

8 Evaluating the educational effectiveness of games

This chapter considers how the educational effectiveness of computer game-based learning applications could be evaluated. With any new teaching technique it is important to consider and test what impact it might have on the learning experience, and ways in which computer game-based learning could be evaluated are considered here. This research aimed to evaluate the effectiveness of the two applications that were produced, the Time Capsule and the Pharaoh's Tomb, to provide evidence that they met their intended learning outcomes and also to compare the learning experience of the students using these two different examples of game-based learning.

The first section of the chapter considers ways to assess educational effectiveness, evaluating learning and engagement. There is a consideration of ways in which learning can be measured, how it is commonly evaluated in comparative experiments examining the differences between two educational interventions, and the advantages and disadvantages of these methods of measurement. The concept of engagement is considered and defined, and the relationship between engagement and learning is discussed and a rationale for using engagement as an indicator of educational effectiveness is presented; the development of a questionnaire to develop post-experiential engagement is also described.

The final two sections of the chapter describe additional studies that were carried out to provide evidence of the educational effectiveness of the two applications. Section 8.2 presents a small comparative study that was carried out to compare the self-reported learning of students using the original face-to-face version of the Time Capsule activity with that of students using the online version. Section 8.3 provides an analysis of the transcripts of students using the Time Capsule and Pharaoh's Tomb activities during the pilot phase of the comparative experiment (described in the following chapter) to provide evidence that the activities support achievement of the intended learning outcomes.

8.1 Evaluating educational effectiveness

One of the research questions of this thesis is concerned with comparing the learning and educational experience resulting from different types of game-based learning application. In order to do this, it is important to consider the alternative ways of measuring the learning resulting from an educational intervention.

The typical way to measure learning from a unit of study is through the assessment for that unit, which should be constructively aligned with the learning outcomes for that unit (Biggs, 2003). In a study involving a larger intervention, a comparison of assessment scores would be a potential way in which to compare learning; however, in the case of the Time Capsule and Pharaoh's Tomb, each relates to only one hour's worth of study, which would only make up a fraction of an assessed course. Therefore the effect of either game on the overall assessment score for a unit is likely to be negligible, and for this reason it was decided that using the assessment score was inappropriate in this instance.

In experimental design studies, the effectiveness of an educational intervention is often measured using a pre-test followed by the intervention, followed by a post-test; differences in the pre- and post-test scores can indicate that the intervention has caused different levels of learning in the target group compared to the control group. This technique has been used for a number of studies on game-based learning; for example, Ebner and Holzinger (2006) tested theoretical knowledge in chemical engineering, Kambouri and colleagues (2006) evaluated basic literacy skills, and Sung and colleagues (2006) examined children's understanding of taxonomic concepts.

Despite this being a common way of evaluating learning in studies of educational interventions, it was decided not to use a pre- and post-test in this research for a number of reasons. The intended learning from the game-based learning studies cited in the previous paragraph was knowledge-based and could be tested quickly by use of a questionnaire; however, the intended learning outcomes from the Time Capsule and Pharaoh's Tomb focused on the development of collaborative skills, which are higher-level cognitive skills than

the acquisition of knowledge (Bloom, 1956) and also involve the development of behaviours and attitude, and would therefore be less appropriate to be tested with a simple questionnaire. While it would be possible to design a pre- and post-test for these type of skills that also examined retention over time and application of learning to other contexts, it would involve undertaking comparable team tasks (which in itself might bias the results by acting as a learning as well as an evaluative activity) and would be time-consuming and impractical to administer. It was thought to be too difficult, if not impossible, to persuade students to give up extra time for this testing. Difficulties with getting students to co-operate with the pre-/post-test model because of the extra work required on their part are not uncommon. Squire (2005), for example, could not persuade the students in his sample group to complete a pre-test at all.

However, despite the inapplicability of the pre-test/post-test design to this research, it was felt that some indication of learning would be useful to triangulate with other findings. A 12-question questionnaire was designed to enable the students to self-report on what they felt they had learned from undertaking the activity (see Appendix 18). This questionnaire asked the students to evaluate any improvement in their own skills in a number of areas, based on the learning outcomes of the application. Two additional questions were added referring to skills that the students were not intended to learn during the activity, which were used to examine the validity of the questionnaire (see Section 9.3).

The self-report questionnaire was not seen, in itself, as a reliable enough indicator of learning to be used as a sole measure for comparison between the two game-based activities. So instead of evaluating learning from the game directly, it was felt that it was more appropriate in this situation to evaluate engagement with the game, as there is evidence that higher levels of engagement with a learning activity lead to increased learning from it. Preece and Jacques (1995) argue that designing interactions to be engaging can encourage and facilitate learning, and Lepper and Malone (1987) provide evidence that there is a link between intrinsic motivation to learn, engagement and instructional effectiveness. It is important, however, to distinguish between engagement with a game, and engagement with the intended learning from the

game. Ideally, educational games should be designed so that the game outcomes are aligned with the learning outcomes so that engagement in the game supports learning. Although the match will not always be perfect, applying the educational design guidelines described in Chapter 5 should help to ensure that this is the case.

For the reasons described, it was decided to use engagement as the primary indicator of educational effectiveness, and to develop a tool to measure relative engagement with a learning experience. The development process is described in the next sub-section.

8.1.1 Evaluating engagement

The most common methods employed to measure engagement in educational settings are the use of questionnaires, and measurements such as time-on-task or attendance rates (e.g. Chapman, 2003). Other techniques include analysis of facial expressions and body language (Hughey, 2002), observations (Read et al, 2002) and voluntary time on task (Virvou et al, 2004).

Since this study is concerned with psychological engagement it was felt that the only way to get a detailed picture of how an individual perceived an activity was by asking the individual him- or herself and trying to understand the individual perceptions of an experience. A questionnaire was used because this was more feasible than interviews in terms of the time available for each student to take part in the evaluation, and using a quantitative measure would enable the use of quantitative statistical analysis to examine a difference in engagement between two activities. The disadvantage of measuring engagement in this way is that there would be no qualitative data available to gain a deeper understanding of the nature of engagement; however, this was unavoidable given the problems of access to students.

In reported examples where engagement has been tested with self-reporting scales, there is often no evidence of the systematic development of the rating instrument (e.g. Davies, 2002) or the method of analysis is not appropriate for ordinal data (e.g. Chapman et al, 1999). In fact, no examples could be found in the literature of engagement questionnaires that had been rigorously developed

(this is not to say that the instruments used were unsound, just that there was no explicit mention of their genesis). For this reason it was decided to develop an original engagement inventory to ensure that the development process, as well as the application and analysis techniques, were sound and appropriate.

To measure the level of psychological engagement with an activity, it was felt that a self-rating questionnaire was a practical, relatively quick and simple to administer, yet not too intrusive, method of measurement. In examining the factors that make up the concept of engagement, Flow theory (Csikszentmihalyi, 1992) was used as a central basis, but acknowledging that flow is an extreme form of engagement and that it is possible to be engaged while not actually in a state of flow. The work of Malone (1980a; 1980b), in terms of challenge, curiosity and control, is also drawn upon. Also taken into consideration are the results from the interviews described in Chapter 4 on factors that appear to universally motivate or demotivate, backed up by adult learning theory regarding adults' motivations for learning (Knowles, 1988). Based on these theories, engagement was hypothesised to be made up of five separate factors; these are shown in Table 8-1.

Factor	Description	Origin
Challenge	The most complex of the factors, consisting of: the motivation to undertake the challenge; clarity as to what the challenge involves; and a perception that the challenge is achievable.	Csikszentmihalyi (1992) Malone (1980a; 1980b) Chapter 4 interviews.
Control	The fairness of the activity, the level of choice over types of action.	Csikszentmihalyi (1992) Malone (1980a; 1980b)
Immersion	The extent to which the individual is absorbed in the activity.	Csikszentmihalyi (1992)
Interest	The intrinsic interest of the individual in the activity or its subject matter.	Malone (1980a; 1980b) Chapter 4 interviews
Purpose	The perceived value of the activity, whether it is seen as being worthwhile and whether feedback is perceived as having a value.	Chapter 4 interviews Knowles (1988)

Table 8-1: Factors hypothesised to increase engagement

It was decided to use a Likert scale questionnaire for a number of reasons. This type of scale is widely used so would be familiar to participants, is relatively straightforward to develop (Robson, 2002) and has established statistical analysis techniques (Greene & D'Oliveria, 1993). It was decided to use a five-point scale because this was considered to provide a meaningful level of discrimination (e.g. between 'agree' and 'strongly agree') without forcing the participant to have an opinion.

A number of potential questions for each factor was generated by the researcher, which were then reviewed by three individuals with experience of developing attitude scales for clarity, ambiguity and language used. The original questions were then revised and refined and an original questionnaire was developed, which contained 42 questions (see Appendix 19 for the questions).

The questionnaire was then piloted by asking participants to play one or more of five games drawn from those examined in Chapter 5 (see Appendix 5). The five games used are shown in Table 8-2.

Activity	Description	Number of responses
Bookworm	Word-building game, testing spelling and vocabulary.	15
The Mystery of Time and Space	Point-and-click adventure game based around puzzle-solving, lateral thinking and investigation.	15
Laser beams	A series of spatial puzzles involving strategy and planning.	11
Typer shark	Typing arcade game testing speed and hand-eye co-ordination.	13
I-sketch	Multi-user picture drawing and guessing game, requiring lateral thinking and social skills.	11
	Total	65

Table 8-2: The games used as part of the testing process for the engagement questionnaire

This questionnaire was piloted with 33 participants, each of whom played at least one game (providing 65 responses in total). The testers were recruited by word of mouth and through email mailings forwarded by colleagues, and were adults who considered themselves to be computer literate. Participants were given instructions by email, asked to play a specific game for 15–20 minutes in their own time and to complete and return the questionnaire immediately afterwards. They were also asked if they would like to play another game, and most participants chose to play more than one game. Games were chosen for testing that were considered to have educational potential, and a range of different types was chosen to elicit variation in response.

In order to generate the final questionnaire for post-testing with students in the final comparative study, the overall responses in each group were first examined to determine whether they were in fact measuring the factor that it was hypothesised they were measuring. This was done by using the SPSS statistical package to first transform the data so that the results of negative questions were reversed, and then to calculate the Kendall Tau rank correlation coefficient (a statistic used to measure correlations between ordinal data) for each pair of questions within each hypothesised factor; a one-tailed test is used as it is hypothesised that the correlations will be positive. A summary of results is shown in Table 8-3; and a more detailed breakdown of the correlation data can be found in Appendix 20.

The questions that did not correlate with all of the other questions in the group at a 0.01 level of significance were removed from further analysis. These tended to be those that were poorly worded, more ambiguous or less clear whether they were a positive or negative influence on engagement.

In order to reduce the number of questions further to an appropriate size for the final questionnaire, the Discrimination Power (DP) of each question was calculated. The Discrimination Power, as described in Robson (2002), is the ability of the question to distinguish between the responses of the upper quartile of respondents overall and the responses of the lower quartile, that is, the degree to which the response to an individual question indicates the overall response to the questionnaire.

Factor	Questions (those underlined do not correlate)
Challenge (motivation)	I wanted to complete the activity I wanted to explore all the options available to me I did not care how the activity ended
Challenge (clarity)	I knew what I had to do to complete the activity The goal of the activity was not clear The instructions were clear I did not find it easy to get started <u>I found using the application easy to learn</u>
Challenge (achievability)	I felt that I could achieve the goal of the activity I had all the things I required to complete the activity successfully I had a fair chance of completing the activity successfully <u>I found the activity difficult</u> I found the activity frustrating From the start I felt that I could successfully complete the activity <u>The activity was challenging</u>
Control	<u>I had lots of choices to make during the activity</u> The types of task were too limited It wasn't clear what I could and couldn't do The activity was too complex The activity would not let me do what I wanted I could not tell what effect my actions had <u>I had lots of potential options available to me</u> I could not always do what I wanted to do
Immersion	I found the activity satisfying I felt absorbed in the activity I felt that time passed quickly <u>I worried about losing control</u> <u>I felt emotion during the activity</u> <u>I felt self-conscious during the activity</u> I felt excited during the activity
Interest	<u>I had to concentrate hard on the activity</u> <u>I knew early on how the activity was going to end</u> I found the activity boring I was not interested in exploring all of the environment I did not enjoy the activity The activity was aesthetically pleasing
Purpose	The activity was pointless The feedback I was given was not useful <u>I did not receive feedback in enough detail</u> I was given feedback at appropriate times It was not clear what I could learn The activity was worthwhile

Table 8-3: The questions hypothesised to measure each factor of engagement; underlined questions are those that do not correlate with all others in the group at the 0.01 level of significance

The questions with the highest discrimination powers in each factor were selected for the final scale with one exception where it was felt that a question with a slightly lower DP was more appropriate because it was less specific (see Appendix 21). Six questions were selected to measure challenge (two from each challenge sub-factor) and an additional three questions were selected from each of the other factors for the final scale, making 18 questions in all. This was felt to be a compromise between the greater reliability gained as the number of questions used to measure each factor is increased, and the decreased propensity for the respondents to complete the questionnaire correctly as the overall number of questions increases. All the questions were reviewed a final time and a small number were altered slightly or reversed to aid clarity. The final questionnaire is shown in Appendix 22.

As a matter of interest, the relative levels of engagement for the five games used for testing were examined, using only the 18 questions from the final scale. The average scores for engagement for each activity can be seen in Table 8-4; it is important to recognise that as absolute values these scores are meaningless but they can be used to compare levels of engagement in different activities.

Activity	Engagement score		
	Min	Mean	Max
Bookworm	40	62	77
The Mystery of Time and Space	30	60	84
Laser beams	38	58	75
Typer shark	53	70	82
I-sketch	29	51	73

Table 8-4: Maximum, mean and minimum engagement scores for each of the five games tested

It is interesting to note that it is the arcade-style game that that appears to be the most engaging, while the collaborative drawing game appears to be the least; but of course, it is difficult to draw any genuine conclusions regarding engagement and game type as the differences could be due to a number of other factors including the design of that specific game, for example the graphic quality, or interaction speed. It is also worth noting that for some games the

range of opinion seemed to be much greater than for others, indicating more polarity of opinion; the *Mystery of Time and Space*, for example, while having a similar average score to *Bookworm* shows much more variation in opinion.

The 18-question engagement questionnaire was used to evaluate the difference between the Time Capsule and the Pharaoh's Tomb in the set of comparative experiments that formed the final part of this research, and is described in the following chapter. The next section of this chapter describes an evaluation that was carried out to compare learning between the face-to-face and online versions of the Time Capsule activity.

8.2 Evaluating learning from the Time Capsule

The opportunity arose to use the Time Capsule exercise with a group of final year undergraduate marketing students as the introductory session to a module on marketing strategy, in which the students had to work in teams. This provided an excellent opportunity to examine any differences in perceived learning between students who used the online version of the activity and those who used the face-to-face version.

As this session was part of an existing course structure, the time available for evaluation on top of the time spent using the activities was very limited, so the only available option was a short self-perception of learning questionnaire immediately after students had undertaken the Time Capsule activity. The questionnaire was made up predominantly of closed questions, but also provided the respondents with the opportunity to make additional comments; it is shown in Appendix 23. Students were informed about the nature of the research and the evaluation before the session and given the opportunity not to complete the questionnaire if they wished, although they were still required to take part in the activity as it was a required part of their course of study.

A total of 60 students participated in the evaluation. Students were randomly placed into groups, with 17 using the online exercise and 43 using the face-to-face version. There was a limited number of computers available for the session, which is why the majority of students used the face-to-face activity. Fisher's exact statistical test was used to evaluate if there was a significant

difference between the responses of students using the online and face-to-face versions for each of the questions. A X^2 test was not considered appropriate because even though the data were nominal, that is, the students could be categorised as either agreeing or disagreeing with each of the statements (Greene & D'Oliveria, 1993), in the case of all questions, either a cell in the contingency table has an expected frequency of less than 1 or over 20% of cells have an expected frequency of less than 5, so X^2 is not applicable (Field, 2005). Therefore, Fisher's exact test is used instead of X^2 in this situation, as it is an appropriate test when expected values are low (Everitt, 2002); and a two-tailed test is used because it is only hypothesised that the variables (i.e. experimental group and perceived improvement) are related, not in which direction. The results of the statistical analysis, as well as a summary of the responses from each group are shown in Table 8-5 below.

Question	Result	F-to-F agree (%)	Online agree (%)
I understand how to make good decisions as part of a group.	All agree	100%	100%
I am more aware of what makes communication effective.	$p=0.676$	82%	88%
Constructive controversy is a good way to make decisions.	$p=0.448$	76%	86%
I understand what makes a group effective.	$p=0.317$	88%	95%
I will be better able to communicate with others in the future.	$p=0.530$	65%	74%
Group reflection is important for effective groups.	$p=0.206$	76%	91%
I will be better able to contribute to group decision-making in the future.	$p=1.000$	76%	76%
I appreciate the benefits of collaborating with others.	$p=0.283$	94%	100%
I can now contribute better to make group work more effective.	$p=0.099$	59%	81%
I enjoyed the exercise.	$p=0.283$	94%	100%
I found the instructions straightforward.	$p=0.393$	82%	91%

Table 8-5: Comparison of learning in the face-to-face (F-to-F) and online Time Capsule activities

These results show that there was no significant difference in the responses to any of the questions from either of the two groups, so there does not appear to be any difference in learning between the two groups. However, there is evidence that the majority of students perceive that they could meet the learning outcomes of the exercise (although it is not clear from the wording of the questions whether they perceive this is because of taking part in the activity; this was resolved in later versions of the questionnaire) and there is stronger evidence that the students found this to be an enjoyable activity to undertake. Although this was not a strictly rigorous trial, it does provide an indication that the Time Capsule is a workable activity within the boundaries of a real teaching situation and that there is no difference in self-perceived learning between the online and face-to-face versions of the activity.

In the final section of this chapter, additional evidence of learning through the Time Capsule and Pharaoh's Tomb activities is provided through analysis of the transcripts from the first pilot study.

8.3 Evidence of learning from transcripts

In order to evaluate the differences in engagement and self-reported learning between the Pharaoh's Tomb and the Time Capsule applications, a comparative experimental study was carried out, and this is described in the following chapter. Before the main study, however, two smaller pilot studies took place; the first of these involved student volunteers and provided the opportunity for the collection of the transcripts, which was not possible when the games were used in real teaching situations because the transcript data were not collected automatically but had to be copy-and-pasted at the end of the session – in the actual teaching situations students had left the game environment before this was possible. It was also felt that if students knew their conversation data were being collected for the main experiment then this might affect their behaviour online and impact upon their experience.

In total, six transcripts were available, three from sessions with participants using the Pharaoh's Tomb and three from the Time Capsule. These transcripts were analysed to see whether there was any evidence of behaviours during the games that would indicate that the intended learning outcomes were being met.

The learning outcomes for both of the activities were:

1. To know what a group is and to be aware of elements that make a group effective.
2. To appreciate the benefits of working as a group and be able to communicate and collaborate successfully with others.
3. To be able to work together to problem-solve and reach effective decisions.

The transcripts were analysed to see if examples could be found of behaviours that supported group effectiveness such as agreeing group goals, friendliness, openness and supporting one another; behaviours that support effective communication such as taking ownership of feelings, asking for feedback, describing behaviour without evaluating and behaviours that support problem-solving and effective decision-making, such as valuing all suggestions, negotiation, compromise and debate around problem-solving. Examples of transcripts of the Time Capsule and Pharaoh's Tomb can be found in Appendices 24 and 25. As well as analysing the contents of the transcripts, a quantitative analysis was carried out to examine the levels of interaction and the contribution rates of the different participants in each game. The results are shown in Table 8-6 below.

Game instance	Word count	Sentences			
		Player 1	Player 2	Player 3	Total
Time Capsule 1	1204	25	46	40	111
Time Capsule 2	622	49	29	54	132
Time Capsule 3	1464	38	33	58	129
Pharaoh's Tomb 1	1684	146	98	55	299
Pharaoh's Tomb 2	1104	30	67	61	158
Pharaoh's Tomb 3	1471	84	89	60	233

Table 8-6: Interaction and contribution rates in the pilot study using the Time Capsule and the Pharaoh's Tomb

It can be seen from this table that in all six cases, all three of the players participated in the game and contributed to the discussion to varying degrees. The total number of statements tended to be higher in the Pharaoh's Tomb but each statement tended to be shorter than in the Time Capsule, this is possibly

because the Pharaoh's Tomb provides more opportunities for interaction with the environment so there is less focus relatively on the communication, whereas the Time Capsule offers limited interaction so there is more focus on negotiation.

An in-depth analysis of the transcripts provided additional evidence that both games were fostering the types of group skills intended. Although this was not a strictly scientific analysis, and can clearly not show whether these skills were present beforehand or have been developed during the activity, the following excerpts from the transcripts do provide some evidence that the games actually did encourage the types of group behaviours that were intended.

There were a number of examples of the players exhibiting behaviours that support group effectiveness, such as clarifying the ground rules and strategies for achieving the group goals:

```
Hilary: we can either pick 2 personal objects each
Hilary: or 6 that are related to the area, my option of
course
Hilary: heritage is important
Hilary: and badgers
Catherine: ok so we pick 2 each at the moment
Felix: I have a few ideas as to what should go in, so let's
see what two we can come up with
Hilary: pick 2 each the and see what we come up with

Felix: Hilary, tell me about the capsule
Hilary: we have £1000 pounds to select 6 items
Hilary: I think it is one of the mayor's ideas
Catherine: we do, can we strike any off of the list straight
away? are there any we believe strongly should or shouldn't
be in the capsule?
```

There were also examples of supportive and friendly behaviour:

```
Jack: i have made a flute with the knife and the reeds
Jack: i will play to the snake
Phil: excellent
Mike: happy days,
Phil: that was clever

Rose: I think thats it!
Dave: bingo!!!!
Dave: well done
Rose: great, have you got a bit of the scarab?
Amir: well done everyone
```

The transcripts also showed that the players exhibited a number of behaviours to support collaboration, such as expressing opinions in an open but non-aggressive manner, and taking responsibility for feelings:

Felix: I am of the opinion that things like bones and historical artefacts should be in museums, but that's just my view

Catherine: yes

Hilary: ah mmmmmmm er good point

Catherine: Who's chosen 3 items?

Hilary: me

Catherine: Thta only leaves me with one choice. I feel thats not very fair like.

Examples could also be found of the participants working together and supporting one another to achieve tasks as part of a team:

Bob: Are you in maze Sam?

Bob: ok

Sam: yea

Jim: Can we help?

Rose: Where is the vase?

Dave: in the big room

Rose: found it!

Rose: we could fill it with water, we havn't used the bucket yet

Rose: I see someone has

Dave: already there lol!!!

There was also evidence that the players were exhibiting behaviours to support problem-solving and decision-making, such as negotiation and compromise:

Felix: We need another thing. Why don't we replace rocks with something also worth £50?

Felix: rocks are cheap and also boring

Catherine: yes the beer

Hilary: no the fudge

Felix: How about the magazine which will not smell bad when it goes off

Felix: But is also about food

Felix: If we go for the yearbooks, we've got £100,00 left for the pub!

Catherine: perfect!

Hilary: I think Badgers are under-represented

Felix: So?

Hilary: ok i will agree

Catherine: while i see that badgers are lovely animals, i still dont see how they are really unique to us, im happy with what we have
Felix: Maybe there's a picture of a badger in the yearbooks.
Hilary: I love badgers
Catherine: yeh, probably
Hilary: badger stew...mmm
Catherine: maybe we could make sure there is? that would work

Other examples include working together to try to solve problems and to make suggestions on how to approach tasks within the games:

Dave: think we need a key???
Amir: i have a key
Rose: There is a locked door here, do you want to try to open it with your key?
Amir: how
Dave: go to the far east room with the key

Sam: how do i apply the whip to the snake?
Jim: drag the whip on screen
Bob: have you tried dragging and dropping it?
Sam: yeah no luck!!

From this small sample, it appeared that the types of behaviours exhibited in the case of people playing the Pharaoh's Tomb and the Time Capsule were different, which is almost certainly due to the differences in design and goals, with the Time Capsule focusing more on negotiation and the Pharaoh's Tomb more on problem-solving. However, there are certainly examples from each of the six transcripts examined that a range of team-building and collaborative behaviours are taking place when the game-based activities are being used, which are clearly related to the anticipated learning outcomes of the activities.

This chapter has aimed to consider the different ways in which the educational experiences of the students using the Time Capsule and Pharaoh's Tomb activities can be evaluated. Measurement of learning was first considered but would be difficult for a number of reasons discussed in Section 8.1; however, a self-rated perception of learning questionnaire was developed. The key measure of educational effectiveness used here was the measurement of post-experiential engagement and the rigorous development and piloting of a Likert questionnaire is described.

In addition, two small studies are discussed, which try to provide some additional evidence that the Time Capsule and the Pharaoh's Tomb actually support the type of learning that is intended. The next chapter builds on these studies and describes the large-scale comparative study that was carried out to compare the two game-based activities using the questionnaires (self-perceived learning and engagement) described in this chapter.