Learning When Serious: Psychophysiological Evaluation of a Technology-Enhanced Learning Game

Ben Cowley[1,2*], Martino Fantato[1], Charlene Jennett[3], Martin Ruskov[3] and Niklas Ravaja[1,4,5]

[1]School of Business, Aalto University, 00076-AALTO, Finland // [2]Cognitive Science Unit, University of Helsinki, P. O. Box-9, 00014, Finland // [3]Department of Computer Science, University College London, WC1E 6BT, UK // [4]Department of Social Research, University of Helsinki, P. O. Box-54, 00014 University of Helsinki, Finland // [5]Helsinki Institute for Information Technology HIIT, P. O. Box-15600, 00076-AALTO, Finland // ben.cowley@helsinki.fi // martino.fantato@aalto.fi // c.jennett@cs.ucl.ac.uk // m.ruskov@cs.ucl.ac.uk // niklas.ravaja@aalto.fi

[*]Corresponding author

ABSTRACT

We report an evaluation study for a novel learning platform, motivated by the growing need for methods to do assessment of serious game efficacy. The study was a laboratory experiment combining evaluation methods from the fields of learning assessment and psychophysiology. 15 participants used the TARGET game platform for 25 minutes, while the bio-signals electrocardiography, electrodermal activity and facial electromyography were recorded. Learning was scored using pre- and post-test question-based assessments Repeated-measures analysis with Generalised Estimating Equations was used to predict scores by tonic psychophysiological data. Results indicate some learning effect, plus a relationship between mental workload (indexed by electrocardiography) and learning. Notably, the game format itself influences the nature of this relationship. We conclude that a high quality of insight is afforded by the combination of subjective self-report and objective psychophysiology, satisfying two of three observable domains.

Keywords

Technology enhanced learning, Serious games, Heart-rate variability, Mental workload, Psychophysiology, Evaluation, Competence development

Introduction

In today's global market, human capital is a recognized strategic asset in companies. Learning and training play a foundational role in talent management, but establishing effective learning strategies across enterprises remains a costly challenge without measurable return.

To support the effort to build such strategies, we report on an evaluation of a game-based Technology-Enhanced Learning (TEL) platform, which teaches about soft skills using project management scenario simulations. Based on positive outcomes from a similar previous study (Cowley, Ravaja, & Heikura, 2013), we designed a combined-methods approach to the evaluation. A multi-trial protocol allowed a repeated measures self-report battery alongside the measurement of physiological signals to index psychological constructs. The two methods combined are a complementary data-gathering tool, because self-report is subjective, discrete and explicit while psychophysiology is objective, continuous and implicit. Thus the study was carried out to examine whether psychophysiological recordings obtained during serious game play predict (short-term) learning outcomes as measured by a pre- and post-test assessment tool. We found, among others, a relationship between mental workload and learning. This has important implications for researchers interested in measuring learning (unobtrusively) in future research.

The experiment used a within subjects design based on adjustment of the participant's knowledge of project management via the learning platform. The independent variables were the physiological effects on the player of exposure to source of topic-relevant education (namely the game), while the dependent variable was the knowledge of the participant. Learning was assessed using questionnaires of two types: one set applied before and after play to test learning performance, and one set of self-report questionnaires to establish the 'felt' experience of the participant.

We begin by describing the state of the art underpinning this experiment. We cover experiment methodology and then detail our results under three themes: psychophysiological predictors of learning; test-based assessment of learning; subjective mood self-report. Finally we offer our discussion and conclusion.

State of the art

Educational game efficacy has been well debated (Egenfeldt-Nielsen, 2006; Gee, 2006). McQuiggan, Lee, & Lester (2006) draw a parallel between the factors describing student engagement and those involved in game play. However learning does not necessarily follow engagement, as argued by Kirschner, Sweller, & Clark (2006) who point out that discovery, problem-based, experiential and enquiry-based techniques are the main tools of games, but all require prior knowledge on the part of the student to evoke learning. Some suggest the solution is in scaffolding the game, i.e., instructional support during learning to reduce cognitive load (O'Neil, Wainess, & Baker, 2005).

Given recent positive results (Blunt, 2009; Ritterfeld, 2009) – and bearing in mind that some form of learning is almost *always* part of play (Koster, 2005) – the relevant question becomes: *how* will a given game work? Will a particular game teach retained, transferable skills which are the ones intended by the designers, or will it teach skills only valuable within the game context? This has become a strong theme in serious games research (Guillén-Nieto & Aleson-Carbonell, 2012).

When examining how such design variants actually work 'in the field', the players' psychological and physiological experience is of central interest, motivating the need to objectively measure this subjectivity. For the assessment of subjective experience, e.g. emotions, there are three observable domains (Bradley & Lang, 2000): i) a subjective cognitive experience (e.g. assessed by questionnaires), ii) behavioural expressions (i.e. actions and behavioural patterns assessed by implicit techniques) and iii) psychophysiological patterns. Three features of the physiology which are particularly interesting in respect of learning are arousal, cardiac and facial musculature – i.e. bodily activation, regulation and pleasure or displeasure in response to the experimental activity. These may act as a sufficient causal explanation (Peters, 1960, p-11) of observed learning.

The basic premise of psychophysiological methods for evaluation is that the study participant/player cannot give inaccurate physical signals (discounting acquisition issues), and the acquisition of signals is non-intrusive, freeing the user/player's attention. Psychophysiological methods are particularly useful for objectively examining experience: because the physiological processes measured are mostly non-voluntary, *the measurements are not contaminated by participant answering style, social desirability, interpretations of questionnaire item wording, or limits of their memory* (Ravaja, 2004).

There is emerging evidence suggesting that the synchronization of activity of different physiological response systems (i.e. response coupling) may reflect the central state of the individual particularly well (Chanel, Rebetez, Bétrancourt, & Pun, 2011), prompting our use of three separate biosignals: electro dermal activity (EDA), electromyography (EMG) and electrocardiography (ECG).

Several studies have shown that digital games (i.e., an active coping task) elicit considerable emotional arousal- or stress-related cardiovascular reactivity in terms of heart rate (HR) and blood pressure (Johnston, Anastasiades, & Wood, 1990). Previously found convergent relations between HR and arousal during digital game playing suggest that HR covaries primarily with emotional arousal during playing (Ravaja, Saari, Salminen, Laarni, & Kallinen, 2006). Henelius, Hirvonen, Holm, Korpela, and Muller (2009) explored the ability of different short-term heart rate variability (HRV) metrics to classify the level of mental workload during a variable task-load computerized test, with good results. Additionally Nishimura, Murai and Hayashi (2011) found that mental workload was well indexed by HRV in a simulator learning task; Hercegfi (2011) reported on a method to assess mental effort of human-computer interaction from HRV and other signals, reporting improvement over existing approaches.

Plotnikov et al., (2012) and Nacke, Stellmach, and Lindley, (2010) have each described approaches to measuring the engagement of players using electroencephalography.

Facial Electromyography (fEMG), direct measurement of the electrical activity associated with facial muscle contractions (Tassinary & Cacioppo, 2000), has the potential to record the unconscious emotional reactions of the player to interaction with in-game stimuli. Recording at the sites of Zygomaticus major (cheek) and Corrugator Supercilii (brow) can index positive and negative valence (Lang, Greenwald, Bradley, & Hamm, 1993; Ravaja, Saari, Turpeinen, et al., 2006; Witvliet & Vrana, 1995). Recording at the Orbicularis Oculi (periocular) can index high arousal positive valence (Ekman, Davidson, & Friesen, 1990).

Arousal is most often measured with EDA (or skin conductance level; also sometimes called galvanic skin response) (Bradley, 2000; Dawson, Schell, & Filion, 2000). EDA is an often-used physiological measure when studying digital gaming experiences (e.g., Mandryk & Atkins, 2007), as it is less susceptible to misinterpretations compared to facial EMG. The neural control of eccrine sweat glands—the basis of EDA—predominantly belongs to the sympathetic nervous system that non-consciously regulates the mobilization of the human body for action (Dawson et al., 2000).

Our primary evaluation approach was to examine the relationship of the psychophysiological signals with learning questionnaires, self-reported game experience and game experience questionnaires. We use tonic values of psychophysiological signals to index various cognitive and emotional processes which can contribute to learning. This approach is novel in its domain, as few psychophysiological studies focus on this area of soft-skills game-based learning (GBL); it is also novel in terms of methodology, as prior studies of GBL mainly used event-related analysis of the biosignals; however because learning in this type of game happens over long time periods, with players who can construct concepts from non-linear relationships in the data they are presented with, a tonic analysis approach is more appropriate.

Methodology

Participants

We enrolled 15 right-handed participants (seven females), randomly sampled from respondents (self-selected volunteers) to advertisements. Their ages ranged from 21-33 years, the mean age being 25.87 years (SD = 3.85). Participants were rewarded with three cinema tickets each. Regarding their background, all 15 participants were novices to project management. 12 from 15 used IT to support their learning, including activities such as online courses, web research, reading online journals, using mind-map programs, watching video lectures. Four out of 15 had used role-play to support their learning: three took part in a course that involved role-play and discussion; one enjoyed *'larping'* (live action role-playing). 10 of 15 currently played computer games of some kind. Two said that they gave up playing such games due to lack of time. Three did not play games.

TARGET platform

The EU-funded Transformative, Adaptive, Responsive and enGaging EnvironmenT (TARGET) project has developed an innovative TEL environment focused on project management. It was an ideal test case for our current work, because universities/enterprise settings that wish to use the TARGET platform might want to measure whether their employees/students are learning without disturbing their playing experience.

The TARGET platform features learning scenarios that employ interactive storytelling and evidence-based performance assessment. In the Stakeholder Management (SM) scenario (see Figure 1) learners get to play the role of the project manager Ingrid. Ingrid is responsible for developing a windmill electricity farm project, and has access to project management tools such as a Gantt chart which are used during the scenario to scaffold the task simulation. Ingrid's task is to negotiate with Jens, a local farmer, to convince him to sell his land for building an access road. Negotiation happens in the form of unstructured dialogue; learners converse via a text field. To help ensure convergence in how didactic ideas are expressed, the dialogue input field provided suggestions for sentence auto-completion. Participants were advised to use these as much as possible. Learners familiar with the scenario and interface usually take up to 10 minutes to negotiate, regardless of success. To ensure such familiarity participants were given a training session.

After the negotiation experience, learners are presented with the evidence-based 'Competence Performance Assessment' module (CPA). The CPA includes a video re-play of the experience synchronised to a line-graph showing the learner's estimated performance in four competences: negotiation, communication, trust-building, and risk and opportunity (see Figure 2). The playback gives learners a chance to review their behaviour, relate their actions to their dynamic competence assessment, and interpret where they went wrong or right. The CPA module is described in detail in (Parodi et al., in press).

Figure 1. Screenshot of the TARGET Stakeholder Management scenario

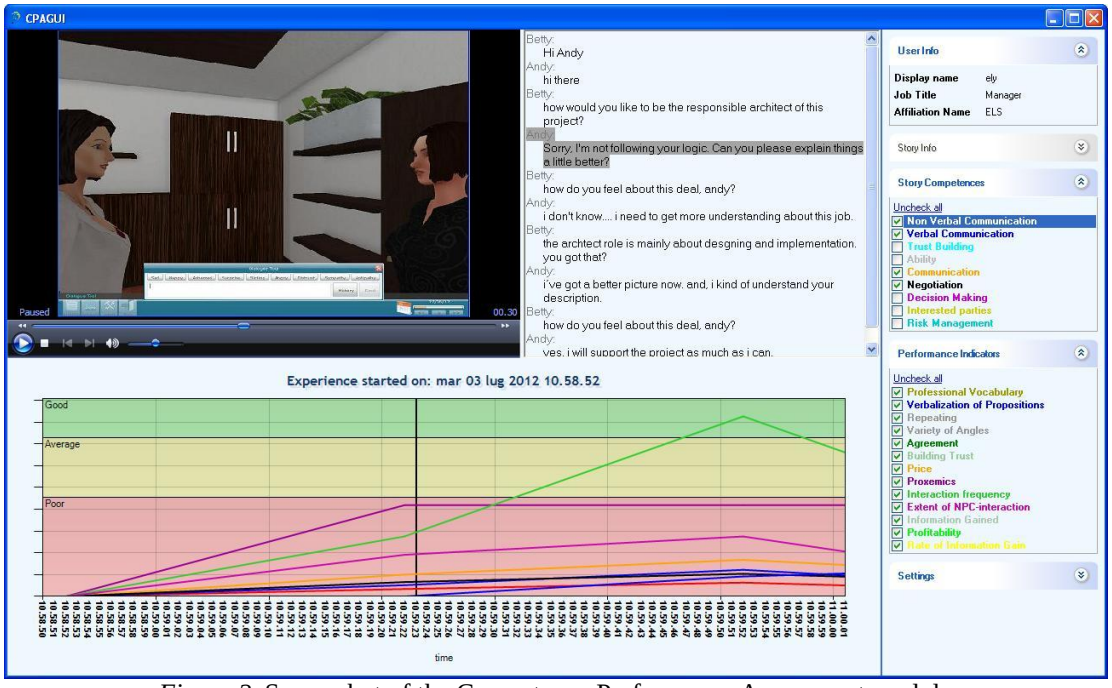

Figure 2. Screenshot of the Competence Performance Assessment module

There are several ways that the TARGET platform aims to achieve learning:

1. By playing the scenario and observing the causal logic of the dialogue system, the learner can see how factors such as style of approach, mood, small talk, etc. influence the negotiation outcome.
2. By reflecting on their performance using the CPA module, they can review their behaviour and think about how they would carry out the negotiation differently next time.
3. The learner is able to replay the scenario multiple times (in our study we instruct learners to play it twice). This allows the learner to try out different strategies, and through repeated play he/she can learn the operative principles for successful negotiation in the scenario.

Psychophysiological measures

We recorded continuously three sets of electrical potentials: EDA is a measure of arousal; EMG when applied facially (fEMG) helps index emotional expression; and ECG measures the heart's potential field and can help index many processes such as mental workload (Cowley et al., 2013).

fEMG was recorded at three separate sites: the muscle regions Corrugator Supercilii (CS, above the brow, indexes negative emotion), Zygomaticus Major (ZM, on the cheek, indexes positive emotion) and Orbicularis Oculi (OO, below the eye, indexes sincere positive emotion). Electrodes were filled with Synapse conductive electrode cream (Med-Tek/Synapse, Arcadia, CA), and impedances were ensured to be below 10kΩ.

To record EDA, electrodes were placed to onto the proximal phalanges of the non-dominant hand. Electrodes were Ag/AgCl filled with TD-246 skin conductance electrode paste (Med Assoc. Inc.). ECG was recorded at the manubrium, lower left rib and neck using electrodes with stud connector 35x45 mm (Spes Medica S.r.l.). For the recordings, the VarioPort ARM was used as DC amplifier and recording device. The specifications of the amplifier can be found in Table 1.

Table 1. Specifications of the Varioport recordings

Signal	Amplification factor	Input range	Resolution (with 16 bit converter)	Frequency Range	Sampling Rate
ECG	214 (±2%)	±5.8mV	0.18µV/bit	0.9-70Hz	64Hz
EMG	4899 (±2%)	±255mV	0.008µV/bit	57-390Hz (±2%)	128Hz
EDA		0-70µS	0.001µS		32Hz

Learning assessment materials

Several learning assessments were designed to test for learning, see Table 2. An 8-item multiple choice questionnaire (MCQ) was administered pre- and post-TARGET. Each question was presented with 5 possible answers. The MCQ was created with the aid of an instructor that used the SM scenario as a teaching exercise with his project management students. The questions were designed to cover key points of negotiation, communication and trust-building.

A 10-item self-assessment of learning questionnaire (SAL) was administered post-TARGET only. For each item, participants are asked to rate their level of agreement on a scale from 1 (strongly disagree) to 7 (strongly agree). Four items refer to competence performance (SAL-performance) and four items refer to competence learning (SAL-learning).

Table 2. Learning materials for the TARGET SM Scenario

Learning Materials	Questions
MCQ	1. What is a Gantt chart used for? 2. What is usually the best way to start a negotiation with a stakeholder? 3. When you are selling a project to the stakeholder, which of the following should be your priority to address? 4. When you want to get commitment from the stakeholder, which is the better approach and why? 5. Why is building trust with a stakeholder important? 6. Which of the following is a way to break trust with a stakeholder? 7. Ideally what is the best way to communicate with a stakeholder? 8. What language should you use when communicating with a stakeholder?

SAL	*Experience*:	
		1. Did you find the Gantt chart useful for managing your project?
		2. Did you enjoy the experience of negotiating with the game character(s)?
	Competence performance:	
		Do you think you performed well for
		3. negotiation?
		4. trust-building?
		5. communication?
		6. risk and opportunity?
	Competence learning:	
		Did playing the game and reflecting on your experience help you to learn about
		7. negotiation?
		8. trust-building?
		9. communication?
		10. risk and opportunity?

Additional materials

We used the self-report item Game Experience Questionnaire (GEQ). The GEQ measures subjective experiences relevant to play including Competence, Sensory & Imaginative Immersion, Flow, Tension, Challenge, Negative affect and Positive affect.

Evaluation design

The evaluation lasted from 2h 10min to 3h 30min (M = 2h 47min, SD = 30min), including preparation and answering questionnaires. The protocol of experiment followed 15 stages:

1. Welcome and briefing (~7min). Participants were informed of the general nature of the evaluation and asked their background information.
2. Pre-test learning measures (~35min). The second stage was to answer the pre-test learning questionnaire, wherein there were no timing restrictions; participants were then trained how to use the platform.
3. TARGET training (~10min). Training material explained the main features and operational details of the TARGET platform.
4. Electrodes and amplifier attached (~30min). After attaching the electrodes and testing the signals, an initial 5-min resting time baseline preceded the playing session of 10 minutes.
5. Baseline (5min). Both baseline and play session times were fixed.
6. Play TARGET game first time (10min). Participants played twice with the same procedure.
7. Self-report: GEQ (~2min). At the end of both game sessions participants had to fill the same self-report survey to assess subjective mood.
8. CPA reflection (5min). In between of the two sessions, there was a 5-minute reflection period, where participants were asked to review their experience using the CPA module and to think about what they did well, what they did poorly, and what they would do differently next time.
9. Break (5min). The short break was used to equalise the physical state of the player between play sessions, and check signals.
10. Baseline (5min). Due to the potential change of physical state after playing, a new baseline was recorded.
11. Play TARGET game second time (10min).
12. Self-report: GEQ (~3min).
13. Electrodes off (~5min).
14. Post-test learning measures (~30min). Finally, participants had to answer the post-test learning questionnaire, which was largely the same as the pre-test.
15. Debriefing (~5min).

Analysis

Learning in the TARGET game was assessed by two pre- post-questionnaire instruments, and one self-assessment, as described above. Four dependent variables (DVs) were derived from these by the following equalities.

- *MCQdiff* = MCQ scores, post-test minus pre-test.
- *EXP* = average scores of SAL experience.
- *Perform* = average scores of SAL competence performance.
- *Learn* = average scores of SAL competence learning.

Psychophysiological data processing procedure

EDA signal was pre-processed using Ledalab (v 3.43) in batch mode: down-sampled to 16 Hz and filtered using Butterworth low-pass filter with cut-off 5 Hz and order 8. The signal was divided in phasic (EDAP) and tonic (EDAT) components using the nonnegative de-convolution method (Benedek & Kaernbach, 2010). Thus phasic EDA and tonic EDA were two of our Independent Variables (IVs); IVs were defined as mean values of the relevant signal over some epoch – in this experiment, epoch length was one minute. Using phasic EDA does not contradict our analysis approach because it is nevertheless a continuous signal component which is *hypothesised* to be the response to discrete events.

ECG signal was pre-processed using Ecglab toolbox for MATLAB (Carvalho, Da Rocha, Nascimento, Souza Neto, & Junqueira Jr, 2002). R-peaks were identified from the original 64 Hz series and corrected for ectopic beats. Inter-beat interval (IBI) time series was obtained by interpolating with cubic splines at 4 Hz. We extracted two features as IVs: Heart Rate (HR) and Heart Rate Variability (HRV).

EMG signals were pre-processed using standard MATLAB functions. They were filtered using 50Hz notch filter, smoothed and square roots of the means squared features were extracted. The resultant signal for each fEMG electrode location was an IV – that is, fEMG 1 = Zygomaticus major, fEMG 2 = Corrugator Supercilii, fEMG 3 = Orbicularis Oculi.

In total there were seven IVs representing the recorded psychophysiology. Every psychophysiological variable also had a baseline, which was derived by the same pre-processing procedure as the mean value over the entire baseline recording period. Baselines are important to correct psychophysiological data which does not conform to absolute ranges – one person, on any given day, may present with greater or lesser baseline levels of activation of any given psychophysiological signal. From all signal files, one minute mean values of each of the IV signals described above were extracted and tabulated with the background and self-report data gathered from participants, into a 2D data matrix suitable for analysis, oriented with repeated measures over time row-wise and variables/factors column-wise.

Statistical analysis

The full list of DVs was *MCQdiff, EXP, Perform* and *Learn*. The full list of IVs included: EDAP, EDAT, HR, HRV, fEMG Zygomaticus, fEMG Orbicularis Oculi, fEMG Corrugator Supercilii. Additional factors included *Gender* and the dichotomous background variables numbered 2-4 in the list above (first paragraph in Methods section): *IT learner*, *role-player* and *game-player*. Covariates were the baseline measures for each of the IVs, and *Age*.

The Generalized Estimating Equations (GEE) procedure in SPSS was used to analyse the data, and separate analyses were carried out for each of the DVs. For every model, we specified participant ID as the subject variable, and game-play trial and minute as the within-subject variables. On the basis of the Quasi-likelihood under Independence Model Criterion (QIC), we specified Independent as the structure of the working correlation matrix. Due to the ordinal nature of the DVs, we specified an ordinal distribution with the Logistic link function.

GEEs are an extension of the generalized linear model, and were first introduced by Liang and Zeger (Liang & Zeger, 1986). GEEs allow relaxation of many of the assumptions of traditional regression methods such as normality and homoscedasticity, and provide the unbiased estimation of population-averaged regression coefficients despite possible misspecification of the correlation structure.

Results

Psychophysiological predictors of learning

Taken from the complete analysis, the statistically significant psychophysical results are summarised in Table 3, showing each physiological variable (IV) that predicted learning (DV) and associated statistics. Analysis of participants' background data showed some important group-level distinctions. A *t*-test of gender across the five DVs mentioned in the previous section showed that men and women scored significantly differently in *MCQdiff*, mean male scores were higher, $t_{(13)} = 2.84$, $p < .05$. Thus we know that men performed significantly better than women on *MCQdiff*, but the small sample size gives little chance of drawing a conclusion from this alone. A *t*-test of levels of *role-playing* across the *Perform* DV showed that those who said they used it in learning scored significantly higher, $t_{(14)} = 2.2$, $p < .05$. These results give context to the analysis, and indicate which of the available data to include as factors or covariates in the GEE models.

Table 3. Summarised statistical results for psychophysiology. Only significant results are reported

IV	B	SE	Wald Chi-Square (df=1)	DV
Task-level HRV ****	-0.59	0.16	12.87	MCQdiff
Tonic EDA *	5.03	2.26	4.95	MCQdiff
fEMG Zygomaticus ***	-1.54	0.53	8.49	MCQdiff
fEMG Orbicularis *	-0.85	0.42	4.03	MCQdiff
HR **	3.19	1.18	7.32	Perform
HRV *	0.49	0.25	3.70	Perform
EDA ***	-2.4	0.85	8.05	Learn

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$, **** $p < 0.001$

GEE models were tested for each of the five learning measurement DVs, with main effects of one IV (out of seven), the baseline (for that particular IV), *Age, Gender,* and the dichotomous background variables *IT learner*, *role-player* and *game-player*. This gave 35 models. Four of the seven IVs for *MCQdiff* were significant.

Task-level *HRV* was negatively associated with *MCQdiff*. That is, *decreased HRV during game playing was associated with better learning*. In contrast tonic *EDA* was positively associated with *MCQdiff*. Thus, *general level of basal arousal was increased for those with better learning*.

Both *fEMG Zygomaticus* and *fEMG Orbicularis* are indices of positive emotion, as explained above. *fEMG Zygomaticus* was negatively associated with *MCQdiff*, and *fEMG Orbicularis* was negatively associated with *MCQdiff*. Thus, *positive emotional responses were decreased for those who showed better learning by multiple choice questions*. Any link between *MCQdiff* and a negative or neutral emotional reaction to the game is not suggested by statistical testing of participants' GEQ self-reports. Two items tested as significant, but only marginally; they were GEQ responses 'I forgot everything around me' (GEQ5) and 'I felt challenged' (GEQ12), by SPSS General Linear Model (GLM) Univariate procedure with *MCQdiff* as DV and each self-report item as covariate.

HR was positively associated with *Perform*. *HRV* was only marginally significant, but was positively associated with *Perform*. Therefore, in the *Perform*-related self-assessments participants with higher HR and HRV rated their performance more highly.

Finally, phasic *EDA* was negatively associated with *Learn*. This implies that the more participants were activated/aroused by the events of the game, the lower that they rated themselves in the *Learn*-related self-assessments. All results are further explored and discussed in the Discussion below.

Question-based assessment of learning

Multiple choice questionnaire (MCQ)

The mean MCQ score before TARGET was 4.67 out of 8 (SD = 1.50). The mean MCQ score after TARGET was 5.20 out of 8 (SD = 1.47). A paired samples *t*-test revealed that this was not a significant difference, $t_{(14)} = -1.372$, *ns.*

Figure 3 shows frequencies of participants that answered each MCQ correctly before and after TARGET. The number of participants that responded after TARGET with the correct answer increased for the MCQs 1, 2, 3, 6, and 7; and decreased for the rest. Paired samples t-tests revealed that there were significant differences for MCQ1 ($t_{(14)} = -2.449$, $p < .05$) and MCQ2 ($t_{(14)} = -3.055$, $p < .01$). This suggests that after playing TARGET, participants had better knowledge of Gantt charts (MCQ1) and the importance of small talk for negotiations (MCQ2). Additionally, for those items which did not have *individually* significant gain scores, we tested the items with positive gains (3, 6 and 7) against those with negative gains (4, 5, and 8). A paired-samples *t*-test for summed gain scores showed significant difference, $t_{(14)} = 2.703$, $p < .05$.

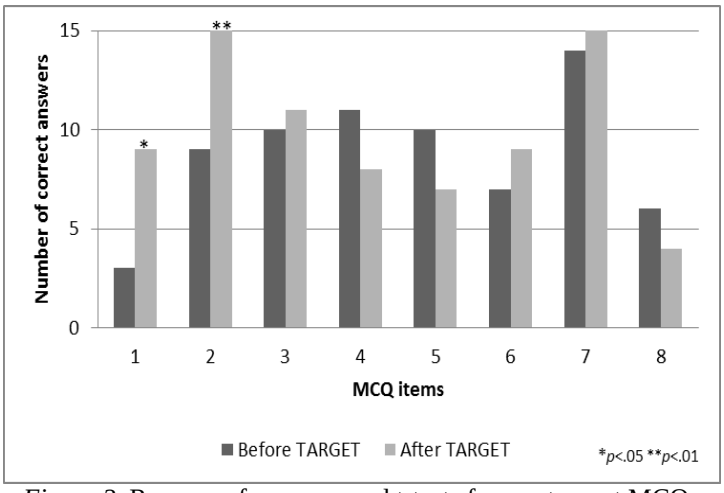

Figure 3. Response frequency and t-tests for pre-to-post MCQs

Self-Assessment of learning questionnaire (SAL)

In Figure 4 we present frequencies of responses, where ratings of 1-3 are grouped under *"disagree,"* 4 as *"neutral,"* and 5-7 are grouped under *"agree."* To specify with respect to the DVs, SAL items 1 and 2 relate to experience; SAL items 3, 5, 7 and 9 relate to competence performance (DV *Perform*); SAL items 4, 6, 8, and 10 relate to competence learning (DV *Learn*). The results suggest that opinions were mixed for experience, competence performance and competence learning.

Although SAL was not a test, we can still gain some insights from the relative distribution of responses. For instance, we can see that the greatest proportion of positive responses, six, were given to items 3 and 4: participants believed they both performed well and learned about negotiation. In contrast the least number of positive responses, two, were given to items 7 and 9: performed well for communication and for risk and opportunity. These two items also had low disagreement rates – therefore mainly there was a large neutral response. Most negative responses, eight, went to item 10: learning about risk and opportunity. Items with the least negative responses were 2, 3 and 7: enjoyed negotiating, performed well for negotiation and communication.

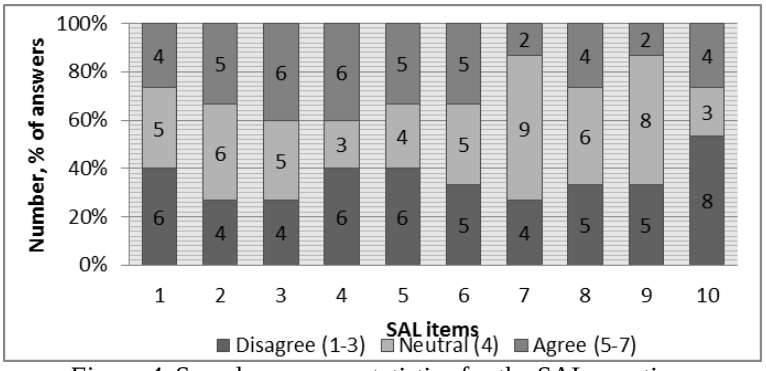

Figure 4. Sample response statistics for the SAL questions

State-based self-reports

The self-report items asked after each game session give a subjective impression of the game experience, complementing the psychophysiological data. Mean and standard deviation is given for each scale item, and figures 4-6 describe the responses in terms of the percentage of respondents corresponding to the low-medium-high levels in the scale (summed over both response times, but missing one participants' data, i.e., 29).

Game experience questionnaire (GEQ)

Participants rated their game experience from 1 to 5, with 14 items which load to seven factors: Competence (M = 2.12, SD = 0.07), Sensory & Imaginative Immersion (M = 2.48, SD = 0.83), Flow (M = 2.22, SD = 0.22), Tension (M = 2.76, SD = 0.29), Challenge (M = 2.59, SD = 0.44), Negative affect (M = 1.97, SD = 0.00), Positive affect (M = 2.69, SD = 0.15).

In Figure 5 we present the response distribution. We used a t-test to determine which response distributions differed significantly from the response 'Not at all'; which we interpret to mean which experiential areas the game elicited some reaction. By this test we can say that a significant (all at $p < .05$) number of participant responses were either moderately or extremely positive for items Positive affect, $t_{(14)} = 2.9$; Tension, $t_{(14)} = 2.3$; Sensory & Imaginative Immersion, $t_{(14)} = 2.4$, and Challenge, $t_{(14)} = 2.8$; contrasting with items Negative affect, Competence, and Flow, for which two-thirds or more responded 'not at all'.

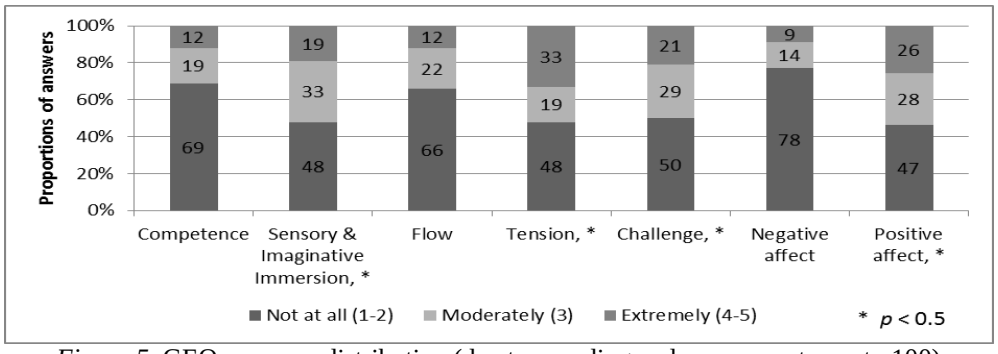

Figure 5. GEQ response distribution (due to rounding, values may not sum to 100)

Discussion

The study aimed to evaluate a novel TEL game, using a combined methods approach with learning assessment and psychophysiology. We propose that the combined approach added value by giving insights where each method alone suffers from ambiguity of interpretation.

Based on MCQ pre- to post-assessment, we conclude that the TARGET game has potential to support learning, due to two positive *MCQdiff* results. However because the gain scores were not significant *overall*, it seems the platform is not yet operating at full potential. We will attempt to explain why using the measurements taken.

The two significant *MCQdiff* results, from MCQs 1-2, contrast in their content with MCQs 2-8, which showed no significant improvement. Specifically, MCQ 1, "what is a Gantt chart…" deals with a more concrete issue reflecting a defined action that the player had to take in the game. For MCQ 2, "the best way to start a negotiation…" is more often realised than the issues from later MCQs because participants start negotiations more often than they successfully conclude them.

Contrasting those MCQs which had positive gains against those with negative gains, MCQs 3, 6 and 7 seem to be more straightforward, dealing with simpler concepts. They may also have been easier to remember from the pre-test, resulting in a gain based on priming.

The *SAL* results, i.e., areas where participants felt that they learned and performed well, imply that they may have had difficulty parsing causal mechanics of game. In other words, they could see that the outcomes of their negotiations were often good, but they did not have much access to the reasoning behind success, i.e., the risks and opportunities underlying their project were not well understood.

For GEQ affective self-report, there is an apparent paradox that relatively good Positive affect and Immersion can occur in the presence of relatively high tension and low competence. We can shed some light on the cause of this puzzle by studying responses for one particular factor. In this (short-form) version of GEQ, Immersion is based on two items: 1 "I was interested in the game's story," and 2 "I found it impressive" For item 1, only 28% responded 'Not at all' (the lowest for all responses); for item 2, this figure was 69%. Thus we see a dichotomy between participants' interest in, and how impressed they were by the game: perhaps indicating that problems were more in the delivery than the content. It appears the quality of the experience was not rewarding, evidenced by low perceptions of success (69% 'Not at all' responses to GEQ:Competence). Oddly, this contrasts with low perceptions of challenge (50% 'Not at all' responses for GEQ:Challenge).

Overall, self-report and testing results show that participants felt intrigued and challenged, but also frustrated by their inability to win the game. These results *may suggest* the interpretation that difficulty in using the game impeded important learning: however the standalone value of self-report evidence is the subject of some debate (Yannakakis & Hallam, 2011), and psychophysiology offers a complementary perspective that helps address the specific concerns of subjectivity and imprecision.

Psychophysiology

Examining the various IVs which predict *MCQdiff* reveals the nature of the relationship between player and game. Recalling that *MCQdiff* is a gain score, similarly to (Cowley et al., 2013), we can draw on the picture established there with respect to HRV. It is evident that learning is not immediately dependent on the particular direction of the HRV relationship. Rather, because *HRV* indexes mental workload, a given game may require a) more or b) less mental workload *relative to baseline values* in order to induce learning. The TARGET game clearly falls in category a), perhaps because it is designed as a business simulator. In our proxy game study (Cowley et al., 2013), we used the educational game *Peacemaker* (Impact Games 2007). The Peacemaker serious game was the designed to teach a peace-oriented perspective on the Israel-Palestine conflict. It is a point-and-click strategy game, where the player acts as a regional leader and must choose how to react to the (deteriorating) situation, deploying more or less peaceful options from a menu in a turn-based manner. It is possible that players who performed better in *Peacemaker* took advantage of cognitive efficiency enabled by the more game-like and polished elements of that game's design, and thus *Peacemaker* falls in category b).

For higher scoring players of the TARGET game, the fact that decreased *HRV* indicates increased mental workload is corroborated by their increased tonic *EDA*. Basal arousal is increased as a function of the task-related workload.

It seems that players who learned may not have found the experience unpleasant, but it was nevertheless not pleasant, was challenging and possibly frustrating, and took effort. This is contextualised by the results for the SAL questions, where participants who were more aroused in response to the events of the game also rated themselves more critically on their performance and learning. If the trend among participants was to perform better with less positive, more negative emotional responding, then it follows that activation in response to game events would have been generally of a negative nature. In other words, failure/setbacks (being more common) aroused a response more often than success.

In general, because four of seven IVs for *MCQdiff* and four of seven items from GEQ were significant, we might conclude that the game was more affecting than not, creating an intriguing if not perfect experience. Combining self-report, testing and psychophysiology, the results suggest a challenging experience for participants as they attempted to learn from the beta TARGET platform.

Issues and future work

Although the sample size may be unusually small in the educational literature, it is reasonable for psychophysiological studies. One can also claim based on a Bayesian inferential argument (Wagenmakers, 2007) that for a given *p* value, small N actually constitutes greater evidence against the null hypothesis.

Some study design issues became apparent on analysis. The dependent metrics used were all based on interstitial *self-report*, a method which has pros and cons. The reports come only at quite infrequent intervals. Using a regular pop-up prompt to gather closer-to-real-time data can increase the frequency, although only at the cost of invasiveness. In this study, using pop-up self-report was infeasible at the time and is a subject for future work.

Further study in the area may benefit from use of more psychophysiological signals such as EEG. While EEG itself indexes several informative features of player experience (Nacke et al., 2010), others have also found evidence that the fusion of several physiological modalities increases emotion recognition accuracy (Chanel et al., 2011) – offering the potential to apply such methods in TEL contexts. Such systems have been developed for the TARGET project (Bedek et al., 2011) and thus early work suggests their applicability also to learning games.

Conclusion

In this study we analysed psychophysiological recordings of players of the TARGET game with respect to their learning, as measured by pre- and post-test questionnaires. Several interesting results were found, pointing to a specific form of interaction for this game, where best results follow an effort of high mental workload and serious play. This may be opposed to types of TEL games where cognitive efficiency builds from the exigencies of playing and thus *reduced* mental workload is predictive of better learning (Cowley et al., 2013). In other words, given that HRV is negatively correlated with workload, for complex games a decrease in HRV can be a prediction of learning, while for simpler games an increase in HRV can be a prediction of learning. The difference may lie in the level of complexity which players are expected to master *without* some scaffolding from entertainment game-design techniques such as skill-chaining (building skills sets from easy component skills mastered one at a time). The closer that a TEL game gets to representing a real-world problem in real-world terms, the greater will be its inherent complexity and required mental workload.

The capability to make this distinguishing insight seems to be afforded by the combination of subjective self-report and objective psychophysiology, suggesting a methodology which can inform directions for future evaluation of such types of learning platforms. While interpretation of evaluation tests is always dependent on the researchers to some degree, the application of empirical techniques can help by corroborating or falsifying our stories.

Acknowledgments

This research was undertaken in the TARGET project, partially-funded by the European Community under the Seventh Framework Programme (Grant Agreement IST 231717).

References

Bedek, M. A., Cowley, B., Seitlinger, P., Fantato, M., Kopeinik, S., Albert, D., & Ravaja, N. (2011, November). *Assessment of the Emotional State by Psycho-physiological and Implicit Measurements.* Paper presented at the International Conference on Multimodal Interaction, Alicante, Spain.

Benedek, M., & Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, *47*(4), 647–658. doi:10.1111/j.1469-8986.2009.00972.x

Blunt, R. (2009, December 1). Do serious games work? Results from three studies. *eLearn*. doi:10.1145/1661377.1661378

Bradley, M. M. (2000). Emotion and motivation. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (2nd ed., pp. 602–642). New York, NY: Cambridge University Press.

Bradley, M. M., & Lang, P. (2000) Measuring emotion: Behavior, feeling and physiology. In R. Lane & L. Nadel (Eds), *The Cognitive Neuroscience of Emotion* (pp. 242-276). Oxford University Press, New York.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. doi:10.1191/1478088706qp063oa

Carvalho, J. L. A., Da Rocha, A. F., Nascimento, F. A. O., Souza Neto, J., & Junqueira Jr, L. F. (2002). Development of a matlab software for analysis of heart rate variability. *Proceedings 6th International Conference on Signal Processing* (pp. 1488–1491). doi: 10.1109/ICOSP.2002.1180076

Chanel, G., Rebetez, C., Bétrancourt, M., & Pun, T. (2011). Emotion assessment from physiological signals for adaptation of game difficulty. *Systems, Man and Cybernetics, Part A: Systems and Humans, 41*(6), 1083-4427. doi: 10.1109/TSMCA.2011.2116000

Cowley, B., Ravaja, N., & Heikura, T. (2013). Cardiovascular physiology predicts learning effects in a serious game activity. *Computers & Education*, *60*(1), 299–309. doi:10.1016/j.compedu.2012.07.014

Dawson, M. E., Schell, A. M., & Filion, D. L. (2000). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (2nd ed.). Cambridge, UK: Cambridge University Press.

Egenfeldt-Nielsen, S. (2006). Overview of research on the educational use of video games. *Nordic Journal of Digital Literacy*, *1*(3), 184–213.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology. II. *Journal of personality and social psychology*, *58*(2), 342–353.

Gee, J. P. (2006). Are video games good for learning? *Nordic Journal of Digital Literacy*, *1*(3), 172–182.

Guillén-Nieto, V., & Aleson-Carbonell, M. (2012). Serious games and learning effectiveness: The case of It's a Deal! *Computers & Education*, *58*(1), 435–448. doi:10.1016/j.compedu.2011.07.015

Henelius, A., Hirvonen, K., Holm, A., Korpela, Jussi, & Muller, K. (2009). Mental workload classification using heart rate metrics. *Engineering in Medicine and Biology Society, Annual International Conference of the IEEE*. doi: 10.1109/IEMBS.2009.5332602

Hercegfi, K. (2011). Heart rate variability monitoring during human-computer interaction. *Acta Polytechnica Hungaria Journal of Applied Sciences*, *8*(5), 205-224.

Johnston, D. W., Anastasiades, P., & Wood, C. (1990). The relationship between cardiovascular responses in the laboratory and in the field. *Psychophysiology*, *27*(1), 34–44.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Work : An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching, *41*(2), 75–86.

Koster, R. (2005). *A theory of fun for game design*. Phoenix, AZ: Paraglyph Press.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, *30*(3), 261–273.

Liang, K.-Y., & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Mandryk, R. L., & Atkins, M. S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International journal of human-computer studies.*, *65*(4), 329.

McQuiggan, S., Lee, S., & Lester, J. C. (2006). Predicting user physiological response for interactive environments: An inductive approach. In J. Laird & J. Schaeffer (Eds.), *Proceedings of the 2nd Artificial Intelligence and Interactive Entertainment* (pp. 60–65). CA: Association for the Advancement of Artificial Intelligence (AAAI).

Nacke, L. E., Stellmach, S., & Lindley, C. A. (2010). Electroencephalographic assessment of player experience: A pilot study in affective Ludology. *Simulation & Gaming*, *42*(5), 632–655. doi: 10.1177/1046878110378140

Nishimura, N., Murai, K., & Hayashi, Y. (2011). Basic study of a mental workload for student's simulator training using heart rate variability, salivary amylase activity and facial temperature. *Proceedings of 6th International Conference on System of Systems Engineering (SoSE)* (pp. 27-30). doi: 10.1109/SYSOSE.2011.5966575

O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. *Curriculum Journal*, *16*(4), 455–474. doi:10.1080/09585170500384529

Parodi, E., Vannucci, M., Jennett, C., Bedek, M., Ruskov, M., & Celdran, J. M. (in press). Analysing players' performances in serious games. *International Journal of Technology Enhanced Learning*.

Peters, R. S. (1960). *The concept of motivation*. New York, NY: Humanities Press.

Plotnikov, A., Stakheika, N., De Gloria, A., Schatten, C., Bellotti, F., Berta, R., ...Ansovini, F. (2012). Exploiting real-time EEG analysis for assessing flow in games. *2012 IEEE 12th International Conference on Advanced Learning Technologies* (pp. 688–689). doi:10.1109/ICALT.2012.144

Ravaja, N. (2004). Contributions of psychophysiology to media research: Review and recommendations. *Media Psychology*, *6*(2), 193–235.

Ravaja, N., Saari, T., Salminen, M., Laarni, J., & Kallinen, K. (2006). Phasic emotional reactions to video game events: A psychophysiological investigation. *Media Psychology*, *8*(4), 343–367.

Ravaja, N., Saari, T., Turpeinen, M., Laarni, J., Salminen, M., & Kivikangas, M. (2006). Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and Virtual Environments*, *15*(4), 381–392.

Ritterfeld, U. (2009). *Serious games: Mechanisms and effects*. New York, NY: Routledge.

Tassinary, L. G., & Cacioppo, J. T. (2000). The skeletomotor system: Surface electromyography. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (2nd ed., pp. 163–199). Cambridge, UK: Cambridge University Press.

Wagenmakers, E. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779 – 804.

Witvliet, C. V, & Vrana, S. R. (1995). Psychophysiological responses as indices of affective dimensions. *Psychophysiology*, *32*(5), 436–443.

Yannakakis, G. N., & Hallam, J. (2011). Ranking vs. preference: A comparative study of self-reporting. *Proceedings of the 4th international conference on Affective computing and intelligent interaction* (pp. 437 – 446). doi: 10.1007/978-3-642-24600-5_47