British Journal of Educational Technology *Vol 48 No 1 2017* 83–100
doi:10.1111/bjet.12342

The effectiveness of three serious games measuring generic learning features

Maartje Bakhuys Roozeboom, Gillian Visschedijk and Esther Oprins

Maartje Bakhuys Roozeboom is a researcher at TNO, Leiden, The Netherlands. Her main interests are in the area of the evaluation research in various domains such as occupational health and serious gaming. Gillian Visschedijk is a researcher at TNO, Soesterberg, The Netherlands. Her main research interests are in the area of serious games in various domains such as education, health, crisis management and military. Her work is characterized by both design projects and validation studies. Esther Oprins is a researcher at TNO, Soesterberg, The Netherlands. Her main research interests are training design, assessment, evaluation and validation research of training with new technology (serious gaming, simulation, simulators), in which she has a PhD, in various domains such as education, aviation and military. Address for correspondence: Ms Maartje Bakhuys Roozeboom, TNO, Schipholweg 99-89, 2316 ZL Leiden, The Netherlands. Email: maartje.bakhuysroozeboom@tno.nl

Abstract

Although serious games are more and more used for learning goals, high-quality empirical studies to prove the effectiveness of serious games are relatively scarce. In this paper, three empirical studies are presented that investigate the effectiveness of serious games as opposed to traditional classroom instruction on learning features as well as learning outcomes. All three studies used a similar longitudinal case-control study design and measured the same set of learning features (control, challenge, feedback, engagement and social interaction). Learning outcomes were measured by self-report and knowledge tests. Results of the three studies show that students that played the serious games scored higher on features associated with high-quality learning. Furthermore, the studies show that serious gaming is more effective on self-reported learning outcomes than traditional classroom instruction. Effects of serious gaming on the knowledge tests were not found. The studies serve as a first step to the development of a generic evaluation framework for serious gaming.

Introduction

Evidence on serious games

The field of serious games with a learning goal is growing, as more and more people experience its benefits compared with more traditional teaching approaches. From a learning perspective, a major advantage of serious gaming is that competences can be acquired in a realistic, attractive and challenging manner (Gee, 2007; Shaffer, 2006; Squire, 2003). Serious games allow learners to learn actively in an authentic, flexible and social learning environment (Squire, 2006, 2011). They are also assumed to be intrinsically motivating and engaging (Csikszentmihalyi, 1990; Malone, 1981; Pavlas, 2010) and give responsibility to the learner. This latter aspect may enhance the learner's self-efficacy (Bandura, 1997) and may facilitate the development of self-directed learning (Percival, 1996; Stubbé & Theunissen, 2008).

Yet as stated by Hays (2005), Sitzmann (2011), Connolly, Boyle, MacArthur, Hainey and Boyle (2012), and Harteveld (2012), well-designed empirical studies proving the effectiveness of serious gaming are still scarce. As a result, many educational institutes wait to implement games. The request for the "proven value" of serious games is increasing, as is the need for evidence on

Practitioner Notes

What is already known about this topic

- Serious games are increasingly accepted as a potentially valuable, efficient and effective alternative for conventional forms of education, training or other applications (Bedwell, Pavlas, Heyne, Lazzara & Salas, 2012).
- Serious games allow learners to learn actively in an authentic, flexible and social learning environment.
- Serious games are also assumed to be intrinsically motivating and engaging (Csikszentmihalyi, 1990; Malone, 1981; Pavlas, 2010) and give responsibility to the learner.

What this paper adds

- High-quality empirical studies that measure the effectiveness of serious gaming are scarce. This paper fills this gap by presenting an empirical longitudinal case-control study on the effectiveness of serious games in relation to classroom instruction.
- This paper describes the result of three studies executed in three different schools using a similar study design. This makes it possible to draw more general conclusions.
- In order to get insight into the "black box" of learning with serious gaming, the mediating role of learning features is investigated as well.

Implications for practice and/or policy

- The results of the three studies have shown that it has a surplus value to measure learning features in combination with learning outcomes. It showed the value of measuring process variables besides outcome variables as this can explain why certain learning outcomes are achieved (Alvarez, Salas & Garofano, 2004; Kraiger, Ford & Salas, 1993; Tannenbaum, Cannon-Bowers, Salas & Mathieu, 1993).
- More insight into the learning mechanism of serious gaming can help game designers to improve their serious games based on proven effects.
- It seems that learning through serious gaming appears to be very qualitative. This study shows that serious gaming has a more positive effect on learning features (being indicative for a high-quality learning process) than classroom instruction.

"what works." That is, for which target groups and type of learning tasks are serious games the optimal alternative to regular educational methods. This type of evidence is strongly desired (Bedwell *et al*, 2012; Mayer, 2012; Oprins & Korteling, unpublished data). There is a natural reason why there are not many well-designed empirical studies. To conduct scientific research with large numbers of learners in educational settings is often too expensive and involves lots of practical problems. Most studies, therefore, are limited to smaller pilots and expert or user evaluations (eg, Oprins & Korteling, unpublished data).

There are basically two purposes of the research studies described in the paper. The first is to contribute to the request for proven value. We do so by presenting three empirical studies investigating the effectiveness of serious games. The second is to provide insight into how the effect of serious games can be measured. A similar experimental design for the three studies is used, and its substantiation and use in a broader sense is discussed.

Effectiveness research on serious games

Empirical studies measuring the effectiveness of serious games usually focus on the learning outcomes, that is, the extent to which the learning objectives have been achieved by playing the game (see various literature reviews, eg, Sitzmann, 2011). Although this is really important, it is not sufficient to improve gaming design in our view. By also measuring generic constructs related to the mechanisms that contribute to the game effectiveness, multiple games can be compared with each other based on their particular design features (Mayer, 2012). In learning evaluation research in general, this approach is already agreed on (Kirkpatrick, 1976, 1994). In this context, a distinction has been made between "process variables" as opposed to "outcome variables" (Alvarez *et al*, 2004; Kraiger *et al*, 1993; Tannenbaum *et al*, 1993). The process variables refer to general characteristics of the learning process, while the outcome variables are domain specific, based on the learning objectives that differ per case. In order to get insight into the "black box" of learning, both types of measures should be included in well-performed effectiveness research. If the same set of process-related learning features is used in multiple studies, the research outcomes make it possible to generalize over more games. This results in a better understanding of the aspects that make serious games effective. In the empirical studies presented in this paper, this approach is followed.

Learning features

This leads to the question which generic process variables should be measured to fulfill this ambition. Therefore, a set of learning features (process variables) has been derived that is enhanced by specific gaming design elements. The idea is that these learning features are the key reasons why serious games are supposed to be effective. Summarizing, the most important learning features, being measured in the empirical studies discussed further are: feedback, control, challenge, rules and goals, engagement and social interaction.

Feedback (eg, Kiili, 2005; Pavlas, 2010; Sweetser & Wyeth, 2005) or assessment (Bedwell *et al*, 2012) refers to the information related to the progress toward goals (how well am I doing?), direct effects of actions, instructional support and learning from own mistakes.

In addition, control (eg, Bedwell *et al*, 2012; Garris, Ahlers & Driskell, 2002; Kiili, 2005; Koster, 2005; Malone, 1981; Pavlas, 2010; Sweetser & Wyeth, 2005) is related to the sense of agency or influence on the learning process by choosing your own pace, order and strategy. That is, the extent to which a game allows freedom of exploration.

Challenge (eg, Bedwell *et al*, 2012; Garris *et al*, 2002; Kiili, 2005; Malone, 1981; Sweetser & Wyeth, 2005) or balance between skills and challenge (Pavlas, 2010), or pleasant frustration (Gee, 2005) are related to the actual content of the game and the problem or challenge the player is faced with. This problem should not be too easy or too difficult; it should progress during play as skills generally improve and it is characterized often by uncertainty or mystery.

Furthermore, rules and goals (Bedwell *et al*, 2012; Garris *et al*, 2002) or clear goals (eg, Kiili, 2005; Pavlas, 2010; Sweetser & Wyeth, 2005) refer to the core of the gameplay. It describes the reasons for which the player interacts with the game world and the motivation for their in-game actions. The rules determine the method by which a player can solve problems in the game, and reach the goals of the game.

Engagement or immersion (Bedwell *et al*, 2012; Pavlas, 2010; Sweetser & Wyeth, 2005) or task involvement (Garris *et al*, 2002) refer to the game player's subjective acceptance of a game's reality and their degree of involvement and focus on the reality/task.

Finally, social interaction (Sweetser & Wyeth, 2005) or human interaction (Bedwell *et al*, 2012) is related to the social interaction with fellow students within the game or outside; for example, during a reflection session (Gee, 2009).

These learning features applied in games are assumed to enhance effective learning for various reasons from a general learning perspective and are therefore appropriate means to use to measure differences in effectiveness between serious gaming and classroom instruction.

Goal of the study

This paper describes three related effectiveness studies using the same methodology and measuring similar general process variables next to game-specific outcome variables. These process variables are the learning features that could be applied at many games used in education, and the motivational features (Ryan & Deci, 2000), that influence learning in general. These three studies all cover economic and business-related games and the target groups are high school and higher education learners (aged 16–20 years). The main research question per study is: How effective is serious gaming in comparison with classroom instruction and why? By measuring the learning outcome as well as the learning process, we were able to study the "black box" of learning. In this way, the results do not only provide insight into the effectiveness per single game but also into the general learning features that positively contribute to learning effectiveness. The research method as such was also evaluated and provides interesting learning points for future studies on serious gaming.

General method

Study design

Table 1 provides an overview of the three experimental studies at the three different schools; each study, including a description of the games, will be explained in more detail in the next section.

Table 1 shows that the content of the three games is somewhat related, but that the games differ in their target group, type of game and duration. Unfortunately, in study 1, the knowledge test could not take place due to practical restrictions in the school. The number of students in the control group (classroom instruction) and experimental group (gaming) varied across the studies.

In all three studies, we used the same experimental design (see Figure 1). The game group was compared with a more traditional (lecture-based) classroom instruction. Learning outcomes were measured by comparing the learning gains (learning outcomes in posttest minus prior knowledge in pretest). We measured the learning process by comparing the learning features of the games (experimental group) with those of the traditional classroom instruction methods

Table 1: Characteristics of the studies

	Study 1: "T-Challenge"	*Study 2: "Ease-it"*	*Study 3: "Currency Exchange"*
Topic	Management skills: entrepreneurship	Business engineering: process management	Economics: currency exchange
Education	Higher education: commercial economy and small business	Higher education: commercial economy and small business	High school: pre-university end years
Type of game	Combination of paper-based game and online game model	Combination of paper-based game and online game model	Paper-based game
Duration of game play	9 play rounds of 1.5 hours	1 afternoon (3 hours)	1.5 hours
Learning outcomes	Self-report	Self-report Knowledge test	Self-report Knowledge test
n	Control group $n = 22$ Game group $n = 88$	Control group $n = 22$ Game group $n = 84$	Control group $n = 41$ Game group $n = 46$

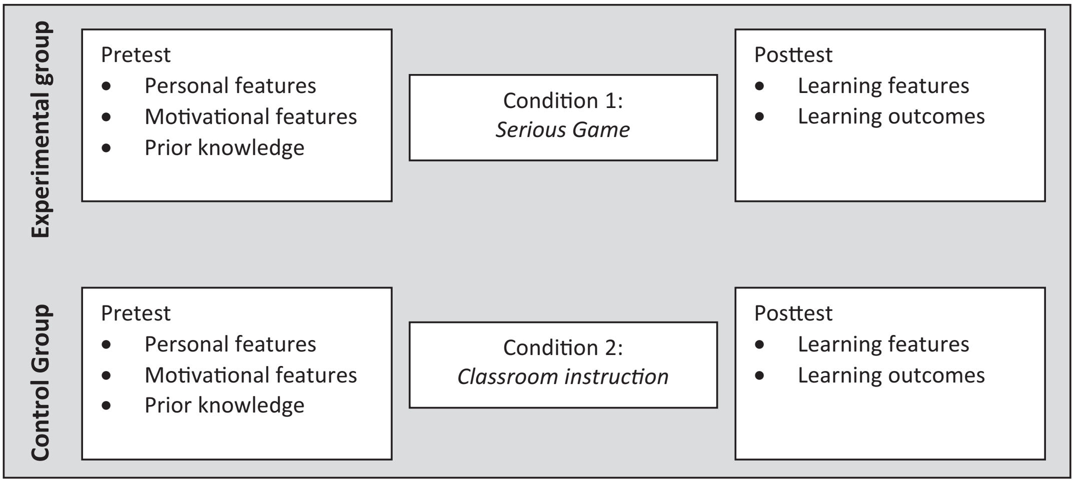

Figure 1: Experimental design

(control group). Personal features (gender, age, class, experience with gaming), measures in a pretest questionnaire, were used as control variables.

Measures

Learning outcomes were measured using a self-report on competences, which gives an impression of the self-efficacy of the student on the specific competences, and a knowledge test (except for study 1). The content of these measures was different for the three studies, yet all used the same format for the self-report. Learners were asked to rate their own general level of competence regarding to the main learning goal on a scale from 1 to 10 (general self-assessment rate): *Rate yourself on a scale from 1 to 10 on how good you are in this topic*. In addition, learners were asked to assess themselves on three to four content-related competences, with each competence consisting of five to seven behavioral indicators (Oprins, 2008). A 5-point rating scale (1 = *poor*, 5 = *good*) was used to rate each indicator. In each of the studies, the question was formulated in a similar way*: Rate yourself on the following competencies and be honest to yourself at the tick of your answers. There are no right or wrong answers. Examples of competences are provided in the study-specific section*. In each of the studies, the self-reports for the pretest and the posttest were similar. The content of the knowledge test for the pretest and the posttest was different; generally, the posttest was more difficult as common in educational settings. The specific measures are elaborated upon in the study-specific section.

The learning process was measured using a posttest questionnaire consisting of five scales on learning features. The choices for the scales were based on the possibility to apply them not only for the game group but also for the classroom instruction group, and its relevancy for the three games we used in our studies. As a result, the learning features as described in the introduction were used with minor adjustments. For instance, "control" was translated more broadly in the sense of self-directedness in the classroom. Also, from "rules and goals," the rules aspect was left out as it was difficult to translate to the classroom instruction group. Instead, challenge and clear goals were put together into one scale. Consequently, the following scales were used: (1) control/self-directedness, (2) challenge/clear goals, (3) feedback, (4) engagement and (5) social interaction.

The questionnaires for both the pretest and posttests were identical for both the serious gaming condition and the classroom instruction condition, except for the word "lesson" or "game." This was also the reason why some features specifically for gaming, argued above, were formulated

Table 2: Reliability of scales, studies 1, 2 and 3 (items translated from Dutch)

	Cronbach's alpha	n	Number of items	Example of items
Control	.82	238	6	I could learn at my own pace.
Challenge	.69	220	9	The content of the game/lesson was too difficult for me.
Feedback	.85	237	8	I always knew how I could improve during the game/lesson.
Engagement	.90	241	7	I felt actively involved in the game/lesson.
Social interaction	.85	240	7	I have learned from the feedback from fellow students.

slightly more generic and applicable also for classroom instruction. All scales consisted of several items (see Table 2) rated on a 5-point scale (1 = *totally disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree*, 5 = *totally agree*). Reliability analyses are performed on combined data from the three studies together. Table 2 shows the Cronbach's alpha's for each scale. According to Nunnally (1978) and Schmitt (1996), the reliability of all scales was good (Cronbach's alpha >.8), except for the "Challenge" scale, which is considered to be acceptable (Clark & Watson, 1995). Therefore, all scales are included in the dataset for further analyses.

Data analyses

Pearson correlation coefficients were calculated for the relationships between the self-report scales and the motivational features. Independent sample *t*-tests were performed to investigate differences on motivational features and learning features in relation to condition. In addition, paired sample *t*-tests were done to investigate the progress on the self-report scales in both conditions.

Next, for all learning outcome variables (the competence measures and the self-assessment rating), a new delta variable (progression on outcome) was calculated by subtracting the mean posttest score per student minus the mean pretest score per student. This variable reflects the progression (or deterioration) on the learning outcomes of a student during the course. Independent sample *t*-tests were performed with the progression-on-outcome variables as dependent variables and condition (experimental group vs. control group) as independent variable.

Finally, to study the influence of the learning features on the progression on the outcome variables, multivariate linear regression analysis was used. The "progression-on-outcome" variables that showed a significant effect ($p < .05$) based on condition were used as dependent variables. For each dependent variable, a separate mediation analysis (Baron & Kenny, 1986) was performed by means of a stepwise multivariate linear regression analyses. In the first step of the analysis, the condition (experimental group vs. control group) was incorporated in the model. In the second step, the learning features were incorporated in the model.

Study 1: T-challenge

Method

Participants

Eighty-three students participated in this study. The experimental group consists of 11 students who are enrolled in a course on Commercial Economy. The control group consists of 72 students who are enrolled in the course on Small Business & Commercial Economy. In total, 46 students completed the posttest (35 in the control group and 11 in the experimental group). The average age of the students is 20 years and 77% of the students are men (Figure 2).

Procedure

Students in the experimental group played an online business game called T-Challenge in the period November–December 2012. The purpose of the game is to make students experience how

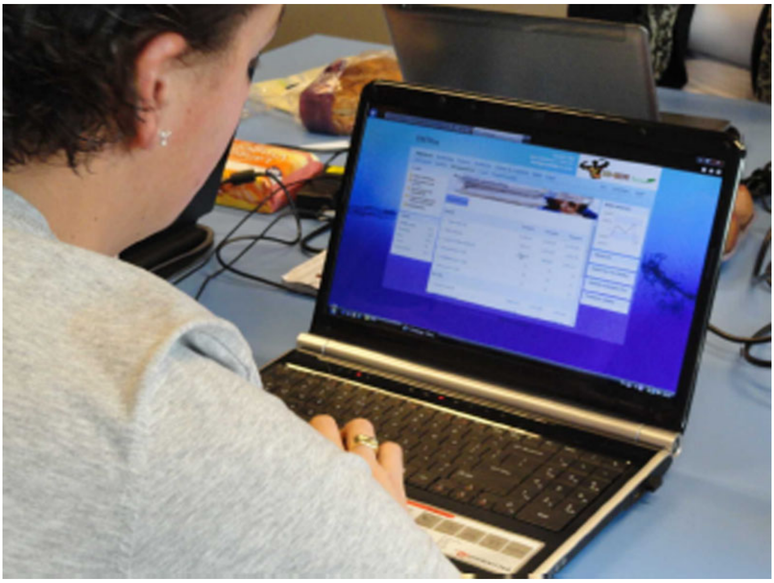

Figure 2: T-challenge

to drive a business, to develop management skills and awareness. By making decisions on (cost) price, salary, new product introduction, promotion budget and personnel, teams must ensure that they build a successful business (a soda factory) and that their business would have better results than other teams. Students played in teams consisting of two or three team members, each of which was accountable for a specific management task. The game was played in nine rounds divided over multiple days. During each round, teams could make tactical decisions in the game. After each round, the effects of changes in the previous round were calculated. The results were shown plenary in a classroom by the teacher by means of a balanced score card. Also, reflection of that round was discussed together with the teacher. During the round, teams had freedom to play the game where they wanted and whenever they wanted. Students in the control group received classroom instruction during the same period of time.

Measures on learning outcomes

Students were asked to assess themselves on four competencies, namely: Strategic Sales & Account Management (SS&AM; consisting of seven indicators), Analysis of Internal and External Environment (AI&EE; consisting of seven indicators), Management Activities (MA; consisting of five indicators), and Design and Implementation (D&I; consisting of five indicators). An example of a behavioral indicator related to the competence SS&AM was "Recognizing strategic, tactical and operational business processes." The students are asked to assess themselves on a 5-point scale (1 = *poor*, 5 = *good*). As stated, these students did not do a knowledge test.

Results

Table 3 presents the results of the paired sample *t*-tests for the differences in means between the pre- and posttest for both the control group and the experimental group. Whereas the control group shows a significant progression on all self-report scales, except for AI&EE, the experimental group shows a significant progression on D&I only. It has to be noted that the values of the means indicate that all students from the experimental group did progress on all self-report scales, except for the general self-assessment rate, but these progressions were not statistically significant. The fact that no significant results are found may be due to the low number of respondents in the experimental group ($n = 11$).

Table 3: Results of the paired sample t*-tests for the differences in means between the pre- and posttest*

T-Challenge	*Learning outcomes: self-report scales*	*Pretest M (SD)*	*Posttest M (SD)*	*t (df)*	*n*
Control group	SS&AM	3.22 (0.57)**	3.55 (0.46)**	3.4 (34)	35
	AI&EE	3.74 (0.46)	3.87 (0.38)	1.56 (34)	35
	MA	3.16 (0.68)***	3.60 ((0.46)***	3.97 (34)	35
	D&I	3.16 (0.71)**	3.63 (0.64)**	3.58 (33)	34
	General assessment rate	5.32 (2.24)**	6.58 (0.77)**	3.24 (30)	31
Experimental group	SS&AM	3.42 (0.49)	3.79 (0.45)	1.52 (10)	11
	AI&EE	3.81 (0.28)	4.17 (0.54)	2.17 (10)	11
	MA	3.58 (0.75)	3.80 (0.30)	1.07 (10)	11
	D&I	3.51 (0.78)*	4.11 (0.28)*	3.08 (10)	11
	General assessment rate	6.78 (0.99)	6.78 (0.84)	0.000 (7)	8

Significant differences between pre- and posttest are marked as follows: ***$p < .001$, **$p < .01$, *$p < .05$. AI&EE, Analysis of Internal and External Environment; D&I, Design and Implementation; MA, Management Activities; SS&AM; Strategic Sales & Account Management.

Table 4: Results of the independent sample t*-tests for the differences in means between the control group and the experimental group*

T-Challenge		*Control group M (SD)*	*Experimental group M (SD)*	*t (df)*
Learning features (posttest)	Control	3.13 (0.49)***	3.82 (0.50)***	–4.02 (44)
	Challenge	3.32 (0.34)**	3.66 (0.31)**	–2.92 (44)
	Feedback	3.16 (0.41)**	3.65 (0.30)**	–3.65 (44)
	Engagement	3.19 (0.55)***	4.12 (0.48)***	–5.01 (44)
	Social interaction	3.48 (0.52)	3.81 (0.60)	–1.75 (44)
Learning outcomes: self-report scales (posttest–pretest)	Δ SS&AM	0.34 (0.59)	0.38 (0.82)	–0.17 (44)
	Δ AI&EE	0.13 (0.48)	0.36 (0.55)	–1.40 (44)
	Δ MA	0.44 (0.66)	0.22 (0.68)	0.97 (44)
	Δ D&I	0.47 (0.77)	0.60 (0.65)	–0.52 (43)
	Δ General assessment rate	1.26 (2.16)	0 (1.51)	1.55 (37)

Significant differences between control group and experimental group are marked as follows: ***$p < .001$, **$p < .01$, *$p < .05$. AI&EE, Analysis of Internal and External Environment; D&I, Design and Implementation; MA, Management Activities; SS&AM; Strategic Sales & Account Management.

Table 4 shows the results of the independent sample *t*-tests that were performed to measure differences between the control group and the experimental group concerning the motivational features at the pretest, the learning features at the posttest, and the progression made on the learning outcomes between the pretest and the posttest. Analyses did not point at significant differences in relation to condition on the level of progression made on the self-report scales. The experimental group scores significantly higher on all learning features, except for social interaction, compared with the control group (Table 4).

As no significant differences were found between the experimental group and the control group in relation to the progression made on the learning outcomes, mediation analyses to study the effect of the learning features on the learning outcomes could not be performed.

Study 2: Ease-it

Method

Participants

One hundred six students in their 2nd year of Commercial Economy and Small Business participated in this study as part of the course Business Engineering. The experimental group consisted

Figure 3: Ease-it

of 84 students and the control group with students from a parallel class consisted of 22 students. The average age of the students is 20 years and 66% of them are men (Figure 3).

Procedure

Students in the experimental group played a simulation business game called Ease-it. The purpose of the game is to provide students insight into principles of process management. Students have to come up with ideas for organizational changes from a process management perspective. The (business) simulation game takes place in the context of a major financial institution (banking and medical insurances) with several departments. Students have to handle claims or applications for funding of (internal) customers correctly and timely. Everyone has their own role in the team, and every team (consisting of approximately 13 participants) starts off with a suboptimal/not very efficient process of handling the claims. They experience this suboptimal process themselves and can reorganize their own business in between rounds. The experimental group played the game during an afternoon within 3.5 hours, in three rounds. In between these rounds, when they came up with ideas how to organize their processes differently, reflective discussions also took place. Students in the control group received traditional education in the same period of time. The game was played in September 2012.

Measures on learning outcomes

The students' level of knowledge regarding the subject was measured in two ways, based on self-reports and based on knowledge tests. Students were asked to rate their own general level of knowledge regarding the topic on a scale from 1 to 10. In addition, students were asked to assess themselves on a 5-point scale (1 = *poor*, 5 = *good*) on three competencies, namely: Strategic Management and Organization (SM&O; consisting of seven indicators), Analysis of Internal Environment (AIE; consisting of seven indicators), and D&I (consisting of five indicators). An example of a behavioral indicator related to the competence SM&O was "Recognizing primary, secondary and administrative business processes." In addition, students were asked to complete a knowledge test before and after the course. Students received five questions for which they could choose between four multiple choice answers. An example of an item is: "When I note that the operation is a loss, I will first . . ." and the multiple choice answers are "reduce costs," "encourage employees to work harder," "ask everyone individually what they think is the cause of the problems" and "seclude myself to think about it." The knowledge test on the pretest was different from the one on the posttest. For each question, a new dichotomous variable was computed, making a distinction between a "correct answer" and a "wrong answer." For each student, the number of correct answers was calculated on the pretest as well as on the posttest.

Results

Table 5 shows the results of the paired sample *t*-tests that were performed to measure the progression on learning outcomes for both the control group and the experimental group. Whereas the control group shows a significant progression on the general self-assessment rate only, the experimental group shows a significant progression on all self-report scales and on the knowledge test.

Table 6 shows the results of the independent sample *t*-tests that were performed to measure differences between the control group and the experimental group concerning the motivational features at the pretest, the learning features at the posttest, and the progression made on the

Table 5: Results of the paired sample t*-tests for the differences in means between the pre- and posttest*

Ease-It	*Learning outcomes*	*Pretest* *M (SD)*	*Posttest* *M (SD)*	*t (df)*	*n*
Control group	SM&O	3.16 (0.59)	3.03 (0.58)	−0.98 (21)	22
	AIE	3.31 (0.49)	3.14 (0.66)	−1.42 (19)	20
	D&I	3.14 (0.81)	2.97 (0.72)	−1.47 (19)	20
	General self-assessment rate	6.14 (0.77)**	5.05 (1.36)**	−3.55 (20)	21
	Knowledge test	2.32 (1.17)	2.55 (1.06)	0.68 (21)	22
Experimental group	SM&O	3.24 (0.59)***	3.78 (0.42)***	7.72 (75)	76
	AIE	3.33 (0.54)***	3.84 (0.47)***	7.04 (73)	74
	D&I	3.18 (0.65)***	3.77 (0.46)***	7.89 (72)	73
	General self-assessment rate	5.81 (1.15)***	7.05 (0.96)***	−8.94 (72)	73
	Knowledge test	2.08 (0.95)**	2.69 (1.21)**	3.13 (73)	74

Significant differences between pretest and posttest are marked as follows: ***$p < .001$, **$p < .01$, *$p < .05$. D&I, Design and Implementation; AIE, Analysis of Internal Environment; SM&O, Strategic Management and Organization.

Table 6: Results of the independent sample t*-tests for the differences in means between the control group and the experimental group*

Ease-It		*Control group* *M (SD)*	*Experimental group* *M (SD)*	*t (df)*
Learning characteristics (posttest)	Control	2.11 (0.73)***	3.60 (0.53)***	−8.90 (27.30)
	Challenge	2.36 (0.55)***	3.48 (0.37)***	−9.00 (26.39)
	Feedback	2.32 (0.59)***	3.75 (0.42)***	−10.68 (27.02)
	Engagement	2.22 (0.74)***	4.13 (0.51)***	−11.43 (26.68)
	Social interaction	3.06 (0.60)***	4.01 (0.48)***	−7.83 (100)
Learning outcomes (posttest–pretest)	Δ SM&O	−0.14 (0.66)***	0.54 (0.61)***	−4.50 (96)
	Δ AIE	−0.19 (0.61)***	0.49 (0.60)***	−4.51 (92)
	Δ D&I	−0.24 (0.73)***	0.59 (0.64)***	−4.99 (91)
	Δ General self-assessment rate	−1.19 (1.53)***	1.19 (1.14)***	−6.60 (26.63)
	Δ Knowledge test	0.23 (1.57)	0.81 (1.67)	−0.95 (94)

Significant differences between control group and experimental group are marked as follows: ***$p < .001$, **$p < .01$, *$p < .05$. D&I, Design and Implementation; AIE, Analysis of Internal Environment; SM&O, Strategic Management and Organization.

learning outcomes between the pretest and the posttest. The experimental group shows significantly more progression than the control group on all of the self-report scales (SM&O, AIE, D&I and the general assessment rate). Differences between conditions regarding the progression on the knowledge test are not statistically significant. In addition, the experimental group scores significantly higher on all learning features compared with the control group.

To study whether differences between the experimental group and the control group in relation to the progression made on the self-reported learning outcomes can be explained by the learning features, mediation analyses were performed. The first mediation analysis shows that the variance of the progression score on SM&O, which is explained by condition in step 1 ($B = 0.78$, $p = .000$), is no longer significant in step 2 ($B = 0.26$, $p = .34$) when the learning features are included in the analysis. Learning features that significantly contribute to the progression on SM&O are engagement ($B = –0.26$, $p = .09$), feedback ($B = 0.37$, $p = .07$) and challenge ($B = 0.39$, $p = .07$). In other words, the differences between the experimental and control groups in terms of the progression SM&O are somewhat explained by feedback, engagement and challenge.

The second mediation analysis shows that the variance of the progression score on AIE, which is explained by condition in step 1 ($B = 0.73$, $p = .000$), is no longer significant in step 2 ($B = 0.02$, $p = .93$) when the learning features are included in the analysis. The learning feature that significantly contribute to the progression on AIE is challenge ($B = 0.41$, $p = .05$). In other words, the differences between the experimental and control groups in terms of the progression of AIE are somewhat explained by challenge.

The third mediation analysis shows that the variance of the progression score on D&I, which is explained by condition in step 1 ($B = 0.88$, $p = .000$), is no longer significant in step 2 ($B = 0.24$, $p = .36$) when the learning features are included in the analysis. Learning features that significantly contribute to the progression on D&I are feedback ($B = 0.42$, $p = .06$) and social interaction ($B = 0.31$, $p = .04$). In other words, the differences between the experimental and control groups in terms of the progression of D&I are somewhat explained by feedback and social interaction.

The fourth mediation analysis shows that the variance of the progression score on the general self-assessment rate, which is explained by condition in step 1 ($B = 2.29$, $p = .000$), is no longer significant in step 2 ($B = 0.16$, $p = .71$) when the learning features are included in the analysis. Learning features that significantly contribute to the general self-assessment rate are feedback ($B = 0.57$, $p = .09$) and challenge ($B = 0.69$, $p = .05$). In other words, the differences between the experimental and control groups in terms of the general self-assessment rate are to a certain extent explained by feedback and challenge.

Study 3: Currency exchange

Method

Participants

The third study was conducted at the Sondervick College, a high school in Veldhoven. Five classes participated in the study, three classes of students in the 5th year of preuniversity education and four classes in the 5th year of higher general secondary education. Students were randomly assigned to either the control group or the experimental group. In total, 41 students participated in the control group and 46 in the experimental group. In the study, 58% of the students are men and the average age is 16 years (Figure 4).

Procedure

Students in the experimental group played the board game called Currency Exchange. The objective of the game is to help students learn about trade and dealing with money exchange. Students play the game in teams of two or three. There are multiple rounds in which team members trade

Figure 4: Currency exchange

goods (for example, cars) with each other. After every round, players decide what the effect of their actions is on the balance of payments and consequently the currency exchange rate of their own currency. The player with the smartest balance between trade and currency exchange has the largest chance of winning the game. The experimental group played the game for 1.5 hours. The control group was taught the same subject matter in a regular classical lesson, in the same period of time. This study was conducted in the period October–November 2012.

Measures on learning outcomes

The students' level of knowledge regarding the subject was measured in two ways, by means of self-reports and knowledge tests. Students were asked to rate their own general level of knowledge regarding to the topic on a scale from 1 to 10. In addition, students were asked to assess themselves on three competencies, namely: Information Literacy (IL; consisting of five indicators), Strategic Thinking (ST; consisting of six indicators) and Cyclical Phenomena (CP; consisting of five indicators). Students are asked to assess themselves on a 5-point scale (1 = *poor*, 5 = *good*). An example of a behavioral indicator related to the competence IL was "Retrieve required information from sources such as text, tables, charts and graphics." In addition, students were asked to complete a knowledge test on the pretest as well as the posttest. Both tests consisted of open questions made and checked by the instructor. An example of an item is: "All international transactions of a country are kept on the balance of payments of the country. Explain how the exchange rate of a country may rise by a balance of payments surplus." The maximum score on the pretest was 12 points and the maximum score on the posttest was 14 points. A new pretest and posttest variable was introduced that expressed the percentage of correct answers. Unfortunately, the difficulty of the pretest and posttest was not similar: the posttest was more difficult. In educational settings, this is quite normal because students are supposed to learn over time, but for experimental purposes this is not ideal as it makes it difficult to determine the net learning effects. We could not change this because the experiment took place in a practical educational setting.

Results

Table 7 shows the results of the paired sample *t*-tests that were performed to measure the progression on learning outcomes for both the control group and the experimental group. Whereas

Table 7: Results of the paired sample t*-tests for the differences in means between the pre- and posttest*

Currency exchange	*Learning outcomes*	*Pretest M (SD)*	*Posttest M (SD)*	*t (df)*	*n*
Control group	IL	2.92 (0.44)	2.87 (0.31)	−0.83 (37)	38
	ST	2.91 (0.38)	2.88 (0.38)	−0.55 (37)	38
	CP	2.39 (0.56)***	2.79 (0.57)***	4.01 (37)	38
	General self-assessment rate	5.17 (1.50)*	5.75 (1.57)*	2.17 (35)	36
	Knowledge test	0.43 (0.19)	0.33 (0.19)	−1.95 (35)	36
Experimental group	IL	3.10 (0.41)	3.11 (0.41)	0.33 (45)	46
	ST	3.00 (0.40)**	3.1 (3.95)**	2.92 (44)	45
	CP	2.64 (0.59)***	3.16 (0.54)***	5.96 (44)	45
	General self-assessment rate	6.13 (1.33)***	6.98 (1.22)***	5.55 (44)	45
	Knowledge test	0.50 (0.17)**	0.41 (0.21)**	−3.46 (44)	45

Significant differences between pretest and posttest are marked as follows: ***$p < .001$, **$p < .01$, *$p < .05$. CP, Cyclical Phenomena; IL, Information Literacy; ST, Strategic Thinking.

Table 8: Results of the independent sample t*-tests for the differences in means between the control group and the experimental group*

Currency exchange		*Control M (SD)*	*Experimental group M (SD)*	*t (df)*
Motivational features	Self-efficacy	3.50 (0.54)**	3.83 (0.51)**	−2.98 (85)
	Intrinsic motivation	3.87 (0.49)	3.87 (0.50)	−0.01 (85)
Learning features	Control	3.02 (0.53)***	3.60 (0.51)***	−5.09 (82)
	Challenge	3.46 (0.35)	3.49 (0.34)	−42 (82)
	Feedback	3.10 (0.50)***	3.70 (0.45)***	−5.83 (82)
	Flow	3.58 (0.43)***	4.02 (0.41)***	−4.71 (82)
	Social interaction	3.24 (0.74)***	3.83 (0.44)***	−4.57 (82)
Learning outcomes	Δ IL	−0.45 (0.33)	0.01 (0.26)	−0.88 (82)
	Δ ST	−0.03 (0.31)*	0.11 (0.25)*	2.20 (81)
	Δ CP	0.40 (0.62)	0.52 (0.58)	−87 (81)
	Δ General self-assessment rate	0.58 (1.61)	0.84 (1.02)	−0.85 (56.51)
	Δ Knowledge test	−0.10 (0.30)	−0.09 (0.18)	−0.09 (79)

Significant differences between control group and experimental group are marked as follows: ***$p < .001$, **$p < .01$, *$p < .05$. CP, Cyclical Phenomena; IL, Information Literacy; ST, Strategic Thinking.

the control group shows a significant progression on CP and the general self-assessment rate, the experimental group, in addition, shows significant progression on ST. Contrary to our expectations, in the experimental group the posttest scores on the knowledge test are significantly lower than the pretest scores.

Table 8 shows the results of the independent sample *t*-tests that were performed to measure differences between the control group and the experimental group concerning the motivational features at the pretest, the learning features at the posttest, and the progression made on the learning outcomes between the pretest and the posttest. Regarding the progression on the learning outcomes, only one significant difference in relation to condition was found. The experimental group showed more progression than the control group on the self-report scale ST. The experimental group scores significantly higher on all learning features, except for challenge, when compared with the control group. In addition, the experimental group scores significantly higher on self-efficacy than the control group.

To study whether differences between the experimental group and the control group in relation to the progression made on the learning outcome ST can be explained by the learning features, a mediation analysis was performed. The mediation analysis shows that the variance of the progression-on-outcome score on ST, which is explained by condition in step 1 (B = 0.24, *p* = .031), is no longer significant in step 2 (B = 0.19, *p* = .14) when the learning features are included in the analysis. The learning feature that significantly contribute to the progression on ST is engagement (B = 0.46, *p* = .000). In other words, the differences between the experimental and control groups in terms of the progression of ST are to a certain extent explained by engagement.

Discussion and conclusion

The aim of the three empirical studies described in this paper was to obtain evidence on the effectiveness of serious games. A fixed research methodology was used and the same set of learning features that is expected to be enhanced with serious gaming was measured. These learning features are the so-called generic "process variables" that were measured in addition to the more frequently used domain-specific "outcome variables" (Alvarez *et al*, 2004; Kraiger *et al*, 1993; Tannenbaum *et al*, 1993). In the experimental design that we chose in the three studies, we measured both types of variables and we related the process variables, cf. learning features, to the learning outcome variables (self-report, knowledge test) in a mediation analysis. The aim was to get insight into the "black box" of learning with serious games based on this set of learning features that should explain why certain learning outcomes were achieved.

The most important conclusions we can draw based on the results of the three studies (Table 9) are that: (1) serious gaming has resulted in more positive effects as compared with classroom instruction, when it comes to *self-reported* learning outcomes, or in other words the students' own feelings of competence; and (2) the features that are known to contribute to learning are clearly in favor of serious gaming compared with classroom instruction. However, positive effects of serious gaming on the (not self-reported) knowledge tests could not be proofed. Explanations for these conclusions and implications for future work are discussed below.

Learning outcomes

The measurement of the learning outcomes consisted of two types of measures: a self-assessment on competences and a knowledge test (except for study 1). In two out of three studies, students who played the game showed significantly more progression on (at least one of) the self-assessed

Table 9: Summary of results of the three studies

	Study 1: T-challenge	*Study 2: Ease-it*	*Study 3: Currency exchange*
1A. Self Assessment *Experimental group: Progression T0–T1)*	+	+	+
1B.Self Assessment *(Experimental group > control group)*	+	+	+
2A. Performance measures *Experimental group: ProgressionT0-T1)*	Not measured	0	–
2B. Performance measures *(Experimental group > control group)*	Not measured	0	0
Learning features *(Experimental group > control group)*	+	+	+
Effect of learning features on learning outcomes	Not measured	Feedback, Engagement, Challenge, Social interaction	Engagement

competences. In the first study, no significant results were found, possibly because of the low number of respondents in the experimental group, but the results pointed at a similar direction. The results from the self-assessments indicate that students who played the game *feel* that they have learned, even more than the students who received classroom instruction. Thus, these higher results on self-assessment might indicate that the effectiveness of learning for serious gaming is higher than for classroom instruction. However, we should take into account that in general, self-assessment is less reliable than objectively measured performance as this is based on own judgments. Self-assessment points at own feelings of competence (Pintrich & de Groot, 1990) and is strongly related to self-efficacy (Bandura, 1997). We may conclude that the students' self-efficacy may be increased by serious gaming. This is consistent with findings of a meta-analysis by Sitzmann (2011) and could be explained by the fact that the students can practice by themselves, in a learning environment that allows self-direction (Pintrich & de Groot, 1990).

Results of gaming on the knowledge test are somewhat inconclusive. Students that played the game "Ease-It" (Study 2) showed a significant progression on the knowledge test. However, this progression could not statistically be proven as higher compared with the classroom instruction group. With "Currency Exchange" (Study 3), students from both conditions deteriorated on the knowledge test. For the students that played the game, this deterioration was even statistically significant.

There are several explanations for the fact that students that played the game did not show better results on the knowledge test. First, the learning gains of serious games are often found in an improved understanding and improved attitude. At the schools, the regular knowledge tests they use consist of a number of multiple choice questions. It can be discussed if this was the best way to measure learning gains. Games are assumed to improve competences because they are more open-ended and self-directed with higher control over own learning processes (eg, Bedwell *et al*, 2012), thus also the test should be more competence-based (Oprins, 2008). In other words, there seems to be a mismatch between the type of learning objectives and type of assessment.

Second, the knowledge test used for the posttest of study 3 is deemed more difficult compared with the pretest used in that study. This is normal in education as students learn over time and could do more difficult tests, but this makes measuring progression over time difficult (Oprins, 2008). In sum, the limitations in the research design concerning the knowledge tests (lack of the knowledge test in study 1; shortcomings of the knowledge tests used in studies 2 and 3) make it difficult to draw conclusions on the effect of serious gaming on more objective learning outcomes. Additional research is needed to further explore these effects.

Learning process

The fact that the learning outcomes did not show clear results concerning the effectiveness of the games, stresses the necessity to also look at the underlying learning process. The comparison between the experimental (gaming) condition with the control (classroom instruction) condition in all three studies resulted in a rather consistent view on the learning features. In general, the students who played the game scored clearly higher on the learning features compared with students who received traditional classroom instruction. The students who played the three different games did have the feeling of more control over their learning process, they received relevant feedback and they felt actively engaged. Students also reported to benefit from social interaction (in two out of three games) and from challenging and useful content (in two out of three serious games). These results are in line with the literature, as mentioned in the introduction section, concerning the gaming elements that are assumed to enhance learning (eg, Bedwell *et al*, 2012; Garris *et al*, 2002; Gee, 2005; Kiili, 2005; Koster, 2005; Malone, 1981; Pavlas, 2010; Sweetser & Wyeth, 2005).

Looking at these results and keeping in mind the theories of how people learn best, it seems that learning through serious gaming appears to be very qualitative. This study shows that serious gaming has a more positive effect on learning features (being indicative for a high-quality learning process) than on classroom instruction.

Investigating the relationship between learning features and learning outcomes provided us with interesting insights. The mediation analyses showed that differences on the self-assessments on competences in relation to condition can partly be explained by some of these learning features. This suggests that gaming has a positive influence on the learning features that contribute to the progression on self-efficacy. In this sense, we can cautiously conclude that this study contributes to the clarification of the mechanism through which serious games stimulate learning based on self-reported measures on learning outcomes. That is, serious games allow learners to learn actively and self-directed in a social, engaging and challenging environment with relevant feedback (see also eg, Gee, 2007; Percival, 1996; Squire, 2003, etc). We did not find clear differences between the various learning features. This implies that none of them was proven more important than others for specific games. In future studies, it would be worthwhile to study the effect of different learning features more closely.

Implications for future research

In sum, the results of the three studies have shown that it has a surplus value to measure learning features in combination with learning outcomes. The fact that the results in the three studies concerning the learning features were so consistent, emphasizes the generalizability of these measures with learning features over the three studies (eg, Mayer, 2012). It showed the value of measuring process variables besides outcome variables as this can explain why certain learning outcomes are achieved (Alvarez *et al*, 2004; Kraiger *et al*, 1993; Tannenbaum *et al*, 1993). At long term, measuring the same generic set of learning features over more games would provide more insight into the design of games that results in optimal learning (Mayer, 2012). This could improve serious gaming design in a broad sense.

In future studies, we would like to make a further improvement in the process variables, that is, going one step “deeper” in the black box, and measure the design mechanics of serious games in more detail. In the current three studies, a generic set of learning features was used that was applicable to the serious game condition as well as the classroom instruction. This was very effective, as it made it possible to compare the two methods concerning their didactical quality. Yet, information that game designers could use in their design process is still lacking. Therefore, it would be interesting to investigate the game design features as well. An example is adding a feature like “game fiction” (Bedwell *et al*, 2012), referring to the game story and game world. Another example is to look in more detail how a feature like “feedback” is designed in the game, instead of only asking how it is experienced by players.

In other words, it would be interesting to measure the effect of specific gaming design features on more general learning features, for instance, by manipulating some gaming mechanisms for one particular game and investigating how this would affect the learning process itself. In a next study (E. Oprins *et al*, unpublished data), we developed a generic evaluation framework for serious gaming and we did a first validation study. This evaluation framework is a further elaboration of the method used in the current study, described in this paper. In this evaluation framework, the gaming design features (eg, game fiction, feedback design methods) and more general learning features (eg, self-directed learning, engagement, intrinsic motivation) comprise the process variables from the idea that the design features of the games directly affect the learning process. The general learning features can be measured in any form of learning including classroom instruction while the gaming design features can be measured over multiple games. Future research, based on this evaluation framework, is focused on the relationship between learning outcomes—

learning features—*and* serious game design features. Then we are able to truly help designers improve their serious games, as they can rely on proven effects (Bedwell *et al*, 2012; Mayer, 2012; Oprins & Korteling, unpublished data). This would really open the black box of learning with serious gaming.

Acknowledgements

The authors would like to thank all the students and staff of the schools for their participation in the experiment. They also want to thank Simagine, Be Involved and TRIQS for making their games available for research purposes.

Statements on open data, ethics and conflict of interest

Data that are collected and analyzed for this study can be accessed by others to check whether results can be replicated.

All students in the study were treated within an ethic of respect. All teachers informed their students about the study. All data gathered in the study were anonymized before the analyses. Students from the control group played the game after the experiment and students from the experimental group received the classroom instruction to make sure that all students received a similar quality of education.

There were no potential conflicts of interest in this study.

References

Alvarez, K., Salas, E. & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review*, *3*, 4, 385–416.

Bandura, A. (1997). *Self-efficacy: the exercise of control*. New York: Freeman.

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Bedwell, W. L., Pavlas, D., Heyne, K., Lazzara, E. H. & Salas, E. (2012). Toward a taxonomy linking game attributes to learning: an empirical study. *Simulation and Gaming*, *43*, 6, 729–760.

Clark, L. A. & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, *7*, 3, 309–319.

Connolly, T. M., Boyle, E. A., MacArthur, W., Hainey, T. & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, *59*, 2012, 661–686.

Csikszentmihalyi, M. (1990). *Flow: the psychology of optimal experience*. New York: Harper & Row.

Garris, R., Ahlers, R. & Driskell, J. E. (2002). Games, motivation, and learning: a research and practice model. *Simulation & Gaming*, *33*, 4, 441–467.

Gee, J. P. (2005). Learning by design: good video games as learning machines. *E-learning*, *2*, 1, 5–16.

Gee, J. P. (2007). *Good video games and good learning: collected essays on video games, learning, and literacy*. New York: Peter Lang.

Gee, J. P. (2009). Deep learning properties of good digital games. How far can they go? In U. Ritterfeld, M. Cody & P. Vorderer (Eds), *Serious games: mechanisms and effects* (pp. 68–82). New York/London: Routledge.

Harteveld, C. (2012). *Making sense of virtual risks: a quasi-experimental investigation into game-based training*. Amsterdam: IOS Press.

Hays, R. T. (2005). *The effectiveness of instructional games: a literature review and discussion* Technical Report 2005-004. Orlando, FL: Naval Air Warfare Training Systems Division.

Kiili, K. (2005). Digital game-based learning: towards an experiential gaming model. *The Internet and Higher Education*, *8*, 1, 13–24.

Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: a guide to human resource development*. New York: McGraw Hill.

Kirkpatrick, D. L. (1994). *Evaluating training programs: the four levels*. San Francisco: Berrett-Koehler.

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

Kraiger, K., Ford, J. K. & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, *78*, 2, 311–328.

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

Mayer, I. (2012). Towards a comprehensive methodology for the research and evaluation of serious games. *Procedia Computer Science*, *15*, 233–247.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Oprins, E. (2008). Design of a competence-based assessment system for air traffic control training (Doctoral dissertation, Maastricht University).

Pavlas, D. (2010). A model of flow and play in game-based learning: the impact of game characteristics, player traits, and player states (Doctoral dissertation, University of Central Florida, Orlando).

Percival, A. (1996). Invited reaction: an adult educator responds. *Human Resource Development Quarterly*, *7*, 131–139.

Pintrich, P. R. & de Groot, W. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*, 1, 33–40.

Ryan, R. M. & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, 55, 1, 68–78.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 4, 350–352.

Shaffer, D. W. (2006). *How computer games help people learn*. New York: Palgrave Macmillan.

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

Squire, K. (2003). Video games in education. *International Journal of Intelligent Simulations and Gaming*, *2*, 1, 49–62.

Squire, K. (2006). From content to context: videogames as designed experience. *Educational Researcher*, *35*, 8, 19–29.

Squire, K. D. (2011). *Video games and learning: teaching and participatory culture in the digital age*. New York: Teachers College Press.

Stubbé, H. M. & Theunissen, N. C. M. (2008). Self-directed learning in a ubiquitous learning environment: a meta-review. *Proceedings of Special Track on Technology Support for Self-Organised Learners*, *2008*, 5–28.

Sweetser, P. & Wyeth, P. (2005). GameFlow: a model for evaluating player enjoyment in games. *ACM Computers in Entertainment*, *3*, 3, 1–24.

Tannenbaum, S., Cannon-Bowers, J., Salas, E. & Mathieu, J. (1993). *Factors that influence training effectiveness: a conceptual model and longitudinal analysis (Tech. Rep. No. 93-011)*. Orlando, FL: Naval Training Systems Center, Human Systems Integration Division.