Chapter XV
Towards the Development of a Games-Based Learning Evaluation Framework

Thomas Connolly
University of the West of Scotland, Scotland

Mark Stansfield
University of the West of Scotland, Scotland

Thomas Hainey
University of the West of Scotland, Scotland

ABSTRACT

The field of games-based learning (GBL) has a dearth of empirical evidence supporting the validity of the approach (Connolly, Stansfield, & Hainey, 2007a; de Freitas, 2006). One primary reason for this is a distinct lack of frameworks for GBL evaluation. The literature has a wealth of articles suggesting ways that GBL can be evaluated against particular criteria with various experimental designs and analytical techniques. Based on a review of existing frameworks applicable to GBL and an extensive literature search to identify measurements that have been taken in relevant studies, this chapter will provide general guidelines to focus researchers on particular categories of evaluation, individual measurements, experimental designs and texts in the literature that have some form of empirical evidence or framework relevant to researchers evaluating GBL environments particularly focusing on learner performance. A new evaluation framework will be presented based on the compilation of all the particular areas and analytical measurements found in the literature.
INTRODUCTION

One of the primary concerns associated with the GBL literature is the dearth of empirical evidence supporting the validity of the approach (Connolly, Stansfield, & Hainey, 2007a; de Freitas, 2006). O’Neil et al. (2005) believe that an essential element missing is the ability to properly evaluate games for education and training purposes. If games are not properly evaluated and concrete empirical evidence is not obtained in individual learning scenarios that can produce generalisable results, then the potential of games in learning can always be dismissed as unsubstantiated optimism.

In the O’Neil study, a large amount of literature was collected and analysed from the PsycINFO, EducationAbs, and SocialAcIAbs information systems. Out of several thousand articles, 19 met the specified criteria for inclusion and had some kind of empirical information that was either qualitative, quantitative or both. The literature was then viewed through Kirkpatrick’s four levels for evaluating training and the augmented CRESST model. The majority of the studies reviewed analysed performance on game measurements. Other studies included observation of military tactics used, observation, time to complete the game, transfer test of position location, flight performance and a variety of questionnaires including exit, stress and motivation questionnaires.

The review of empirical evidence on the benefit of games and simulations for educational purposes is a recurring theme in the literature and can be traced even further back. For example, Randel, Morris, Wetzel, and Whitehill (1992) examined 68 studies from 1963 comparing simulations/games approaches and conventional instruction in direct relation to student performance. Some of the following main discoveries were made:

• 38 (56%) of the studies found no difference; 22 (32%) of the studies found a difference that favoured simulations/games; 5 (7%) of studies favoured simulations/games however control was questionable; 3 (5%) found differences that favoured conventional instruction.
• With regards to retention simulations/games induced greater retention over time than conventional techniques.
• With regards to interest, out of 14 (21%) studies, 12 (86%) showed a greater interest in games and simulations over conventional approaches.

Although lack of empirical evidence supporting GBL is not a new issue, the growing popularity of computer games in conjunction with recent advances in games and hardware technology, the emergence of virtual worlds and massively multiplayer online games (MMOGs), reinforces the need for a flexible evaluation framework that can be used by Evaluation researchers. This chapter presents such an evaluation framework.

In the next section, we examine previous research and, in particular, discuss the types of evaluation that can be used and their applicability and importance in the field of GBL. We also examine previous evaluation frameworks that have been presented in the literature that could be applicable to GBL and follow that with the results of an extensive literature review to identify studies that performed some form of evaluation and attempted to take appropriate measurements through various experimental designs particularly focusing on learner performance. On the basis of these reviews, we then present a flexible framework for GBL evaluation from a pedagogical perspective. In the final section we discuss future validation of the new framework.

PREVIOUS RESEARCH

Evaluation

Ainsworth (2003) divides evaluation into two main types: evaluation to inform design encompassing
cognitive walkthrough, heuristic evaluation and formative evaluation and evaluation to assess an end product or determine the best use for that product, encompassing summative evaluation. As we discuss shortly, we found very few academic articles in the GBL literature that actually addressed formative evaluation and even less that addressed summative evaluation. According to Ainsworth the most important general questions of evaluation are:

- What is to be done with the information collected?
- What are the appropriate forms of measurement?
- What is the most appropriate experimental design?
- What is an appropriate form of comparison?
- What is an appropriate context?

Formative Evaluation

Ogle (2002) states that formative evaluation is “a systematic and empirical process although rarely a scientific one”. Citing Tessmer (1993), Ogle highlights that instructional designers generally carry out formative evaluations as they have intimate knowledge of the material and are also the most qualified to put comments made to productive use. One advantage of formative evaluation is that it also allows evaluation of the instruments of evaluation. For example, evaluation of a prototype GBL application may be performed a number of times in different contexts and learners can assist in highlighting any ambiguity or weaknesses in the instruments of evaluation, such as ambiguous questions in a questionnaire. This allows the experiment to be revised and improved each time to gather more productive results. There are several variations of formative evaluation, however the main ones are as follows:

- **Expert Reviews** – Conducted very early on in the evaluation process and is particularly focused on instructional content, technical quality or accuracy. The primary goal is to get expert reviewers to highlight things that are not right and offer correctional advice.
- **One-To-One Evaluations** – Designed for the developer to work with a number of potential learners from the intended user base, primarily to assess the learners’ reaction to content and assess particular indicators of performance. Discussion techniques and questions should be used at this stage to obtain information.
- **Small Group Evaluations** – Conducted using small groups in which the instructor interacts with the learners in the same type of environmental context as the intervention will be used. The main goal is to refine the instruction by the collection of descriptive and quantitative feedback.
- **Field trials** – Also known as field-tests or beta tests, field trials are designed to see if the changes made from the small group evaluation were effective and whether the intervention can be used in the intended context. It consists of the instructor acting as an observer while the intervention is used with a larger group (e.g. 20 - 40 learners) in a “situated evaluation”.

Braden (1992) believes that a primary weakness of instructional design models is that a formative evaluation is performed at the end, if at all, and should be performed during the entire process and be consistently iterative.

Summative Evaluation

Ainsworth (2003) points out that the aim of summative evaluation is to assess an end product and as a result this type of evaluation is usually performed by external evaluators. The two main types of summative evaluation identified by
Towards the Development of a Games-Based Learning Evaluation Framework

Harpel (1978) are: research-oriented and management-oriented. Research-oriented summative evaluations are to validate and improve programs. Management-oriented summative evaluations are to assess cost and whether the programs did what they were supposed to do.

Problems Evaluating GBL

Dondi and Moretti (2007) identify two general and very important issues associated with evaluation of GBL:

- Producing a general framework of evaluation is difficult unless it is abstract. This is primarily because the evaluation processes are inextricably linked to the main goal of the evaluation, which leads to the question of “what exactly is being evaluated about the GBL intervention?”
- It is difficult to distinguish between single aspects of the GBL environment and holistic aspects related to “the processes of analytical measurement/evaluation”. If the evaluation is more holistic and attempts to take everything into account then it is more likely that the evaluation will be general, not particularly detailed or rigorous, leading to other methods being required to establish quality standards. If the evaluation is based on single aspects then the holistic view of the GBL intervention cannot be taken into account. To overcome this problem it is clear that prioritization of particular evaluation areas is required, which leads to the question of “what is the most important objective that the GBL intervention must achieve?”

Approaching the concept of evaluation of GBL from a pedagogical perspective assists in refining the process. The literature review presented shortly shows that there is no general method of evaluation for GBL or simulations. When evaluation does take place the most common method appears to be surveying the participants and certainly the majority of GBL evaluations encountered in the research literature has some form of survey as an evaluation instrument. The largest disadvantage of this is that surveys are sensitive to weak methodological design and can be influenced by particular wording and exogenous factors (Saari, Johnson, McLaughlin, & Zimmerle, 1988).

Previous Evaluation Frameworks and Models

When developing an evaluation framework for GBL, it seems logical to design the framework from a pedagogical perspective as the entire ideology of GBL is using games/simulations to motivate and engage learners, resulting in more effective learning even at a supplementary level. There are very few evaluation frameworks in the literature that specifically address the effectiveness of GBL from this perspective and ask questions such as: Does the GBL environment increase knowledge acquisition? Does it improve learner performance? Does it assist in the formation of metacognitive strategies? The majority of available frameworks are focused on either e-Learning or commercial games such as World of Warcraft.

Two examples of these frameworks are based on Nielsen’s Heuristic Evaluation developed in 1990 (Nielsen & Molich, 1990). Heuristic Evaluation consists of ten recommended heuristics and is supposed to be performed by a small team of evaluators. It is a Human Computer Interaction (HCI) technique that focuses on finding interface usability problems and has been extended with additional heuristics to encompass website specific criteria. The technique has also been expanded and developed to produce a framework for web-based learning (Ssemugabi & de Villiers, 2007) and a framework for heuristic evaluation of Massively Multi-player On-Line Role Playing Games (MMORPGs) (Song & Lee, 2007). One of the main difficulties associated with frameworks
Towards the Development of a Games-Based Learning Evaluation Framework

developed from Heuristic Evaluation is that the quality of a Heuristic Evaluation is dependent on the knowledge of the expert reviewer. By extending frameworks to encompass web-based learning and MMORPGs, suitable reviewers would have to have sufficient knowledge of HCI and games to perform a high-quality evaluation. In addition, from a GBL perspective the main difficulty is that these frameworks do not specifically focus on pedagogy.

Tan, Ling, and Ting (2007) reviewed four GBL frameworks and models including: the design framework for edutainment environments, the adopted interaction cycle for games, the engaging multimedia design model for children and the game object model. According to their results one framework significantly addressed pedagogy and game design: the game object model developed to allow identification of suitable game elements to be supported by valid pedagogical elements (Amory, Naicker, Vincent, & Adams, 1999). The game object model (GOM) has been further developed using theoretical constructs and developments in the literature to become the game object model version II framework (GOM II) (Amory, 2006). This particular framework can be used from both a game design perspective and an evaluation perspective. The original GOM (Figure 1) has several spaces:

- **Game Space** – Embodies all of the components. Components are represented by a square and contain different discrete interfaces that can be either abstract or concrete. The interfaces are displayed within the component. The components are either free standing or part of other components. The inner components inherit all of the outer components interfaces and inner components interfaces are concrete whereas the outer components interfaces are more abstract. The interfaces have been listed in this model from the most important to the least important. The game space component encompasses all of the other components and has the following interfaces: play, exploration, challenges and engagement.
- **Visualisation Space** – This component encompasses the elements and problems components and has the following interfaces:

Figure 1. Adapted from the game object model (GOM) (1999)
Towards the Development of a Games-Based Learning Evaluation Framework

story line, critical thinking, discovery, goal formation, goal completion, competition and practice.

- **Elements** – This component has the following interfaces: fun, graphics, sound and technology.

- **Problems** – This component has the following interfaces: manipulation, memory, logic, mathematics and reflexes.

- **Abstract interfaces** – Promote educational objectives and appear in the diagram in bold. They are the following: fun, critical thinking, discovery, goal formation, goal completion, competition, practise, play, exploration, challenges, and engagement.

- **Concrete interfaces** – Enable the realization of the educational objectives. They are the following: graphics, sound, technology, manipulation, memory, logic, mathematics, reflexes and story line.

The GOM II is divided up into several interrelated interfaces and consists of a number of complex objects developed from contemporary practices and educational theories. The GOM II is a far richer model and one of the main developments is the introduction of a new social space to support the development of on-line communities. The study describes the model as an “idiosyncratic, homological, inclusive ideology and represents one of many ways of seeing educational computer game development. Therefore the model should be viewed as a means of structuring discussions and could easily be reconceived to suit different, or alternative viewpoints.”

Kirkpatrick’s four level framework (1994) (Figure 2), takes pedagogy into account. It was originally developed in 1994 as a framework for evaluating training but it has also been proposed for the evaluation of business simulations as educational tools (Schumann, Anderson, Scott, & Lawton, 2001).

The CRESST model of learning (Figure 3) is composed of five families of cognitive demands and can be used in a motivational learning view for the evaluation of games and simulations (Baker & Mayer, 1999). Each family in the CRESST model is composed of a task that can be used as a skeletal design for testing and instruction. The CRESST model is divided into content specific and content independent variables. Content specific variables include: content understanding and problem solving. Content independent variables include: collaboration/teamwork, communication and self-regulation.

Dondi and Moretti (2007) reviewed Uni-Game (Games-based Learning for Universities and Life Long Learning) and SIG-Glue (Special Interest Group for Game-based Learning in Universities and Lifelong Learning), two projects funded by the European Commission. This led to the development of a ‘classification of games by learning purposes’ and an ‘evaluation framework for assessing games’. The evaluation framework takes into account that “a learning game should be a ‘good game’ through which the player will achieve the stated learning objectives” and covers both pedagogical and technical criteria (Table 1). A reduced version is provided to give an illustration of the content with a focus on the pedagogy, context and evaluation criteria sections:

A further framework specifically for games and simulations that addresses pedagogy is the Four Dimensional Framework (FDF) (de Freitas & Oliver, 2006). The FDF is “designed to aid tutors selecting and using games in their practice. The framework includes: context, learner specification, pedagogy used and representation as four key aspects for selecting the correct game for use in learning practice” (de Freitas, 2006). The FDF is displayed in Figure 4. The four dimensions are not designed to be considered in isolation but all dimensions should be considered as a collective whole. The first dimension of the FDF focuses on context with macro-level factors such as political, economic and historical and micro-level factors such as specific tool, resources and general availability. Context is a highly important factor that
Towards the Development of a Games-Based Learning Evaluation Framework

**Figure 2. Adapted from Kirkpatrick’s four levels for evaluating training (1994)**

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1: REACTION</td>
<td>This level assesses how learners feel about participation in a learning experience. It is an evaluation of learner satisfaction measuring “how those participating in the program react to it”. Kirkpatrick highlights that positive reaction to a training program does not necessarily indicate or ensure that learning will take place, however a negative reaction will almost certainly lead to learning not taking place. General methods of collecting learner reaction data are, for example, satisfaction questionnaires.</td>
</tr>
<tr>
<td>Level 2: LEARNING</td>
<td>Kirkpatrick describes this level as “the extent to which participants change attitudes, improve knowledge, and/or increase skill as a result of attending the program”. Learning is measured by quantifying the differences in attitudes, increase in knowledge and the increase in skills. The measurements are typically taken within the context of a training session as dictated by the learning outcomes. Campbell and Stanley (1963) note that the best experimental methodology for establishing whether learning has taken place is a pre-test, post-test, experimental, control group methodology.</td>
</tr>
<tr>
<td>Level 3: BEHAVIOUR</td>
<td>This level assesses whether the learners are actually applying what they have learned within the training session in their working environment. Kirkpatrick (1998) describes behaviour as “the degree to which learners have changed their behaviour outside of the learning environment because of their participation in the learning activities”. Behaviour can be measured by surveying individuals who have the opportunity to observe the behaviour of the learners in different settings to rate the extent to which the attitudes, knowledge and skills are utilized.</td>
</tr>
<tr>
<td>Level 4: RESULTS</td>
<td>This level measures the final benefits to the company. Kirkpatrick defines it as “final results that occurred because the participant attended the program”. The final results can be in relation to a number of attributes such as increased profit, increased turnover, decrease in number and severity of accidents.</td>
</tr>
</tbody>
</table>

**Figure 3. Adapted from Baker and Mayer’s CRESST model of learning: families of cognitive demands (1999)**

- Learning
  - Content specific variables
    - Content understanding
    - Problem solving
  - Content independent variables
    - Collaboration/teamwork
    - Communication
    - Self-regulation
can enable or impede learning depending on difficulty of delivery. The second dimension of the FDF focuses on particular learner or learner group attributes such as learner level, learning styles, preferences, background and age. Research has indicated that different game types can be used to acquire or learn different skills; for example simulations are good for teaching tactical and strategic planning (Dempsey, Haynes, Lucassen, & Casey, 2002). The third dimension of the FDF focuses on the “internal representational world or diegesis – of the game or simulation” and covers immersion and fidelity, presentation mode and the interactivity. de Freitas and Oliver (2006) emphasise that this is a particularly important dimension of the framework as it highlights the distinction between immersion in the game and the critical reflection process that takes place out with the game. The fourth dimension of the FDF focuses on the learning processes during informal and formal curricula based learning. It is designed to promote reflection of the practitioners in terms of frameworks, models, methods and theories used to support learning practice. The dimension supports differentiated learning provided by the availability of e-Content, e-Assessment and new software tools by considering how learning content is personalised and embedded.

**LITERATURE SEARCH**

This section presents the literature search that was carried out in the Summer of 2008 to identify previous evaluation approaches for games-based learning.

**Method Used to Collect Data**

An extensive literature search was performed by reviewing various electronic databases including: ACM, ABIINFORM Global Database, Academic Search Premier, ScienceDirect, Blackwell Synergy, EBSCO (consisting of Psychology and

---

**Table 1. SIG-Glue quality criteria framework**

<table>
<thead>
<tr>
<th>Target groups and prerequisites identification</th>
<th>Learning objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correspondence of established objectives and objectives that can be reached using the game</td>
<td>Clearly defined objectives</td>
</tr>
<tr>
<td>Context of usage</td>
<td>Context suggestions</td>
</tr>
<tr>
<td>Coherence with the targeted context</td>
<td>Didactic strategy</td>
</tr>
<tr>
<td>Indications of average play time</td>
<td>Incentives and support to motivation</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Coherence of the social and collaborative activity with objectives</td>
<td>Communication and media</td>
</tr>
<tr>
<td>Clear, user-friendly tone and language</td>
<td>Quality of the interaction</td>
</tr>
<tr>
<td>Coherence between media used and established objectives of target group</td>
<td>Evaluation</td>
</tr>
<tr>
<td>Clear identification of evaluation criteria and procedures</td>
<td>Adequate number and distribution of evaluation activity, during the game and at the end</td>
</tr>
<tr>
<td>Type of evaluation activity proposed</td>
<td>Quality of the feedback of the evaluation</td>
</tr>
<tr>
<td>Relevance of evaluation activity and consistency with the objectives and/or the contents</td>
<td>Supporting the reflective process</td>
</tr>
</tbody>
</table>
Towards the Development of a Games-Based Learning Evaluation Framework

Figure 4. Adapted from the four dimensional framework (de Freitas & Oliver 2006)

Towards an Evaluation Framework for GBL

This section presents a compiled evaluation framework for GBL based on the key measurements we identified in the literature search. The previous frameworks reviewed were highly instrumental in attempting to make sure there were no omissions and also assisted in the categorisation process. The only direct similarity is the concept that the categories do not necessarily have to be considered in isolation like the FDF. The highest abstraction of the framework is displayed in Figure 5.

The purpose of the framework is to identify what can potentially be evaluated in a GBL application. The literature review identified existing frameworks of evaluation and the particular attributes that researchers have attempted to measure during a GBL environment or simulation intervention. Drawing on the frameworks in the existing literature and particular identified measurements, GBL can be evaluated in terms of learner performance, learner/academic motivation, learner/academic perceptions, learner/academic preferences, the GBL environment itself.
Towards the Development of a Games-Based Learning Evaluation Framework

Figure 5. Evaluation framework for effective games-based learning

and collaboration between players where appropriate. Like the Four Dimensional Framework (de Freitas & Oliver, 2006) presented earlier, the categories do not necessarily have to be viewed in isolation but as a collective whole depending on what is to be evaluated. This is the greatest direct similarity between the new framework and frameworks identified above. The framework can be used in both a developmental sense to inform design during the implementation and embedding of a GBL application into curricula in a formative evaluation sense and also points to examples of individual analytical measurements already present in the literature for focusing on an evaluation at the end of development in a summative evaluation sense.

### Learner Performance

This category encompasses pedagogy from the perspective of the learner and evaluates aspects of learner performance. Considering that a GBL intervention can be used in both educational institutions and industrial settings the word ‘learner’ is meant to encompass the two settings or indeed any setting in general. The category is primarily concerned with whether there is an improvement in the performance of the learner as a result of the intervention. The improvements are of course inextricably connected to the learning outcomes of the GBL intervention and can be: improvement in knowledge acquisition (procedural, declarative, general), the formation of metacognitive strategies, and improvement in the formation of skills etc.

### Learner/Instructor Motivation

This category is primarily concerned with the particular motivations of the learner for using the intervention, the learner level of interest in participating in the intervention, participation over a prolonged period of time and determining what particular motivations are the most important (Connolly, Boyle, & Hainey, 2006; Connolly, Boyle, & Hainey, 2007b). Are the learners participating extrinsically or intrinsically (Deci & Ryan, 1991)? What particular features of the GBL environment or simulation are the most interesting? Are the learners distracted in anyway? Are the learners willing to use the GBL environment or simulation more than once? When considering Kirkpatrick’s four levels for evaluating the effectiveness of business simulations in particular curricula it is important to identify the motivations that not only apply to the learner but also to the instructor. Therefore, it may be important to identify what motivates the instructors to attempt to assimilate a GBL approach into their curricula.
Towards the Development of a Games-Based Learning Evaluation Framework

**Learner/Instructor Perceptions**

This category mainly encompasses perceptions associated with the learners such as their perception of the overview of time within a game or simulation, how real the game is and its correspondence with reality, for example whether the GBL intervention represents a holistic view of a particular organisation or process, perception of game complexity, advice quality and level of self-reported proficiency at playing games. The category also encompasses the learners’ perception of how the GBL intervention can assist them and whether confusion is experienced. The instructor would also have similar perceptions depending on their particular involvement. If the instructor was simply incorporating content into the GBL intervention or simulation then their perceptions may be more important in terms of whether the intervention is fitting well into the particular context. Perceptions are once again extremely dependent on the learning outcomes and what particular perceptions are considered important in the evaluation criteria.

**Learner/Instructor Preferences**

This category considers learner and instructor preferences during a GBL intervention. Learners like to learn in different ways and have different learning styles (Kolb, 1984) therefore different learners will have different preferences. This category could include: learner preference for media when teaching the material, preference of conventional or GBL training, preference and utilization of particular game features, most preferred positive and negative aspects of the game and preference for different competitive modes (Yu, Chang, Luit, & Chan, 2002). For an instructor, this category could include when to introduce the GBL intervention in their particular course or whether they prefer to teach with the GBL intervention.

**Collaboration**

The GBL evaluation framework is designed to be a set of general guidelines for conducting an evaluation of a GBL intervention. The framework has a particular emphasis on pedagogy, as learning is the most important goal. The framework can be customised to particular requirements that are generally optional depending on what particular analytical measurement is necessary to be indicative of effective GBL. Collaboration is optional depending on whether the game is played on an individual level, cooperative group level, competitive group level or multiple cooperative groups competing against each other. It may very well be the case that collaboration does not require evaluation. If collaboration is to be evaluated, the main ways are through achievement of learning outcomes or particular goals, log files monitoring interaction, mapping team aspects to learner comments, measuring the regularity and
level of collaboration and learner group reflection essays.

**GBL Environment**

This category encompasses all aspects that could potentially be evaluated about the GBL environment. It is the most complicated of all categories as it can be divided into the following five identified subcategories from the literature: virtual environment, scaffolding, usability, level of social presence and deployment. In terms of the actual virtual environment itself the evaluation criteria may be the following: validating the background environment and characters including virtual agent expressiveness (Dugdale, Pallamin, & Pavard, 2006), evaluation of factors with regards to environmental alteration, advice importance within the environment, the context of the environment in terms of real-world decision making support and general game difficulty.

Scaffolding refers to the advice and resources within the environment to support the learner in completing their learning outcomes. Scaffolding can be evaluated through monitoring of appropriate realism, feedback, learner perception of the quality of advice, an expert review of the quality of advice and monitoring of the utilisation of resources and advice. Usability can be analysed by looking at particular task completion times, average task completion times, the ease of the task, the number of errors made while performing a task and the ranking of the tasks by the learners.

Usability can also be evaluated through conversation analysis, correlation of the learner demographics to the susceptibility of the problem that is to be overcome by the GBL intervention. In particular relation to developing a GBL intervention, player reactions to initial and incremental prototypes in an iterative fashion may be monitored to evaluate increased and decreased usability aspects. Level of social presence is to do with the immersion and interaction in the game world. It can be monitored by looking at relationship frequencies, player evaluation of game character personalities, attitude and mood statements towards characters and events in the game indicating a social presence.

Deployment is intended to encompass the most effective method of incorporation of the GBL application into the educational context and can also mean the preference of different gaming conditions; i.e. particular format of delivery in a technical sense and also with regards to embedding the GBL environment into the curriculum.

**Measurements Encountered Specifically Associated with the Learner Performance Category**

Each category in the evaluation framework can be extended and has measurements associated with it. The GBL environment category has already been explored in a previous study (Connolly, Stansfield, & Hainey, 2008). Due to the complexity of the framework this study will primarily focus on measurements associated with the learner performance category, which is based on cognition and divided into knowledge and skills (Wouters, van der Speck, & van Oostendorp, in press). All measurements associated with it are displayed in Table 2.

**CONCLUSIONS AND FUTURE DIRECTIONS**

This chapter has highlighted the requirement for empirical evaluation evidence in the GBL literature and presented a new GBL evaluation framework to help researchers evaluate GBL applications. The learner performance category of the framework has been explored in detail in this chapter (the GBL environment category has been reviewed Connolly, Stansfield, and Hainey (in 2008)). The chapter has highlighted that there is a missing link between recognised methods of
<table>
<thead>
<tr>
<th>Learning outcome</th>
<th>Area</th>
<th>Game</th>
<th>Study</th>
<th>Methodology</th>
<th>Measurements</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognition</td>
<td>Knowledge</td>
<td>KMQuest</td>
<td>Christoph, Sandberg, &amp; Wielinga (2005)</td>
<td>Experimental/control group + pre-test/post-test + survey for metacognitive strategy formation.</td>
<td>Knowledge gained. Metacognitive strategies</td>
<td>The total increase of groups with and without the task model were: declarative knowledge increase - (pre-test 0.51 (SD 0.11), post-test 0.62 (SD 0.10)). Procedural knowledge increase - (pre-test 0.49 (SD 0.13) post-test 0.62 (SD 0.09)). Effect of task value model on learning outcomes was inconclusive. Metacognition was inconclusive.</td>
</tr>
<tr>
<td></td>
<td>Knowledge Management</td>
<td>KMQuest</td>
<td>Christoph (2007)</td>
<td>Three studies were performed. Two experimental + pre-test/post-test + survey for metacognitive strategy formation studies. One Experimental/control group + pre-test/post-test + survey for metacognitive strategy formation.</td>
<td>Knowledge gain. Comparison of metacognitive strategy formation in relation to knowledge gain.</td>
<td>The total increase of the groups without a task value model were: declarative knowledge increase - (pre-test 0.51 (SD 0.11), post-test 0.64 (SD 0.10)). Procedural knowledge increase - (pre-test 0.49 (SD 0.13) post-test 0.62 (SD 0.09)). General procedural increase - (pre-test 0.51 (SD 0.13) post-test 0.59 (SD 0.09)). Specific procedural increase - (pre-test 0.48 (SD 0.16) post-test 0.63 (SD 0.13)). Very little relationship between self reported metacognition and learning measures.</td>
</tr>
<tr>
<td></td>
<td>Knowledge Management</td>
<td>KMQuest</td>
<td>Leemkuil &amp; de Hoog (2005), Leemkuil (2005)</td>
<td>Three studies were performed. One experimental group + pre-test/post-test + additional transfer test (a homework assignment with 10 marks associated with it). One experimental/control group pre-test/post-test + additional transfer test (a homework assignment with 10 marks associated with it). One experimental/control group pre-test/post-test.</td>
<td>The importance of advice within the game. Explicit and implicit knowledge items.</td>
<td>Experimental group knowledge increase. Implicit knowledge increase - (pre-test 3.48 (SD 1.59), post-test 4.14 (SD 1.36)). Explicit knowledge increase - (pre-test 8.93 (SD 2.36), post-test 8.93 (SD 1.98)). Advice within the game was ineffective.</td>
</tr>
<tr>
<td>Topic</td>
<td>Game/Methodology</td>
<td>Methodology</td>
<td>Findings</td>
<td>Notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Management</td>
<td>Incredible Manager Dantas, Barros, &amp; Werner (2004)</td>
<td>Experimental group + pre-test/post-test.</td>
<td>Knowledge acquisition. Interest in subject. Interest in GBL.</td>
<td>100% of participants say that learning occurred. 87% say interest in the subject was increased.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Architectural energy efficiency</td>
<td>Architectural energy efficiency game Cowan (2007)</td>
<td>Two experimental groups and one control group + pre-test/post-test. The post-test was conducted one month after the intervention.</td>
<td>Knowledge gained.</td>
<td>Participants in gaming group were more consistently successful in improving.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemistry and Science</td>
<td>Not specified - Computer Animated Instruction Talib, Matthews, &amp; Secombe (2005)</td>
<td>Experimental/control group + pre-test/post-test.</td>
<td>Knowledge gain in high and low level achievers. Conceptual change.</td>
<td>Greater level of knowledge gain for those students exposed to constructivist animations Mean = 13.6, (SD 3.63) as opposed to traditional techniques Mean = 10.2 (SD 3.90). Heightened conceptual change for high and low levels.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Game</td>
<td>Methodology</td>
<td>Findings</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet Peering</td>
<td>Knowledge gain, team essays and individual essays.</td>
<td>Achieved a positive learning experience in general. No control group was used as the game was designed to reinforce concepts from the lectures. The game was reported to reinforce concepts however new information was not necessarily introduced.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Komisarczuk &amp; Welch (2007)</td>
<td>Experimental group/outcomes + control group + pre-test post-test.</td>
<td>Knowledge gain. The game leads to at least similar learning as traditional methods. Post-test playing = 1.83. Game feels like incidental learning with 99% of students playing at least once more.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Squire, Barnett, Grant, &amp; Higginbotham (2004)</td>
<td>Experimental group + pre-test post-test.</td>
<td>Learning in games-based learning environments. The game leads to similar or equal learning outcomes compared to traditional methods. If playing the game voluntarily feels similar to incidental learning.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ebner &amp; Holzinger (2007)</td>
<td>Experimental group + pre-test post-test.</td>
<td>The game leads to at least similar learning as traditional methods. Post-test playing = 1.83. Game feels like incidental learning with 99% of students playing at least once more.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheng et al (2007)</td>
<td>Game group/existing material group/tutorial group + pre-test post-test.</td>
<td>Participants in the game condition performed better. Existing training material group (pre-test = 0.66, post-test = 0.74). Game group (pre-test = 0.65, post-test = 0.80).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheng et al (2007)</td>
<td>Game group/existing material group/tutorial group + pre-test post-test.</td>
<td>On a scale of 1-5 (1-not at all and 5-definitely), the average rating for the perception of the game teaching new process knowledge was 2.5. On a scale of 1-5 (1-not at all and 5—very much so), the average rating for whether the participants believe that the game teaches software engineering was 3.7.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oh &amp; Van der Hoek (2005)</td>
<td>Experimental group + post-test survey.</td>
<td>If the game teaches software engineering processes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software Development</td>
<td>SimJavaSP</td>
<td>Experimental group + post-test survey</td>
<td>If the games teaches software development.</td>
<td>67% of respondents perceive the game as being able to teach software development lifecycles.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>----------</td>
<td>--------------------------------------</td>
<td>--------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>Metalloman</td>
<td>Four experimental groups + post-test survey for usability.</td>
<td>Effects of interactivity and media richness on learning.</td>
<td>No significant knowledge gain.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>ASTRA EAGLE</td>
<td>Two experimental groups (cooperative condition and competitive condition)/one control group + two pre-tests/post tests. One associated with performance and another associated with attitudes towards the subject area.</td>
<td>Cooperative game play would increase math performance.</td>
<td>Results from a post hoc pair-wise comparison on the adjusted post-test means indicated no significant difference in math performance between cooperative and competitive conditions. Adjusted mean of cooperative condition = 61.2 Adjusted mean of competitive condition = 59.9. Both conditions performed significantly higher than the control group. Adjusted means for the control group condition = 55.3.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>River City</td>
<td>Two experimental groups (guided social constructivist group (GSC) and expert modeling and coaching group (EMC))/control group + two pre-tests/post-tests. One for understanding of content and one for attitudes.</td>
<td>Efficacy of test subject. Understanding and content knowledge.</td>
<td>Increase in Biology knowledge in test group by 32% – 35%. 17% increase in control group. Inquiry content results improved more in the control group than the other two experimental groups. Control group increased by 20%. Guided social constructivist group increased by 18% and the expert modeling and coaching group increased by 16%.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geography</td>
<td>VR Engage</td>
<td>Two studies performed. Two experimental groups/two control groups + two pre-tests/post-tests</td>
<td>Improvement in mistakes made in relation to different levels of academic performance (poor, mediocre, good).</td>
<td>Consistent improvement in the percentage of mistakes made. For students of previously poor academic performance the experimental group = 48.97% (SD 10.94), the control group = 31.57% (SD 7.72). For students of previously mediocre academic performance the experimental group = 38.50% (SD 10.06), the control group = 31.64% (SD 5.08). For students of previously good academic performance the experimental group = 33.80% (SD 9.66), the control group = 32.84% (SD 9.67).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>Game</td>
<td>Author(s)</td>
<td>Methodology</td>
<td>Findings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>--------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>River City</td>
<td>Nelson (2007)</td>
<td>Two experimental groups (one with moderate guidance and one with extensive guidance) + control group + pre-test/post-test.</td>
<td>Knowledge gain with different levels of guidance: No Guidance, Extensive Guidance, Moderate Guidance. However, advice was ineffective. Moderate Guidance group showed predicted gain of 0.45. No Guidance group had a gain of 0.14, while Extensive Guidance had a gain of 0.13, identical to the control group.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology - genetics/epidemics</td>
<td>Virus and Live Long and Prosper (LLAP)</td>
<td>Klopfer et al. (2004)</td>
<td>Four experimental groups: one tags for LLAP, one palms for LLAP, one tags for Virus and Live, one palms for Virus and Live + two pre-tests/post-tests.</td>
<td>Self-assessment of learning about content, technology, experimental design and whether the participants believed that the technology positively impacted learning. Participants rated their learning of content (mean = 3.64), technology (mean = 3.72) and experimental design (mean = 3.64) highly. They expressed strong agreement (mean = 3.95) with the statement that the technology positively impacted their learning.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Childhood violence prevention</td>
<td>Invasion of the Foodles</td>
<td>Fontana &amp; Becker-</td>
<td>Experimental/control group + pre-test/post-test.</td>
<td>Learner knowledge towards human behaviour and conflict resolution strategies. Significant change was not recorded for most of the areas of human behaviour and conflict resolution in relation to knowledge.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>ASTRA EAGLE</td>
<td>Ke (2006)</td>
<td>Three experimental groups (individualistic, competitive and cooperative) + pre-test/post-test for performance and attitudes.</td>
<td>Knowledge gain with different goal structures: cooperative, competitive and individualistic. There was no significant effect of gaming goals structures. Adjusted post-test means for each group are: Individualistic = 60.5, competitive = 59.2 and cooperative = 60.7.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Process Management</td>
<td>Containers Adrift</td>
<td>Experimental group + post-test.</td>
<td>Ability to drawing a visual representation, negotiating parameters, build and running a simulation model and assessing design performance.</td>
<td>The tool was actively used and positive observations were made.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------</td>
<td>---------------------------------</td>
<td>-----------------------------------------------------------------</td>
<td>-----------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Management</td>
<td>Incredible Manager</td>
<td>Experimental + pre-test/ post-test.</td>
<td>Project Management skills.</td>
<td>Skills gained. 100% of participants believed that the game increase their project management skills.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>Zombie Division</td>
<td>Experimental + pre-test/ post-test.</td>
<td>Learner fluency. Learner change in exercising a skill overtime.</td>
<td>Inconclusive.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>SKILLS ARENA</td>
<td>Two experimental groups, one handheld game group and a delayed group tested 10 weeks after the intervention + two pre-tests/post-tests. One for math skills the other for attitudes towards the subject area.</td>
<td>Mathematical skill acquisition.</td>
<td>The handheld game group outperformed those in the card game group. Handheld group - pre-test mean = 37.06 (SD 14.31), post-test mean = 44.71 (SD 12.62). Card game group - pre-test mean = 37.05 (SD 15.58), post-test mean = 39.95 (SD 17.00).</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
evaluation and the GBL literature with the majority of studies surveying participants with very little mention of formative or summative evaluation methodologies.

Future work and validation of the evaluation framework will firstly include a pilot study evaluation of a software development game designed to teach requirements collection and analysis at tertiary education level (Connolly, Stansfield, and Hainey, 2007a). Other categories of the framework will be refined and investigated in depth next as we have primarily focused on the learner performance category of the framework in this study. The framework will also be made available to researchers who are developing and evaluating GBL interventions for the purposes of refinement and also to assess the broader question of whether the framework was beneficial in directing their research interests.

REFERENCES


Towards the Development of a Games-Based Learning Evaluation Framework

Games. Information Technology in Childhood Education Annual, (pp. 49–62).


Towards the Development of a Games-Based Learning Evaluation Framework