U., Shen, C., Wang, H., Nocera, L., & Wong, W. L. (2009). Multimodality and Interactivity: Connecting properties of serious games with educational outcomes. *CyberPsychology & Behavior*, 12(6), 691–697.

Robinson, M. D., Watkins, E. R., & Harmon-Jones E. (Eds.) (2013). *Handbook of Cognition and Emotion*. The Guilford Press.

Rogatko, T. P. (2009). The Influence of Flow on Positive Affect in College Students. *Journal of Happiness Studies*, 10, 133–148.

Rohleder, N., Wolf J.M., Maldonando, E.F., & Kirschbaum C. (2006). The Psychosocial Stress-Induced Increase in Salivary Alpha-Amylase is Independent of Saliva Flow Rate. *Psychophysiology*, 43, 645-652.

Ross, S. M., Morrison, G. R. (1989). In search of a happy medium in instructional technology research: Issue concerning external validity, media replications, and learner control. *Educational Technology Research and Development*, 37, 19 - 34.

Roozendaal, B. (2002). Stress and Memory: Opposing Effects of Glucocorticoids on Memory Consolidation and Memory Retrieval. *Neurobiology of Learning and Memory*, 78, 578–595.

Schommer, N. C., Hellhammer, D. H., & Kirschbaum, C. (2004). Dissociation between reactivity of the hypothalamus–pituitary–adrenal axis and the sympathetic–adrenal–medullary system to repeated psychological stress. *Psychosomatic Medicine*, 65, 450–460.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591-611.

Sharpley, C.S., Kamen, P., Galatsis, M., Heppel, R., Veivers, C., & Claus, K. (2000). An Examination of the Relationship between Resting Heart Rate Variability and Heart Rate Reactivity to a Mental Arithmetic Stressor. *Applied Psychophysiology and Biofeedback*, 25, 143–54

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology*, 64, 489–528.

Skosnik, P. D., Chatterton, Jr., R.T., Swisher, T., & Park, S. (2000). Modulation of attentional inhibition by norepinephrine and cortisol after psychological stress. *International Journal of Psychophysiology.*, 36 (1), 59–68.

Smolej-Fritz, B. & Avsec, A. (2007). The experience of flow and subjective well-being of music students. *Horizons of Psychology*, 16(2), 5–17.

Stroud, L. R., Salovey, P., & Epel, E. S. (2002). Sex differences in stress responses: social rejection versus achievement stress. *Biological Psychiatry*, 52(4), 318–27.

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.

Tsujita, S., & Morimoto, K. (1999). Secretory IgA in Saliva can be a Useful Stress Marker. *Environment Health and Preventive Medicine*. 4(1): 1–8

Um, E., Plass, J. L., Hayward, E. O., & Homer, B. D. (2011). Emotional Design in Multimedia Learning. *Journal of Educational Psychology*, 104(2), 485-498.

van Dijk, V. (2010). *Learning the triage procedure: Serious gaming based on guided discovery learning versus studying worked examples* (Unpublished master's thesis). Universiteit Utrecht, Utrecht, the Netherlands.

Vollmeyer, R., & Rheinberg, F. (2006). Motivational effects on self-regulated learning with different tasks. *Educational Psychology Review*, 18, 239–253.

Wang, J., Korczykowski, M., Rao, H., Fan, Y., Pluta, J., Gur, R. C., McEwen, B. S., & Detre, J. A. (2007). Gender difference in neural response to psychological stress. *Social Cognitive and Affective Neuroscience*, 2(3),227–239.

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.

Wastiau, P., Kearney C., & den Berge, W. V. (2009). How are digital games used in schools? Complete Results of the Study. European Schoolnet. http://games.eun.org/upload/gis-full_report_en.pdf. Accessed 16 Dec 2013.

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.

Wilensky, U. (1999) NetLogo, Center for Connected Learning and Computer-Based Modeling, Northwestern University. http://ccl.northwestern.edu/netlogo/ Accessed 16 Dec 2013.

Wingfield, J. C., & Sapolsky, R. M. (2003). Reproduction and Resistance to Stress: When and How. *Journal of Neuroendocrinology*, 15, 711–724.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games. *Journal of Educational Psychology*, 5(2), 249-265.

Figures

A note: all the figures can be printed in black & white.

This page was intentionally left empty. All the figures are in the main text for the sake of the review process.

Figure captions separately

This page was intentionally left empty. All the figure captions are in the main text for the sake of the review process.

Tables

This page was intentionally left empty. All the tables are in the main text for the sake of the review process.

Appendix A – Treatments' Description

A.1 EU-comp Treatment Detailed

This section extends Sec. 2.1.1.

In the diplomatic layer of *Europe 2045*, each player has his/her own project, which is a vision of how the EU should look in the future and it is formally defined by: a) a set of policies that should be put in place, b) a set that should be suspended, and c) a set to which the project is indifferent (e.g., the Green Europe project supports environmental protection and investment into alternative energy resources, while the Conservative Europe project strives to preserve traditional values). From the gaming perspective, projects present roles the students can play. Because some projects agree or disagree upon the same subset of policies, each player can find a teammate to support his/her intended particular policy change. Thus, the game features both collaborative and competitive aspects at the same time. The final appearance of Europe at the end of each game session is thus the result of intense negotiations and voting in a given player group. In this study, the game offered eight different projects; one for each student. Every project had exactly four policies associated with it. The following points are crucial because other EU* treatments differ in those:

1. General framing: In the first two "tutorial" rounds, the players were familiarized with the game's mechanisms and rules and with controlling its user interface. They were informed they would compete against each other in order to win; but they were also informed that they would also need to collaborate to win (in this study, only the diplomatic layer's outcome could influence the game ranking of the players).
2. Project selection and role-playing: The players had three minutes for reading brief, textual descriptions of eight projects. They then stated which three projects they would most like to play and were assigned the projects based on their preferences (one project can be played by one player only; therefore, the players had to select three possible projects in order to avoid conflict). At that point, each player was also assigned a member state to play (each project was always coupled with the same state) and given a flag badge and a small flag stand so as to better identify with his/her state.
3. Project introduction: Each player was given the description of his/her project and its policies. The players then had *exactly* eight minutes for reading their project description. They each then had exactly one minute for presenting the project's main visions to their fellow players (an hourglass was used for timing and the students saw it; thus the timing of presentations was indeed very precise).

In each of the subsequent four rounds (the 3rd to the 6th), the following happened.

4. Players were able to briefly control their states (i.e., play the economic layer).
5. Policy selection: Four players proposed a draft for a policy change. The players chose the draft.
6. Policy presentation: Each of these four players had exactly eight minutes to read expository texts about his/her proposed policy. Meanwhile the other four players could engage in one of the following two activities, or a combination of them. First, they could control their state. Second, they could read materials about policies associated with their own projects or about policies proposed by the other four players. After the eight minutes had passed, the discussion started. Students moved away from the computers and presented their drafts for policy

changes. During presentations, students sat in chairs; usually organized in a round or square formation. Each student proposing a policy change had exactly 1.5 minute to introduce the policy and present its benefits. At the beginning of Round 3, an assistant from the experimental team demonstrated what should be said during these 1.5 minutes (he used an unrelated example). Opponents or other proponents could then react/ask questions during a discussion moderated by the teacher (approx. 2-3 minutes).

7. Negotiation: After the four presentations, the negotiation for or against support of the proposed policy changes started (5 minutes). The teacher encouraged students to stand up, make small clusters, secretly negotiate outside of the classroom, etc. Small groups of students often informally engaged in "collaboration agreements", which usually lasted more than one round and under which the students mutually supported each others' proposals.
8. Voting: The students voted on each draft presented. The results were presented at the beginning of the next round; including the current game ranking of the players.

We operationalized "learning effectiveness" by means of the amount of knowledge about a) the player's own project; b) all other projects; c) policies each player presented himself/herself; and d) the process of negotiations on policy changes. These types of knowledge could be acquired, respectively:

a) from reading expository texts about one's own project;

b) from reading expository texts about other projects, including associated policies, and by observing the behavior of players playing the respective projects and listening to them;

c) from reading expository texts about the policies associated with the player's own project;

d) by participating in presenting drafts for policy changes and in subsequent negotiations.

We would like to add two remarks. First, if the teacher was asked a question about any policy, he said that the answer could be found in the textual materials but did not answer the question directly. This was not what a real teacher would do, but in this experiment, we wanted to measure knowledge acquired by students themselves from reading the expository texts and presenting the drafts of policy changes. We were not interested in knowledge acquired by means of the teacher's occasional answers (the questions and the number thereof were hard to predict and therefore the answers were impossible to standardize/control; however, the texts were always the same) or brief contextualizing lectures that a teacher might have given in a real class. Second, since every project had four policies associated with it in this study and each student presented a policy draft exactly twice, the student had to choose exactly two out of four policies of his/her own free will.

A.2 EU-class Treatment Detailed

This section extends Sec. 2.1.2.

Looking at the list of types of knowledge that can be acquired through *Europe 2045's* diplomatic layer (see Sec. A.1), one realizes that these types of knowledge can be acquired also in a typical classroom setting where a teacher engages students in reading the same expository texts and presenting and discussing them as in the EU-comp treatment. In other words, the constrained version of *Europe 2045*, the EU-comp treatment, can be considered as a gamified version of a project day at a school that capitalizes on the frontal teaching model augmented by reading and discussions. The EU-class treatment models such a project day in a controlled laboratory environment.

The components of the EU-comp's game learning mechanics were replaced as follows:

1. General framing: The EU-class students were told that we were investigating a new "discussion-based teaching model." The word "game" was carefully avoided. There was no competition during the whole treatment.

2. Project selection and role-playing: Each EU-class learner was paired with an EU-comp (or EU-no-comp) learner and was assigned the peer's project (i.e., no choice). Moreover, the EU-class learners did not represent their projects/states, they did not receive flag badges/stands and they were instructed "to study a project" rather than "to play a project role."

3. Project introduction: Exactly as in the EU-comp groups, the EU-class learners were instructed to read short project descriptions for three minutes; however, they were meant to select three projects that were the most in line with their own ideas/beliefs. They were then given eight minutes to study a detailed textual description of their assigned project (the project description was the same as in the EU-comp treatment) and one minute was provided to present the project's main visions to their peers.

4. Economic layer: it was absent.

5. Policy selection: Each EU-class learner was assigned a policy to study and to present, based on what his/her peer had chosen in the EU-comp group.

6. Policy presentation: The EU-class learners had eight minutes to study the assigned policy and 1.5 minutes to introduce the policy and present its benefits (as in the EU-comp group; using the same textual materials). The tables and chairs were arranged in rows as in a regular classroom, not in a round/square formation. After each presentation, the teacher invited other students (especially those who, earlier in the day, had presented a project related to the policy that had just been presented) to express their opinion regarding whether the policy should be put into force in the EU or not, when considering the context of "their" project. They could express positive as well as negative opinions and ask questions. The discussion was moderated by the teacher and it was stopped by him after 2-3 minutes. Before the first presentation started, the teacher demonstrated what should be said during the 1.5 minutes, using an unrelated example (the same example as in the EU-comp groups).

7. Negotiation: It was replaced by a discussion started by the following instruction from the teacher: "Now, please think about how the political tendency/view you read about today at the beginning of the class (i.e., the project), is related to the policies that have just been presented. For instance, it can relate to them positively, neutrally or negatively." The teacher called upon students to express their opinions about a few policies, at least, and encouraged them to discuss them (note that students were sometimes quiet).

8. Voting: It was absent in the EU-class treatment. The time allotted to voting (and playing the economic layer, Point 4) in the EU-comp groups was filled in by an unrelated short film about an EU topic at the very end of the workshop (around 20 minutes) and two short breaks in the middle.

Finally, the introduction to the game was replaced by an unrelated 40-minute-long frontal lecture on the EU using PowerPoint slides and by an unrelated 20-minute-long, pen-and-paper "heat up" mini-game on the topic of the EU and EU law.

We add two remarks. First, recall that in the EU-comp treatment, four students prepared themselves for policy presentations, over a period of eight minutes, while the other four read materials about policies associated with their own projects or read materials about policies proposed by their peers (but mostly played the economic layer of *Europe 2045*). Our pilot study showed that this format did not work well in the EU-class treatment. This was because not only could the other four players not play the economic layer, but also they were not motivated by the game behind the diplomatic layer and therefore they did not read the respective materials carefully and they tended to become bored and irritated (as they would in a regular class). Therefore, we had to replace four rounds of the EU-comp treatment with two “rounds” in the EU-class treatment. In both of these “rounds,” each participant prepared him/herself for the presentations that directly followed. For the EU-class treatment, this two-round setting was more ecologically valid than the four-round setting.

Second, note that knowledge we tested (see Sec. A.1) could be acquired neither from the game introduction nor from the voting *per se*. Thus, the EU-class treatment was not put at a disadvantage.

Appendix B – Prior Knowledge Questionnaire for EU* treatments

A1. I follow events on the international political scene:

a. not at all
b. once a week: *(select whatever options apply)* TV, online, radio, print media, other sources……………………..
c. 2-3 times per week *(select whatever options apply)* TV, online, radio, print media, other sources……………………..
d. daily *(select whatever options apply)* TV, online, radio, print media, other sources……………………..

A2. Are you able to explain what the accession criteria are for a country wishing to join the EU? *(indicate your ability on a scale of 1 (not at all) - 5 (definitely yes))*

A3. On topics related to the European Union I consider myself to be: *(select one answer)*

a. Beginner. I know a little about it.
b. Slightly advanced. I have average knowledge.
c. Advanced. I know quite a bit.
d. I don’t know anything. I am not interested in this topic.

A4. When I hear about political events in the EU, I can imagine what influences political decisions. *(indicate your ability on a scale of 1 (not at all) - 5 (definitely yes))*

A5. Subject – The Basics of Social Science: *(select one answer)*

a. This my favourite subject.
b. I find it generally interesting. I am often interested in the topics discussed.
c. I am not really interested. Most topics do not interest me.
d. It’s my least favourite subject. I literally have a negative relationship to the subject.

A6. Who is the current president of the European Commission? *(select one answer)*

a. Herman Van Rompuy
b. Catherine Margaret Ashton
c. Vladimír Špidla
d. José Manuel Durão Barroso

A7. How many member-states does the EU currently have? *(select one answer)*

a. 12
b. 15
c. 27
d. 28

A8. When did the Czech Republic join the EU? *(select one answer)*

a. 1998
b. 2001
c. 2003
d. 2004

A9. Štefan Füle is the Czech European Commissioner for: *(select one answer)*

a. Employment, Social Affairs and Inclusion
b. Enlargement and European Neighbourhood Policy
c. Agriculture and Rural Development
d. Health and Consumer Policy

Appendix C – Forming Subgroups for EU* treatment

--- Insert Table 10 around here ---

The assignment to condition was as follows. The optimal number of participants in each subgroup was eight. Table 10 shows how large the subgroups were when a number of participants other than 16 or 24 arrived. Participants were matched based on their pre-test score in the following way: in cases of 19 or less participants, pairs and usually also a few singles were formed (see the table). Singles were selected randomly. In cases of 20 or more participants, trios and usually also a few pairs or singles were formed. Members of the pairs/trios were then assigned to the subgroups randomly. Singles were assigned according to the table. In case this random assignment resulted in a situation in which the boys/girls ratio in the subgroups differed and could be improved by a swap, the assignment of one or two randomly chosen mixed-sex pairs/trios was swapped. Sometimes, one or two students had to leave before the experiment's end: in that case, the student was assigned to the EU-class condition.

Table 10 Arrangement of participants into groups for the EU* treatments.

The size of the group	EU-comp or EU-no-comp	EU-class	EU-no-comp
15 (2x)	**8**	**7**	
16 (2x)	**8**	**8**	
17	**8**	**9**	
18	8	10	
19	8	11	
20	8	6	6
21	8	7	6
22	**8**	**8**	**6**
23	8	7	8
24	8	8	8
25	8	9	8
26	**8**	**10**	**8**

Note: Only the numbers denoted by **bold** actually occurred.

Appendix D – Analysis per EU* Treatments

For completeness, we also computed correlations between affective variables in each EU* treatment separately (Tables 11 – 13). In addition, correlations of cortisol differences with affective variables per each EU* treatment are in Table 14.

While many correlations were relatively consistent across treatments, especially among Flow, PANAS+ and PANAS–, and also among cortisol variables, SIAS and RCI.cont, several others were not, such as among cortisol variables and PANAS–. Although it would be interesting to explore differences between variants of Europe 2045, we stress that these results should be considered as exploratory and interpreted with utmost caution. Because the groups were relatively small (e.g., EU-no-comp had 26 participants included in the analysis), correlation coefficients could be easily influenced by outliers. For example, if we look at Figure 14 showing relationship between *3–1* variable and PANAS– for the EU-no-comp group, we can see that slope of regression line changes as we remove one outlier from the analysis (denoted by a triangle). In our opinion, these data can be useful in guiding future research, but they should not be interpreted as firm results.

--- Insert Table 11 around here ---

Table 11 Correlation matrix of affective variables and Test score for the EU-comp treatment.

EU-comp	Flow	PANAS+	PANAS-	Like	Test score
Flow	-				
PANAS+	0.67***	-			
PANAS-	-0.50**	-0.35*	-		
Like	0.46**	0.31.	-0.26	-	
Test score	0.41*	0.32	-0.15	0.00	-

·*p* < .1 **p* < .05 ***p* < .01 ****p* < .001

--- Insert Table 12 around here ---

Table 12 Correlation matrix of affective variables and Test score for the EU-class treatment.

EU-class	Flow	PANAS+	PANAS-	Like	Test score
Flow	-				
PANAS+	0.62***	-			
PANAS-	-0.38**	-0.02	-		
Like	0.45***	0.55***	-0.05	-	
Test score	0.23	0.25	0.26	0.16	-

***p* < .01 ****p* < .001

--- Insert Table 13 around here ---

Table 13 Correlation matrix of affective variables and Test score for the EU-no-comp treatment.

EU-no-comp	Flow	PANAS+	PANAS-	Like	Test score
Flow	-				
PANAS+	0.69***	-			
PANAS-	-0.41*	-0.31	-		
Like	0.57**	0.58**	0.10	-	
Test score	0.06	0.40	0.10	0.33	-

**p* < .05 ***p* < .01 ****p* < .001

--- Insert Table 14 around here ---

Table 14 Correlations of affective variables with the *3–1* and *4–3* variables for the individual EU* treatments.

	EU-comp		EU-class		EU-no-comp	
	3-1	4-3	3-1	4-3	3-1	4-3
SIAS	0.27 (34)	-0.20 (37)	0.35*(51)	-0.40**(52)	0.34. (24)	-0.34 (24)
RCI.comp	-0.55**(28)	0.38*(31)	-0.06 (44)	0.11 (45)	-0.21 (21)	0.21 (21)
RCI.cont	-0.24 (28)	0.39*(31)	-0.25.(44)	0.30*(45)	-0.44*(21)	0.43*(21)
Like	-0.04 (34)	0.14 (37)	0.10 (51)	0.02 (52)	-0.10 (24)	-0.10 (24)
Flow	-0.09 (34)	0.19 (37)	-0.24.(50)	0.14 (51)	-0.06 (24)	0.03 (24)
PANAS+	0.15 (34)	-0.08 (37)	0.12(50)	-0.00 (51)	-0.27 (24)	0.12 (24)
PANAS-	0.11 (34)	-0.20 (37)	0.53***(50)	-0.45***(51)	0.01 (24)	0.24 (24)
Test score	-0.05 (22)	0.13 (24)	-0.05 (32)	-0.17 (33)	-0.13 (12)	-0.02 (12)

·*p* < .1 **p* < .05 ***p* < .01 ****p* < .001

--- Insert Figure 14 around here ---

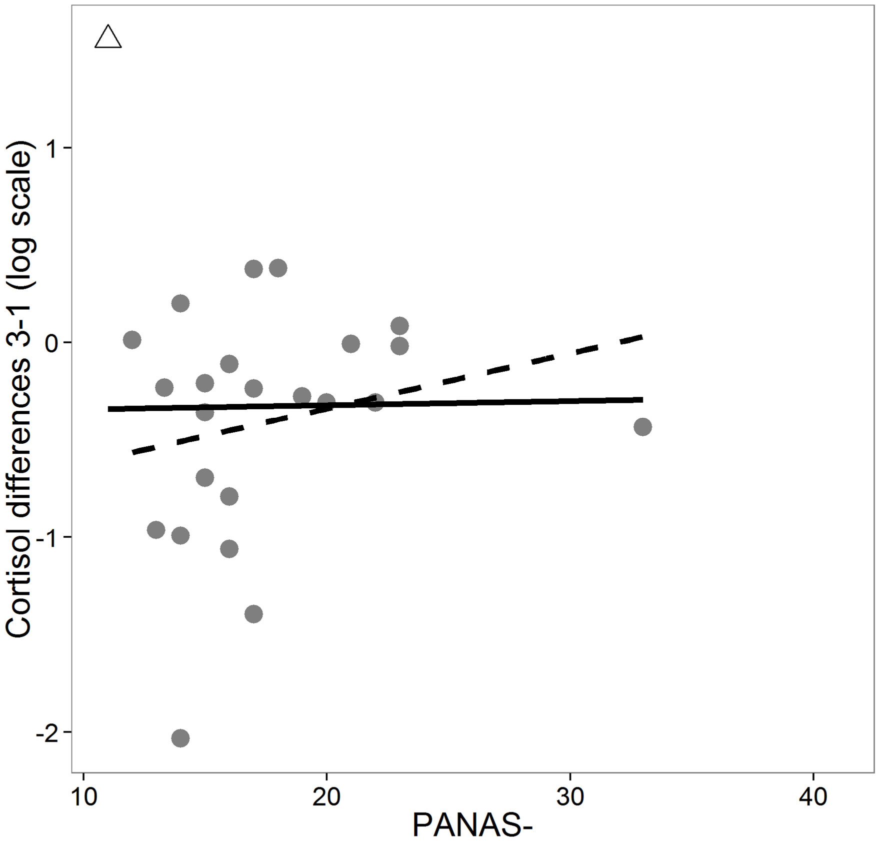


Fig. 14 Scatter plot showing relationship between the PANAS– and the *3–1* variable with regression line before (dashed line) and after removal (solid line) of the outlier denoted as triangle (EU-no-comp group).