The research and evaluation of serious games: Toward a comprehensive methodology

Igor Mayer, Geertje Bekebrede, Casper Harteveld, Harald Warmelink, Qiqi Zhou, Theo van Ruijven, Julia Lo, Rens Kortmann and Ivo Wenzler

Igor Mayer is an associate professor in Public Administration and Gaming in the Faculty of Technology, Policy and Management, and director of the Serious Gaming Research Center, TU Delft, The Netherlands. Geertje Bekebrede is an assistant professor in Policy, Organization and Gaming in the Faculty of Technology, Policy and Management, TU Delft, and a senior game designer at Tytron Serious gaming, The Netherlands. Casper Harteveld is an assistant professor of Art and Design at Northeastern University. Harald Warmelink is an assistant professor in Policy, Organization and Gaming in the Faculty of Technology, Policy and Management, TU Delft. Qiqi Zhou is an assistant professor in Policy, Organization and Gaming in the Faculty of Technology, Policy and Management, TU Delft. Theo van Ruijven is an assistant professor in Policy, Organization and Gaming in the Faculty of Technology, Policy and Management, TU Delft. Julia Lo is a PhD researcher in Policy, Organization and Gaming in the Faculty of Technology, Policy and Management, TU Delft. Rens Kortmann is an assistant professor in Systems Engineering in the Faculty of Technology, Policy and Management, TU Delft. Ivo Wenzler is an associate professor in Policy, Organization and Gaming in the Faculty of Technology, Policy and Management, TU Delft, and an associate partner at Accenture. Address for correspondence: Dr Igor Mayer, Faculty of Technology, Policy and Management, TU Delft, PO Box 5015, 2600 GA, Delft, The Netherlands. Email: i.s.mayer@tudelft.nl

Abstract

The authors present the methodological background to and underlying research design of an ongoing research project on the scientific evaluation of serious games and/or computer-based simulation games (SGs) for advanced learning. The main research questions are: (1) what are the requirements and design principles for a comprehensive social scientific methodology for the evaluation of SGs?; (2) to what extent do SGs contribute to advanced learning?; (3) what factors contribute to or determine this learning?; and (4) to what extent and under what conditions can SG-based learning be transferred to the real world? In the Netherlands between 2005 and 2012, several hundred SG sessions with 12 SGs were evaluated systematically, uniformly and quantitatively to create a dataset, which comprises data on 2488 respondents in higher education or work organizations. The authors present the research model, the quasi-experimental design and the evaluation instruments. This focus in this paper is on the methodology and dataset, which form a sound foundation for forthcoming publications on the empirical results.

Introduction

The growing interest in (digital) Game-Based Learning (GBL), Serious Games and Simulation-Gaming (both abbreviated as SG)\(^1\) entails a growing need to know the effects of what we are doing and promoting. To meet this need, we require proper methods, tools and principles that the fragmented communities agree upon, can validate and apply; in other words, we must take a step

\(^1\)The oxymoron Serious Games was first used as title of a book by Clark Abt (1970); the simulation-gaming discipline spurred a.o. by Duke’s (1974) *Gaming: the Future’s Language*. There’s not much light between Abt’s and Duke’s view on gaming, although later serious games mainly refer to digital nonentertainment games (as artifact). In the context of this research, we will not make a semantic or taxonomic distinction between serious games and simulation games and will use the abbreviation SG for both.
• geographical fragmentation: aligning for instance countries and geographical markets in the Europe or USA.
• disciplinary fragmentation: aligning disciplines like engineering, humanities and social sciences.
• institutional fragmentation: aligning for instance different research institutes; and
• business chains and networks: aligning for instance producers, consumers, financers, sponsors, publishers, stakeholders, etc.

Considerable efforts and resources are being devoted to researching and evaluating GBL and SGs. As a result, both the number and the quality of such evaluations are increasing (see for a discussion below). However, there are still considerable weaknesses, for example:
• a lack of comprehensive, multipurpose frameworks for comparative, longitudinal evaluation (Blunt, 2006; Meyer, 2010; Mortagy & Boghikian-whitby, 2010; Vartiainen, 2000);
• few theories with which to formulate and test hypotheses (R. E. Mayer, 2005; Noy, Raban & Ravid, 2006);
• few operationalized models to examine “causal” relations (eg, in structural equations models) (Connolly, Stansfield & Hainey, 2009; Hainey & Connolly, 2010);
• few validated questionnaires, constructs or scales, either from other fields (eg, psychology) or newly constructed for SG and GBL (Boyle, Connolly & Hainey, 2011; Brockmyer et al, 2009; Mayes & Cotton, 2001);
• a lack of proper research designs, other than randomized controlled trials that can be used in a dynamic, professional learning context (Connolly, Boyle, MacArthur, Hainey & Boyle, 2012; Kato, Cole, Bradlyn & Pollock, 2008; Knight et al, 2010; Szturm, Betker, Moussavi, Desai & Goodman, 2011; van der Spek, 2011; van der Spek, Wouters & Van Oostendorp, 2011); and

Practitioner Notes
What is already known about this topic
• There are a great many evaluation studies on the efficacy of games for learning.
• There is a plethora of approaches and methods.
• There is an increasing number of reviews on evaluation studies.

What this paper adds
• This publications gives requirements for science and methodology of game-based learning (GBL).
• This paper gives conceptual frameworks and models for comparative evaluation.
• This paper gives structure and examples for comprehensive dataset composition.

Implications for practice and/or policy
• The authors make a strong plea and recommendation to look at evaluation of GBL in a more systematic way.
• The paper gives constructs and items to (re-) use in evaluation of GBL.
• an absence of generic tools for unobtrusive (“stealth”) data gathering and assessment in and around SGs (Kickmeier-Rust, Steiner & Albert, 2009; Shute, 2011; Shute, Masduki & Donmez, 2010; Shute, Ventura, Bauer & Zapata-Rivera, 2009).

In short, although there is a promising increase in publications, methods, tools and findings, we lack an overarching methodology for SG research. Aligning a fragmented serious gaming community to evaluate and research gaming for learning in a comparative, systematic fashion using procedures, frameworks and methods that can be validated, checked and reproduced represents an enormous challenge.

This is the wider context of our SG research project (2005–12) on the evaluation of SGs for advanced learning, which distinguishes it from Games in K12 education or Video games in the classroom. The challenge is to gather data on the quality, application and outcomes of a broad range of SGs on different topics and with different objectives, used in and for different institutional contexts, at different moments in time and under uncontrolled conditions. Although it is valuable to find the effects of playing games with students in a lab, it is essential to know the effects of GBL in uncontrolled circumstances and for objectives that truly matter for real-life performance (ie, emergency management and leadership), as the latter is usually the case in professional learning and training.

In search of a methodology

A social–scientific discipline of SGs research would include a critical and reflective discourse on:

1. frames and discourses: the multiple, often conflicting ways in which we perceive and discuss SGs and GBL (Chong & Druckman, 2007; Shaffer, 2006; Squire, 2002);
2. methodology: the rationale and principles on which SGs and GBL research is founded (Mackenzie & Knipe, 2006);
3. research designs and data gathering: what works, why and when? (De Vaus, 2001; Schneider, 2005);
4. validated research instruments and tools: questionnaires, surveys, logging and tracking instruments, including their validation, for SGs and GBL (Boyle et al., 2011; Brockmyer et al., 2009; Chertoff, Goldiez & LaViola, 2010; Mayes & Cotton, 2001; Wright & Marsden, 2010);
5. a dynamic body of knowledge identifying the state of the art and knowledge gaps (Ma, Williams, Prejean & Richard, 2007; Raphael, Bachen, Lynn, Baldwin-Philippi & McKee, 2009; Young et al., 2012); and
6. professional ethics of the SG designer, the SG advocate, the SG seller, the SG interventionist, etc. (Babbie, 2007; Chandler & Torbert, 2003).

This is highly needed because of:

1. accountability: “users” (clients, players and learners) are becoming more exposed to and familiar with SGs. They have the right to know what they are actually buying, using or playing, for what reason and what the effects or consequences of the application of SGs and gamification\(^2\) are. We expect that users will also become more demanding, critical and skeptic; and
2. responsibility: the opposite of accountability. A discipline that advocates the use of SGs and gamification to repair a broken reality (McGonigal, 2011), especially when vulnerable groups in society are involved (eg, children, patients and immigrants) has a great responsibility to critically reflect on the short- and long-term value and structural consequences of the games and gamification tools they are developing, promoting and using.

\(^2\)The use of SG can be part of a larger gamification approach; but not all forms of gamification include the use of SG. Gamification (McGonigal, 2011) refers to the incorporation of one or more principles of “play” (eg, participant feedback to create competition to give engagement to increase performance, etc.) into (organizational, commercial, social and political) reality.
Research objective

The aim of our project is to develop a generic evaluation methodology for serious gaming that comprises a framework, conceptual models, research design(s), data gathering techniques and evaluation constructs and scales. Our ambition is to resolve the dilemma between the generality and standardization that are necessary for comparative, theory-based research, and the specificity and flexibility that are necessary for evaluating singular cases. The main research questions are:

1. What are the requirements and design principles for a comprehensive methodology for evaluating serious gaming?
2. To what extent do SGs contribute to learning in a real-world context?
3. What factors contribute to or determine this learning?
4. To what extent and under what conditions can serious game-based learning be transferred to the real world?

Cases

Between 2005 and 2012, more than 300 SG sessions with 12 SGs were evaluated as systematically, uniformly, quantitatively and qualitatively as possible, by TU-Delft in cooperation with various partners. All of the games discussed below targeted the learning and training of professionals (to be), in many cases (future) engineers, on things like project management, complex decision making, planning, leadership and team work, combined with specific content and context-related professional knowledge. Advanced learning makes declarative, specialized and basic knowledge acquisition less significant than personal mastery, combining knowledge, skills and attitudes. Generally, we would not use SG to teach or train university students or professionals to do basic mathematical calculations or what have you, because such things can much better be taught with lectures, books, tutorials and cases. We use SG to let professionals gain more insight into social-technical complexity, and how to handle it for instance when they are put in the position of project leader. This does not make substance and knowledge acquisition irrelevant because it is the locus—not the focus—of their professional mastery. It should therefore be included in the narrative and game-play. Many of the games were not only played with university students but also with experienced and senior professionals. All games were played in a “facilitated” mode.

From the research and evaluation efforts of the different games, an evaluation method emerged. Over the years, step by step, we were first reusing evaluation items that worked well. Later, we started testing the items for things like construct validity (scaling, etc.), improved our data collection (see Figure 2 in later section), structured the constructs in an evaluation framework (see Figure 1 in later section), and now leading to structural equation modeling (see, eg, Bekebrede, 2010; Harteveld, 2012; I. S. Mayer, Warmelink & Bekebrede, 2013). At the time of writing, the still expanding dataset contains data on 2488 respondents (comprising male and female students and professionals, aged between 17 and 75) and 960 original variables on such aspects as player background, session, learning and/or policy context, game quality, player experiences, and immediate game results and effects. These data were gathered before, during and after the sessions in a quasi-experimental fashion. Table A1 contains an overview of the background characteristics of the games. Video impressions of four games can be found at http://signaturegames.nl; most other games can be visited on company or game websites. The games are summarized below:

(2) SimPort–MV2 (TU Delft, Tygron Serious Gaming, & Port of Rotterdam, n.d.): a computer-supported, multiplayer strategic planning game for higher education and advanced professional learning that revolves around the development of the second Maasvlakte area in the Port of Rotterdam, the Netherlands (Bekebrede, 2010).

(3) Ventum Online (“Ventum Online,” n.d.): a computer-supported, multiplayer management game for engineering students and professional project managers that centers on the development of an offshore wind farm.

(4) Construct.it (TU-Delft 2009): a computer-supported, multiplayer planning game for students in higher education concerning the urban reconstruction of a seaport area in a real Dutch town.

(5) Climate Game (“Climategame,” n.d.): a 3D computer-supported, multiplayer, strategic planning game for students in higher education and professional policymakers about integrated water management in relation to urban planning, climate change, etc. (Zhou, Mayer, Bekebrede, Warmerdam & Knepflé, 2013).

(6) SprintCity (“SprintCity,” n.d.): a computer-supported, multiplayer policy game focused on transit-oriented development around railway stations for professionals working in the field.

(7) Cyberdam (Stichting Rechten Online, 2007): an online, game-based virtual learning environment (VLE) in which teachers in higher education can create their own online role-playing games. Cyberdam is a platform in which 17 different games in 12 institutions have been developed and tested (I. S. Mayer et al, 2013).

Shark World (“Sharkworld,” n.d.): a single-player, multimedia, digital game for lower and higher vocational education in project management.


TeamUp (TU-Delft 2010): a digital, 3D, multiplayer game centered on team communication and leadership.

SimVenture (Venture Simulations Ltd, n.d.): a single-player, computer-based game centered on business entrepreneurship.

Table A2 presents the important respondent and background variables in the dataset on the following aspects per game and total for all games: (1) number of player-respondents (2488 in total); (2) first and most recent dates of play (between 2005 and 2012); (3) nationality (161 non-Dutch); (4) age (between 17 and 75); (5) gender (1650 male; 630 female); (6) student/professional (1831 students, 597 professionals); (7) education (1471 university; 219 middle voc. educ.; 500 higher voc. education); (8) frequency of playing analog games (never–daily); and (9) frequency of playing digital games (never–daily). In addition, and mainly for illustration, we include means (standard deviations) of seven basic multiple item constructs (Cronbach α in Table A2) to give an indication of player satisfaction with the quality of the SG (see Table A2): (1) clarity of the SG (five items); (2) realism of the SG (two items); (3) learning satisfaction (various items); (4) team engagement (three items); (5) facilitator quality (two items); (6) user interaction (two items); and (7) identification with role.

Comparative and longitudinal research
In theory, comparative, longitudinal research into GBL has a great many advantages: it provides the opportunity to vary the context or the intermediating variables, such as the number or the intrinsic motivation of students (Paas, Tuovinen, Merriënboer & Aubteen Darabi, 2005), modes of dissemination, the quality or experience of the game facilitators and the institutional facilities (eg, the quality of the classroom or computer infrastructure) (Kenny & Gunter, 2011; Kenny & McDaniel, 2011).

In practice, however, comparative, longitudinal research into GBL presents a great challenge. It requires a high level of discipline and synchronization among the stakeholders, the use of a common evaluation framework that is standardized and robust enough to compare but also flexible enough to allow adjustments to local conditions, institutions, course topics, games and times. It also requires efficient game-based evaluation tools and techniques to gather rich data on a wide spectrum of variables. And while doing the research, the games, the questions, the methods and the tools evolve. In this paper, we discuss the design of the methodology in eight steps, namely:

1. framing;
2. foundations and requirements;
3. conceptual framework;
4. quasi-experimental research design;
5. contextualization;
6. research questions and hypothesis;
7. operationalization; and
8. data reduction and analysis.

3For reasons of space and focus, in this paper we do not go into detail about the (validity and reliability) of the items (questions) nor the underlying statistical analysis.
Step 1: framing
Like “learning,” the notion “game” in a context of research or science can have different meanings. Consider the differences and possible overlap in the use of “game” as . . .

1. research theory: game theory as in economics, political science, etc. (Leyton-Brown & Shoham, 2008; Shubik, 1999);
2. research concept: organization, management, decision making as a strategic or political game (Scharpf, 1997; Steunenberg, Schmittden & Koboldt, 1999);
3. research object: studying game cultures, game economics, game politics, etc. (Castronova, 2005; Ermi & Mäyrä, 2003; Salomon & Soudoplatoff, 2010);
4. design artifact: game as a socio-technical design, as an artifact, etc. (Björk & Holopainen, 2005; Harteved, 2011; van der Spek, 2011);
5. research method: game as a research method comparable with simulation or experimentation (Barnaud, Promburom, Trebuil & Bousquet, 2007; I. S. Mayer, Carton, de Jong, Leijten & Dammers, 2004; Tykhonov, Jonker & Meijer, 2008);
6. intervention method: game as therapy, educational, learning, change- or decision-support method (Geurts, Duke & Vermeulen, 2007; Preschl, Wagner, Forstmeier & Maercker, 2011);
7. data gathering method: game as an environment for observation, group interview and data modeling (Cooper et al., 2010; Good & Su, 2011; Wood, Griffiths & Eatough, 2004).

Step 2: foundations and requirements

Literature overview
A great many PhD theses and related academic papers on the effects of one or a few GBL and/or serious gaming experiments have now been published (Bekebrede, 2010; Bekebrede, Warmelink & Mayer, 2011; Blunt, 2006; Bremson, 2012; Egenfeldt-Nielsen, 2005; Kato et al., 2008; Kuit, 2002; Leemkuil, 2006; Squire, 2004; van der Spek, 2011).

Several review articles on GBL have also been published, and such articles are now appearing with increasing frequency (Adams, 2010; Barlett, Anderson & Swing, 2008; Boyle, Connolly, Hainey & Boyle, 2012; Connolly et al., 2012; Coulthard, 2009; Egenfeldt-Nielsen, 2006; Girard, Ecalle & Magnan, 2012; Gosen & Washbush, 2004; Greenblat, 1973; Hays, 2005; Jenson & de Castell, 2010; Ke, 2009; J. Lee, 1999; I. S. Mayer, 2009; Papastergiou, 2009; Randel, Morris, Wetzelf & Whitehill, 1992).

However, few publications provide high-quality evaluation frameworks for what to measure in a comparative fashion and how to do so, taking into account the real-life and dynamic setting of the project (De Freitas & Oliver, 2006). A useful summary and review of 11 evaluation models have been presented by Hainey (Hainey, 2010) and colleagues (Connolly et al., 2009; Hainey & Connolly, 2010). Among the 11 frameworks reviewed are the four-dimensional evaluation framework proposed by De Freitas and colleagues (De Freitas & Oliver, 2006; de Freitas, Rebolledo-Mendez, Liarokapis, Magoulas & Poulovassilis, 2010) and Kirkpatrick’s four levels for evaluating training (Kirkpatrick, 1994, 2006). This overview is directed at the evaluation of (digital games) in (formal, K12) education and aiming at the construction of a model for GBL evaluation (Hainey, 2010). This may explain why other types of evaluation models like the Technology Acceptance Models for serious game adoption (Yusoff, Crowder & Gilbert, 2010) and, more importantly, Kriz and Hense’s framework for theory-based evaluation—used for simulation-games, also nondigital—are not taken into consideration (Bekebrede, 2010; Kriz & Hense, 2004, 2006).

Limitations of existing frameworks
The models and frameworks discussed above are high-level models. They specify a limited number of generic concepts that can or should be taken into consideration when evaluating SGs. With relevance in their own right, models like De Freitas’ four dimensional framework or Kirkpatrick’s
four levels of evaluation are not easy to use for exploratory or explanatory hypothesis generation and testing. Moreover, there are even fewer evaluation frameworks of GBL in higher education, let alone professional, in-company training or group and organizational learning. In sum, the existing models (we know) give:

1. few indications how to use the models, for what purpose, with what scope and under what conditions;
2. few procedures to validate the conceptual research/evaluation model;
3. few research hypotheses and research designs;
4. few definitions of or relations and interrelations between the concepts in the model; and
5. few operationalizations and validations of constructs.

Furthermore, in the application of the models, we see:

1. a dominance of single case-studies, one game, one context of application;
2. a lack of information on the questionnaires used;
3. a focus on the GBL of children in formal education; little attention to advanced–professional learning outside education; and
4. a focus on the learning of individuals in formal training or the educational context; little attention to the learning of teams, groups, organizations, networks or systems in a policy or organizational context.

Requirements
An important question, therefore, is what the requirements are for a good evaluation framework for serious game evaluation research. A generic evaluation framework (and corresponding procedures), for GBL and SGs research ideally (and based upon our experience with the evaluation of the 12 games between 2005 and 2012) has the following characteristics:

1. broad scope: takes into account the broad range of educational contexts, games, learning objectives and topics;
2. comparative: able to use certain data from different games for comparison;
3. standardized: in order to use a pre-/quasi-experimental research design, materials and procedures should be standardized;
4. specific: measuring data precisely by pinpointing variables;
5. flexible: as game play cannot be always predicted, data gathering should be flexible for measurement; however, still standardized, specific, etc.;
6. triangulated: using a mixed method approach with qualitative and quantitative data;
7. multileveled: individual, game, team, organization and system levels;
8. validated: validated research methods, for example, research method and game design;
9. expandable: possibility of measurements on new variables;
10. unobtrusive: using gaming for systematic and extensive data gathering (research, comparative or theory-based evaluation, etc.) needs to be unobtrusive;
11. fast and nontime consuming: using real-world cases for data gathering implies that tools and methods need to be fast and nontime consuming, because in real-world projects, not much time and resources can be devoted to research; and
12. multipurposed: persuading stakeholders to do data gathering beyond the obvious and minimal.

In practice, no such framework exists and trade-offs need to be made. A GBL or an SG evaluation needs to be broad in scope but light in operation; it must address both the formative and the summative purposes of evaluation (Bloom, Hastings & Madaus, 1972) and the evaluation interests of the designers, players, financers and other stakeholders. At the other end of the spectrum, the data need to be suitable for more in-depth analysis, in order to understand what happens and why.
**Step 3: conceptual framework**

**Elements**

A generic model for the social scientific research, evaluation and assessment of SGs in a real-world context should provide:

1. a flexible and generally applicable research model from which we can derive:
   (a) a set of research questions and hypotheses;
   (b) a research design for applying the model;
   (c) a suite of research tools and instruments; and
   (d) some guidelines, practices and rules for applying, falsifying, validating and improving the above.

2. Empirical testing of the robustness of the model.

**Comparative, theory-based evaluation**

Before the learning effectiveness and contributing factors can be established, an evaluation framework is required that allows:

(1) the operationalization of independent, dependent and mediating or context variables, like “engagement” (in this case, independent), “learning effectiveness” (in this case, dependent) and age (mediating), or psychological safety (context);

(2) a systematic, unobtrusive process of data gathering and data analysis; and

(3) the formulation of research questions and hypotheses based on a conceptual research model.

The core of the model is a deconstruction of GBL into:

1. The pregame condition: the subject’s attitudes, knowledge, skills and behavior relevant to GBL and SGs and/or the case at hand before playing the game. In the 12 cases, we measured a variety of items and constructs ranging from, for example, attitudes toward GBL to organizational commitment (see 3.1–3.4 in Figure 1).

2. The quality of the GBL intervention: subdivided into the quality of the game design itself, the game play, the interaction with the facilitator/teacher and the interaction with the digital game environment (see 4.1–4.2 in Figure 1).

3. The postgame condition: the subject’s attitudes, knowledge, skills and behavior relevant to the GBL etc. (see 5.1–5.4 in Figure 1).

4. Background variables referring to the person, student or professional as a participant:
   (a) Socio-demographic variables: age, sex, nationality, etc. (see 1.1 in Figure 1).
   (b) Professional and student characteristics: position, work experience, level of education, etc. (see 2.1 in Figure 1).

5. Mediating variables
   (a) Individual as a participant (eg, personality characteristics; Big 5, Hexaco) (see 1.2 in Figure 1).
   (b) Individual as a learner (learning styles, etc.) (see 1.3 in Figure 1).
   (c) Individual as a gamer: (eg, game skill, game experiences, game attitudes, game-play style, etc.) (see 1.4 in Figure 1).
   (d) Professional/student as a serious gamer (eg, previous experiences with SGs in a professional context) (see 2.4 in Figure 1).
   (e) Professional/student as a participant (eg, intrinsic/extrinsic motivation) (Ainley & Armatas, 2006).

6. Context variables: organizational/institutional climate in which the GBL/SG takes place (eg, commitment to the organization, identification with leader or organization, psychological safety) (see 6.1 in Figure 1).
7. First-order learning: direct influence of playing the game on the individual, small group attitudes, knowledge, skills or behavior (see 7 in Figure 1).
8. Second-order learning: direct/indirect, short- or long-term influence of the game in the large (including design process, sessions, discussions, publications, other interventions, etc.) on the group, network, organizational and system levels (see 8 in Figure 1).

**Step 5: quasi-experimental research design**

Now the model can be translated into a quasi-experimental design: from the simple “post-test only,” to a “pre-test/post-test” design, a “randomized (R),” “control group (C),” “repeated measurement” design (Campbell & Stanley, 1963; Cook & Campbell, 1979; Creswell, 2002). Figure 2 illustrates the basic translation of the conceptual model into a quasi-experimental design (R and C not included in figure):

**Step 6: contextualization**

**Data gathering**

One of the special features of SGs for advanced learning is that they provide excellent environments for mixed method data gathering, or triangulation: from crowd sourcing to panel discussions, surveys and observations/video observations. Figure 3 gives a visual impression of methods that can be mixed with SGs.

For the games listed in Table A1, evaluation data were gathered through mixed methods, mostly combining pregame and postgame surveys among the players, live or video observations, transcripts of after-action reviews and game results. In a few cases, methods were applied more rigorously with in-game knowledge tests or network and communication analyses from video.
observations. Table 1 gives an overview of how to mix the various methods in pregame, in-game and postgame stages.

**Step 6: defining questions and hypotheses**

We classify the types of research questions and research hypotheses that can guide GBL and SGs research into:

1. **design-oriented research (artifact): “making it (better)”**:
   - (a) The validation of specific and generic game-based artifacts and events; and
   - (b) The development and validation of design theories, methods and tools.
2. **Intervention-oriented research (learning, change, policymaking, management): “making it work”**:
   - (a) The learning effectiveness/impact of game-based interventions; and
   - (b) The transfer of game-based interventions to the real world.
3. **Domain-oriented research (healthcare, military, energy, etc.): “making it matter”**:
   - (a) The effectiveness of the use of SGs to understand the complexity, dynamics in specific domains.
<table>
<thead>
<tr>
<th>How</th>
<th>What?</th>
<th>Self-reported</th>
<th>Qual.</th>
<th>Quant.</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-game</td>
<td>Personality, player experiences, context, etc.</td>
<td>Interviews, focus group, logbook.</td>
<td>Survey, questionnaires</td>
<td>In-game log file analysis, network analysis, data mining, biometric observation.</td>
<td></td>
</tr>
<tr>
<td>In-game</td>
<td>Socio-demographic, opinions, motivations, attitudes, engagement, power, influence, learning, reputation, network centrality, learning satisfaction, etc.</td>
<td>Game-based behavioral assessment.</td>
<td>Psychometric, socio-metric tests: eg, personality, leadership, team roles, IQ.</td>
<td>Video, audio personal observation, ethnography, maps, figures, drawings, pictures, etc.</td>
<td></td>
</tr>
<tr>
<td>Post-game</td>
<td>Qual.</td>
<td>Behavior, skills, etc.</td>
<td>Quant.</td>
<td>Qual.</td>
<td>Quant.</td>
</tr>
<tr>
<td></td>
<td>Values, knowledge, attitudes, skills, personality, power</td>
<td>Behavioral performance of student, professionals, player and/or facilitator, others; decisions, strategies, policies, emotions, conflicts, etc.</td>
<td>Behavioral performance of student, professionals, player and/or facilitator, others; decisions, strategies, policies, emotions, conflicts, etc.</td>
<td>Biophysical–psychological responses, like stress (heart freq., perspiration).</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>How</th>
<th>What?</th>
<th>Self-reported</th>
<th>Qual.</th>
<th>Quant.</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-game</td>
<td>Personality, player experiences, context, etc.</td>
<td>Interviews, focus group, logbook.</td>
<td>Survey, questionnaires</td>
<td>In-game log file analysis, network analysis, data mining, biometric observation.</td>
<td></td>
</tr>
<tr>
<td>In-game</td>
<td>Socio-demographic, opinions, motivations, attitudes, engagement, power, influence, learning, reputation, network centrality, learning satisfaction, etc.</td>
<td>Game-based behavioral assessment.</td>
<td>Psychometric, socio-metric tests: eg, personality, leadership, team roles, IQ.</td>
<td>Video, audio personal observation, ethnography, maps, figures, drawings, pictures, etc.</td>
<td></td>
</tr>
<tr>
<td>Post-game</td>
<td>Qual.</td>
<td>Behavior, skills, etc.</td>
<td>Quant.</td>
<td>Qual.</td>
<td>Quant.</td>
</tr>
<tr>
<td></td>
<td>Values, knowledge, attitudes, skills, personality, power</td>
<td>Behavioral performance of student, professionals, player and/or facilitator, others; decisions, strategies, policies, emotions, conflicts, etc.</td>
<td>Behavioral performance of student, professionals, player and/or facilitator, others; decisions, strategies, policies, emotions, conflicts, etc.</td>
<td>Biophysical–psychological responses, like stress (heart freq., perspiration).</td>
<td></td>
</tr>
</tbody>
</table>
4. Disciplinary research (methodology, ethics, explanatory and interpretative theories): “making it understandable”:
   (a) the sociological, economic, political, cultural, etc. frames on SGs;
   (b) Theory construction on GBL and SGs;
   (c) Methodology: design and validation of research methods and tools; and
   (d) Reflection and ethics.

Depending upon the case at hand, a pregame, an in-game and a postgame instrument for measuring or observing relevant variables can now be defined or constructed. There is a great variety of games, players and learning contexts, and trade-offs need to be made between time, resources and the focus of the evaluation (see above). The first-order effects can be established as changes between pregame and postgame measurements, with or without a control group.

Knowledge acquisition for instance can be evaluated with knowledge tests (eg, exam style) making it even stronger when combined with self-reporting and, or in-game performance measurements. An excellent example of such a multimethod assessment of knowledge acquisition among professionals can be found in Harteveld (2012). The ex-post, self-reporting or self-assessment of change and learning however is quite common and often sufficient. Recent studies found a strong and significant correlation between self-reported learning, in-game performance improvement and pregame, postgame testing (Harteveld, 2012). Yet, even when based upon self-reporting, high-quality questionnaires with items, constructs and scales for comparative and longitudinal measurements of knowledge acquisition and learning are not commonly available.

**Step 7: operationalization**

The operationalization of the generic conceptual model (Figure 1) in the context of a dynamic, multistakeholder project can be a real challenge. First, most educational or client organizations have their own procedures and preferences for evaluations; sometimes a set of course evaluation questions is mandatory. Paying clients are not always inclined to evaluate beyond their own immediate needs (eg, “Did we get what we expected and were promised?”). In many cases, we need to convince stakeholders that for the purpose of advancing serious gaming, a thorough evaluation should be done. We need to emphasize and ensure privacy, safety and nonobtrusiveness. Another handicap is that the client usually expects case-specific evaluation questions. Ad hoc and case-specific evaluation questions stand in the way of comparative, longitudinal research. Through the years, we have found flexible ways of working, trying to validate and reuse as much as possible the items, constructs and scales. In our comparative research (Tables A1 and A2), we have gradually build up a set of validated and reusable questions for the following constructs and items:

**Pregame**


2. Previous experiences/skills: with computers, games, VLEs, etc. (Erfani et al, 2010; Harper et al, 2007; Mortagy & Boghikian-white, 2010).


5. Behaviors (behavioral intentions)
6. Group, team, organizational characteristics: team/group conflict, psychological safety, psychological collectivism, team and organizational commitment, etc. (Brockner & Higgins, 2001; Carmeli, Brueller & Dutton, 2009; Edmondson, 1999; Ferris, 2005; Jackson, Colquitt, Wesson & Zapata-Phelan, 2006).

In-game
7. Game performance: based upon in-game scores, such as time, avoidable mistakes, etc. (Baba, 1993; Blumberg, 2000; Oslin, Mitchell & Griffin, 1998; Tallir, Lenoir, Valcke & Musch, 2007; Trepte & Reinecke, 2011).
8. Game play: effort; dominance, influence, power, etc.

Postgame
10. Game experience: engagement, fun while playing the game, etc. (Boyle et al, 2012; Mayes & Cotton, 2001; Schuurink, Houtkamp & Toet, 2008).
11. Player satisfaction with: the game (eg, clarity, realism); user interaction (eg, attractiveness, ease of use, computer malfunctions, support, etc.); the quality of the facilitator (eg, supportive, player identification with facilitator); the interaction with other students (eg, player efforts, motivation); identification of players with role; team engagement (Olsen, Procci & Bowers, 2011; Reichlin et al, 2011) (see also Table A2 and below).
12. First-order learning (short term, individual, participants).
   (a) Player learning satisfaction, self-reported, self-perceived learning, for example broad range of items.
   (b) Measured changes in knowledge, attitudes, skills and behaviors (behavioral intentions).
13. Second-order learning (medium term, long term, collective, participants and nonparticipants):
   (a) Self-reported, case-based, reconstructive: asking clients, participants, etc. how the results of the GBL have been implemented.
   (b) Measured changes in team, group, organizational characteristics: safety, commitment, performance, performance, etc.

Step 8: data reduction and analysis
A final and important concern is data reduction and analysis. Over the years, we have varied and changed items, questions and constructs. Data reductions through factor analysis and reliability analysis of scales increasingly allow us to select the influential and distinguishing items and construct scales. The present dataset now contains a total number of 960 variables about the 12 games.

As indicated above, we frequently use the same scales or constructs to measure “game design quality” and “player satisfaction” through constructs like clarity, realism, facilitator quality, user interaction etc. Table A2 gives an indication of how the 12 games individually, on average and comparatively score on seven basic constructs. The end goal of comparative research is to test the efficacy of GBL and SGs through structural equation models (forthcoming publications; see also I. S. Mayer et al, 2013) using also validated or newly constructed psychometric scales and constructs for the broad range of constructs listed above.

Conclusions
We have formulated both the requirements for and a conceptual research model that can be translated into quasi-experimental research designs and operationalized into an evaluation model for specific cases and contexts of GBL. We have also demonstrated the principles and workings of the model on the basis of a comparative case of 12 SGs. The strengths of our model
we believe is that it finds a balance between broad application scope through flexibility and cross-case comparative research through a modular evaluation framework. The framework does not rigidly superimpose hypotheses about sGBL but does support the generation, exploration and testing of hypothesis and structural models (I. S. Mayer et al., 2013).

We are aware of at least one weakness in our approach to evaluate the 12 games: we had multiple roles and potential interests in their evaluation as designers, users, teachers, facilitators, entrepreneurs, etc. We have done our best over the years to separate our roles as much as possible and to approach evaluation as systematically and critically as conditions allowed. In our role as designers or teachers, we would like to see confirmed that the games we designed and used were engaging and effective learning tools; as researchers and evaluators however we had to live with the fact that sometimes they were not. For the designer or teacher this can be painful; for the researcher it raises questions about if, what, when and how serious gaming works (G. J. Hofstede, De Caluwe & Peters, 2010). By and large, we see three ways forward in our research:

1. to perform a comparative analysis of the data in order to find the factors that influence the efficacy of GBL and SGs;
2. to enrich and improve the constructs and scales for GBL and SGs; and
3. to use a digital tool for the quasi-experimental design research into SGs and GBL that allows the immediate coupling of a variety of pregame, in-game and postgame data.

Acknowledgements

The authors wish to thank all persons, companies and institutions that have contributed to the design, playing and evaluation of the games mentioned in this paper. Parts of the game design and research were funded by the Next Generation Infrastructures foundation (http://www.nextgenerationinfrastructures.eu). Demo videos of the games, details about their designers and clients can be found at http://signaturegames.nl/gamelab and referenced game and institutional websites.

References


Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools (chapter 1) (pp. 1–33). IGI.


## Appendix

### Table A1: Overview of game cases

<table>
<thead>
<tr>
<th>Game name</th>
<th>Start year</th>
<th>End year</th>
<th>Current status</th>
<th>Client organization</th>
<th>Game concept</th>
<th>Game platform / engine</th>
<th>Game objective</th>
<th>Education (TU-Delft) students</th>
<th>Professional learning &amp; training</th>
<th>Support policymaking &amp; management</th>
<th>Support academic research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Veerkracht (Resilience)</td>
<td>2011</td>
<td>2013</td>
<td>Pending</td>
<td>Rijkswaterstaat, the Netherlands Ministry Infrastructure Environment, the Netherlands, Hekom, Ospar, ICES</td>
<td>Custom made</td>
<td>Custom made</td>
<td>Leadership and personal competences training</td>
<td>√</td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>MSP (Marine Spatial Planning) Challenge 2011</td>
<td>2011</td>
<td>Continues</td>
<td>Pending</td>
<td>Rijkswaterstaat, the Netherlands Ministry Infrastructure Environment, the Netherlands, Hekom, Ospar, ICES</td>
<td>Custom made</td>
<td>Custom made</td>
<td>Leadership and personal competences training</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Shark World</td>
<td>2010</td>
<td>Continues</td>
<td>Used occasionally by client</td>
<td>TU-Delft</td>
<td>Custom made</td>
<td>Custom made</td>
<td>Training project management skills</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Sprintstad</td>
<td>2009</td>
<td>Continues</td>
<td>Used frequently by client</td>
<td>Vereniging Deltametropool,</td>
<td>Custom made</td>
<td>Custom made</td>
<td>Insight in complexity of urban planning and transportation</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team-Up</td>
<td>2009</td>
<td>Continues</td>
<td>Used occasionally</td>
<td>Vereniging Deltametropool,</td>
<td>Custom made</td>
<td>Mod</td>
<td>Team role performance and competences training</td>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>TopSim</td>
<td>2009</td>
<td>Continues</td>
<td>Used occasionally</td>
<td>Royal Dutch Shell, corporate learning</td>
<td>Custom made</td>
<td>MOD</td>
<td>Training skills and theory on policy making</td>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Supervisor / Hazard Recognition</td>
<td>2008</td>
<td>2011</td>
<td>Research</td>
<td>Royal Dutch Shell, corporate learning</td>
<td>Custom made</td>
<td>Mod</td>
<td>Training of operational safety and emergency procedures</td>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Construct.IT</td>
<td>2008</td>
<td>2009</td>
<td>Closed</td>
<td>3TU construction management</td>
<td>Custom made</td>
<td>Custom made</td>
<td>Insight and skills for construction management projects</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Watergame</td>
<td>2006</td>
<td>Continues</td>
<td>Used frequently by client</td>
<td>MOD</td>
<td>Custom made</td>
<td>Custom made</td>
<td>Insight in complexity of water management, urban planning</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Cyberrdam 3D</td>
<td>2006</td>
<td>2007</td>
<td>Closed</td>
<td>n.a.</td>
<td>Custom made</td>
<td>COTS</td>
<td>Training skills and theory on policy making</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Sim Port-MV2</td>
<td>2005</td>
<td>Continues</td>
<td>Used frequently by client</td>
<td>Port of Rotterdam (PoR)</td>
<td>Custom made</td>
<td>Custom made</td>
<td>Insight in complexity of port planning</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Levee Patroller</td>
<td>2005</td>
<td>Continues</td>
<td>Used actively by client</td>
<td>Deltavers, Water Boards</td>
<td>Custom made</td>
<td>Mod</td>
<td>Training of operational safety and emergency procedures</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Sim-MV2</td>
<td>2004</td>
<td>2005</td>
<td>Closed</td>
<td>Port of Rotterdam</td>
<td>Custom made</td>
<td>Custom made</td>
<td>Insight in complexity of port planning</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Sieberrdam : railway area</td>
<td>2004</td>
<td>2009</td>
<td>Closed</td>
<td>n.a.</td>
<td>Custom made</td>
<td>COTS</td>
<td>Training skills and theory on policy making</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Ventum Online</td>
<td>2003</td>
<td>2012</td>
<td>Used occasionally</td>
<td>n.a.</td>
<td>Custom made</td>
<td>Custom made</td>
<td>Training skills and theory on project management</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Game name</td>
<td>Virtual</td>
<td>Technology use</td>
<td>Stand alone / networked</td>
<td>Single / multiplayer</td>
<td>External realism</td>
<td>Player interaction</td>
<td>Panfercula</td>
<td>Min no players per game instance</td>
<td>Optimal no of participants in session</td>
<td>Average (un)interrupted duration of game in hours</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------</td>
<td>----------------------</td>
<td>-------------------------</td>
<td>----------------------</td>
<td>-----------------</td>
<td>--------------------</td>
<td>-------------------------------</td>
<td>-----------------------------------</td>
<td>-----------------------------------</td>
<td>----------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Veerkracht (Resilience)</td>
<td>2D visuals</td>
<td>Multimedia supported</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>High</td>
<td>Role play</td>
<td>Game manual; game maps;</td>
<td>15</td>
<td>15–30</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>MSP (Marine Spatial Planning) Challenge 2011</td>
<td>2D virtual</td>
<td>Computer supported</td>
<td>Stand alone</td>
<td>Multiplayer</td>
<td>High</td>
<td>Role play</td>
<td>Game manual; game maps;</td>
<td>100</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Shark World</td>
<td>2D virtual</td>
<td>Multimedia supported</td>
<td>Networked</td>
<td>Single player</td>
<td>Low</td>
<td>Competition</td>
<td>Game manual; game maps;</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprintstad</td>
<td>2D virtual</td>
<td>Computer supported</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>High</td>
<td>Role play</td>
<td>Game manual; game maps;</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Team-Up</td>
<td>3D virtual</td>
<td>Computer supported</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>Low</td>
<td>Competition</td>
<td>Game manual; game maps;</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TopSim</td>
<td>2D visuals</td>
<td>Computer supported</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>Medium</td>
<td>Competition</td>
<td>Game manual; game maps;</td>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervisor / Hazard Recognition</td>
<td>3D virtual</td>
<td>Computer-based</td>
<td>Stand alone</td>
<td>Single player</td>
<td>High</td>
<td>Action</td>
<td>Game manual; game maps;</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construct.IT</td>
<td>2.5D virtual</td>
<td>Computer supported</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>High</td>
<td>Role play</td>
<td>Game manual; game maps;</td>
<td>15</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watergame</td>
<td>2.5D virtual</td>
<td>Computer based</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>High</td>
<td>Role play</td>
<td>Computers and Internet; game software</td>
<td>4</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyberdam 3D</td>
<td>3D virtual</td>
<td>Computer based</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>Medium</td>
<td>Role play</td>
<td>Game manual; game maps;</td>
<td>?</td>
<td>n.a. online</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sim Port-MV2</td>
<td>2.5D visuals</td>
<td>Computer-based</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>High</td>
<td>Role play</td>
<td>Computers and Internet; game software</td>
<td>5</td>
<td>10–15 (2/3 teams)</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Leeve Patroller</td>
<td>3D virtual</td>
<td>Computer based</td>
<td>Stand alone</td>
<td>Single player</td>
<td>High</td>
<td>Action</td>
<td>Computers and Internet;</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sim-MV2</td>
<td>2D virtual</td>
<td>Computer based</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>High</td>
<td>Role play</td>
<td>Game manual; game maps;</td>
<td>5</td>
<td>10–15 (2/3 teams)</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Sieberdam : railway area reconstruction</td>
<td>2D visuals</td>
<td>Computer supported</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>Medium</td>
<td>Role play</td>
<td>Game manual; game maps;</td>
<td>20</td>
<td>n.a. online</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ventum Online</td>
<td>2D visuals</td>
<td>Computer-based</td>
<td>Networked</td>
<td>Multiplayer</td>
<td>Medium</td>
<td>Role play</td>
<td>Game manual; game maps;</td>
<td>8</td>
<td>12</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Table A2: Dataset (on January 10, 2013)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>n = freq. in row</td>
<td>482</td>
<td>683</td>
<td>635</td>
<td>65</td>
<td>36</td>
<td>45</td>
<td>23</td>
<td>347</td>
<td>96</td>
<td>26</td>
<td>26</td>
<td>24</td>
</tr>
<tr>
<td>n = % of total row</td>
<td>19.4%</td>
<td>27.3%</td>
<td>25.4%</td>
<td>2.6%</td>
<td>1.4%</td>
<td>1.8%</td>
<td>0.9%</td>
<td>13.9%</td>
<td>3.8%</td>
<td>1.0%</td>
<td>1.0%</td>
<td>1.0%</td>
</tr>
<tr>
<td>Latest date of play</td>
<td>13-3-2012</td>
<td>5-6-2012</td>
<td>20-4-2012</td>
<td>6-1-2010</td>
<td>24-11-2011</td>
<td>22-6-2010</td>
<td>15-12-2010</td>
<td>29-11-2012</td>
<td>20-6-2012</td>
<td>20-6-2012</td>
<td>20-6-2012</td>
<td>20-4-2012</td>
</tr>
<tr>
<td>Freq. in % subcategory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total freq.</td>
<td>145</td>
<td>114</td>
<td>132</td>
<td>100</td>
<td>99</td>
<td>4.0</td>
<td>99</td>
<td>4.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nationality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>114</td>
<td>132</td>
<td>145</td>
<td>14</td>
<td>14</td>
<td>24</td>
<td>29</td>
<td>27</td>
<td>44</td>
<td>18</td>
<td>14</td>
<td>114</td>
</tr>
<tr>
<td>Other</td>
<td>12</td>
<td>2</td>
<td>29</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>27</td>
<td>67</td>
<td>12</td>
<td>12</td>
<td>a.a.</td>
</tr>
<tr>
<td>Missing</td>
<td>470</td>
<td>681</td>
<td>606</td>
<td>65</td>
<td>45</td>
<td>23</td>
<td>291</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of game sessions (days with one or more instances)</td>
<td>114</td>
<td>145</td>
<td>132</td>
<td>100</td>
<td>99</td>
<td>4.0</td>
<td>114</td>
<td>145</td>
<td>14</td>
<td>14</td>
<td>24</td>
<td>29</td>
</tr>
<tr>
<td>Age</td>
<td>27.6 (8.9)</td>
<td>23.7 (1.9)</td>
<td>21.3 (4.0)</td>
<td>23.9 (2.5)</td>
<td>40.2 (12.8)</td>
<td>37.9 (2.2)</td>
<td>21.2 (1.6)</td>
<td>35.7 (9.8)</td>
<td>41.7 (11.0)</td>
<td>24.8 (2.1)</td>
<td>24.8 (2.1)</td>
<td>45.8 (7.7)</td>
</tr>
<tr>
<td>Student/professional</td>
<td>2428</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>326</td>
<td>662</td>
<td>632</td>
<td>65</td>
<td>10</td>
<td>0</td>
<td>23</td>
<td>61</td>
<td>0</td>
<td>26</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>Professional</td>
<td>136</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>45</td>
<td>0</td>
<td>284</td>
<td>63</td>
<td>0</td>
<td>24</td>
<td>597</td>
</tr>
<tr>
<td>Secondary education</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Middle vocational education</td>
<td>9</td>
<td>0</td>
<td>210</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td></td>
<td>219</td>
</tr>
<tr>
<td>Higher vocational education</td>
<td>155</td>
<td>0</td>
<td>345</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td></td>
<td>500</td>
</tr>
<tr>
<td>University</td>
<td>126</td>
<td>662</td>
<td>555</td>
<td>63</td>
<td>—</td>
<td>—</td>
<td>23</td>
<td>27</td>
<td>15</td>
<td>1471</td>
<td>39.9</td>
<td></td>
</tr>
</tbody>
</table>
Table A2: Continued

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Playing analogue games</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1493</td>
<td>100</td>
</tr>
<tr>
<td>1. Rarely</td>
<td>51</td>
<td>47</td>
<td>62</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>17</td>
<td>1</td>
<td>40</td>
<td></td>
<td>190</td>
<td>12.7</td>
<td></td>
</tr>
<tr>
<td>2. Couple of times a year</td>
<td>131</td>
<td>225</td>
<td>224</td>
<td>29</td>
<td>7</td>
<td>12</td>
<td>40</td>
<td>3</td>
<td>40</td>
<td></td>
<td>668</td>
<td>44.7</td>
<td></td>
</tr>
<tr>
<td>3. Monthly</td>
<td>79</td>
<td>155</td>
<td>128</td>
<td>11</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td>40</td>
<td></td>
<td>391</td>
<td>26.2</td>
<td></td>
</tr>
<tr>
<td>4. Weekly</td>
<td>37</td>
<td>97</td>
<td>54</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>34</td>
<td></td>
<td>209</td>
<td>14.0</td>
<td></td>
</tr>
<tr>
<td>5. Daily</td>
<td>7</td>
<td>18</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td></td>
<td>35</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>Playing digital games</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1853</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>1. Rarely</td>
<td>53</td>
<td>90</td>
<td>81</td>
<td>11</td>
<td>14</td>
<td>3</td>
<td>1</td>
<td>82</td>
<td>33</td>
<td></td>
<td>372</td>
<td>20.1</td>
<td></td>
</tr>
<tr>
<td>2. Couple of times a year</td>
<td>98</td>
<td>154</td>
<td>139</td>
<td>21</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>81</td>
<td>27</td>
<td></td>
<td>562</td>
<td>30.3</td>
<td></td>
</tr>
<tr>
<td>3. Monthly</td>
<td>55</td>
<td>110</td>
<td>95</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>42</td>
<td>6</td>
<td></td>
<td>344</td>
<td>18.6</td>
<td></td>
</tr>
<tr>
<td>4. Weekly</td>
<td>63</td>
<td>119</td>
<td>114</td>
<td>12</td>
<td>6</td>
<td>1</td>
<td>7</td>
<td>46</td>
<td>6</td>
<td></td>
<td>400</td>
<td>21.6</td>
<td></td>
</tr>
<tr>
<td>5. Daily</td>
<td>36</td>
<td>49</td>
<td>52</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>20</td>
<td>6</td>
<td></td>
<td>175</td>
<td>9.4</td>
<td></td>
</tr>
</tbody>
</table>

**Game profile (example constructs)**

| Construct 1: Clarity of the SG (5 items) | Cronbach's α | n | 1.8 (0.6) | 3.4 (0.6) | 3.3 (0.7) | 3.7 (0.5) | 3.9 (0.9) | 3.5 (0.5) | 3.2 (0.7) | 3.8 (0.9) | 4.1 (0.6) | 3.4 (0.8) | 3.6(0.8) | 4.2 (0.5) | n = 1170 | 0.76 |
| Construct 2: Realism of the SG (2 items) | Cronbach's α | n | 3.7 (0.7) | 3.5 (0.7) | 3.2 (0.8) | 4.1 (0.7) | 3.3 (0.9) | 2.7 (0.8) | 4.1 (0.8) | 4.0 (0.8) | 3.5 (0.9) | 4.1 (1.1) | 3.6 (1.1) | n = 1277 | 0.71 |
| Construct 3: Learning satisfaction (var. items) | Cronbach's α | n | 3.0 (0.8) | 3.2 (0.6) | 3.2 (0.6) | 3.1 (0.7) | 4.1 (0.9) | 3.4 (1.1) | 2.8 (0.7) | n.a. | 3.6 (0.8) | 4.0 (0.5) | 4.0 (0.5) | 3.2 (0.5) | n = 1151 | 0.88–0.91 |
| Construct 4: Team Engagement (3 items) | Cronbach's α | n | 3.8 (0.7) | 3.6 (0.7) | 3.3 (0.9) | 3.6 (0.7) | 3.6 (0.8) | 3.4 (0.9) | — | 4.2 (0.6) | 3.5 (1.0) | 4.2 (0.7) | — | n = 1145 | 0.74 |
| Construct 5 Facilitator quality (2 items) | Cronbach's α | n | 4.0 (0.6) | 3.8 (0.7) | 3.4 (0.9) | 3.5 (0.7) | 4.1 (0.8) | 4.0 (0.5) | 2.6 (0.8) | 4.1 (0.9) | 4.1 (0.9) | 4.3 (0.6) | 3.6 (0.9) | 3.5 (0.9) | n = 1170 | 0.76 |
| Construct 6: User interaction (2 items) | Cronbach's α | n | 3.7 (0.6) | 3.7 (0.7) | 3.4 (0.7) | 3.8 (0.9) | 3.3 (0.9) | 3.5 (0.8) | 4.1 (0.7) | 3.8 (0.7) | n.a. | 3.2 (0.8) | — | n = 1634 | 0.69 |
| Construct 7: Identification with role | Cronbach's α | n | 3.9 (0.6) | 3.7 (0.6) | 3.6 (0.8) | 3.9 (0.6) | 3.5 (0.8) | — | 3.3 (0.8) | 4.1 (0.7) | — | 3.6 (0.9) | 4.0 (0.8) | 4.2 (0.5) | n = 1181 | 0.7 |