Narrative and Content Combine in a Learning Game for Virtual Heritage

Jeffrey Jacobson,[1] Kerry Handron,[2] and Lynn Holden[1]

[1] PublicVR, Boston. USA.
[2] Carnegie Museum of Natural History, Pittsburgh. USA.

Abstract

We find the “game” paradigm an efficient and familiar way for students to interact with virtual heritage content (3D models, etc.) and with the supporting information. The student/player works though the interactive narrative, striving towards goals which matter in the context of the content itself. The rewards and challenges must be part of the content, and not some visual sugar or meaningless allocation of “points” for the experience to be effective. Accordingly, *Gates of Horus* is a virtual heritage learning game based on an ancient Egyptian temple. The student learns from a virtual priest, who also challenges the student to demonstrate knowledge. The student’s reward is entry into successively deeper and more mysterious parts of the temple. Several categories in a knowledge post-test demonstrated that *Gates of Horus* is an effective learning tool ($p < 0.0016$, $p < 0.0044$, and $p \approx 0$).

***Keywords**: virtual heritage, game, learning, education, research, Egypt, school, museum, experiment.*

1 INTRODUCTION

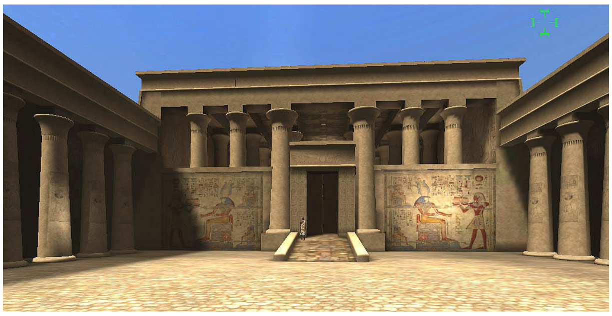


Figure 1. *Gates of Horus: the courtyard.*

Gates of Horus is an educational game based on our Virtual Egyptian Temple.[1] The student assumes the role role of a young prince being schooled in the mysteries of the temple by asking the high priest for explanations of temple features. In turn, the priest tests the student’s understanding by asking his own questions, and rewards accurate responses by allowing the student to go further into the temple. The student “wins” by answering a final set of questions in the inner sanctuary, which elicits spoken praise from the divine image of the temple god (Horus).

The game works equally well on a desktop computer and in visually immersive (dome) displays, from one-person mini-domes to all-digital planetarium theaters. The temple is the subject of a human-guided, virtual-guided tour at the Earth Theater, a digital dome, at the Carnegie Museum of Natural History in Pittsburgh. *Gates of Horus* software and source materials are all free to the public and require only low-cost equipment.[2]

We constructed *Gates of Horus* (fig. 1) for our case study on learning activities as game-like narratives. As with other forms of activity-based learning, the student works through specific challenges to achieve well-defined goals. The student must construct new personal knowledge by understanding the lesson in terms of his or her prior knowledge[3] and perception of the world.[4]

The game responds to the student’s performance as s/he progresses. This makes game-based learning a special case of adaptive media, where the software tries to adapt to the needs of the user.[5] Furthermore, the adaptivity in games gives the student some control over the learning process, the central tenant of Constructivist learning theory.[6] Many learning applications based on virtual reality claim a constructivist approach to learning,

[1] Jeffrey Jacobson and Lynn Holden, “The Virtual Egyptian Temple,” paper presented at the World Conference on Educational Media, Hypermedia and Telecommunications, Montreal, Canada, June–July, 2005. http://publicvr.org/downloads/Jacobson2005e.pdf.

[2] PublicVR, http://publicvr.org.

[3] B. S. Bloom, *Taxonomy of Educational Objectives* (Chicago, New York: Longman, Inc., 1956).

[4] William Winn, “Learning in Artificial Environments: Embodiment, Embeddedness and Dynamic Adaptation,” *Cognition and Learning* 1 (2003): 87–114.

[5] Peter Brusilovsky and C. Peylo, “Adaptive and Intelligent Web-based Educational Systems,” *International Journal of Artificial Intelligence in Education* 13 (2003): 156–169.

[6] D. H. Jonassen, “Constructivism and Computer-mediated Communication in Distance Education,” *American Journal of Distance Education* 9 (2) (1995): 7–26.

especially when goal-seeking activities are effectively cast in the form of a game[1] (see figure 2). Educational games have attracted a great deal of attention and research.[2]

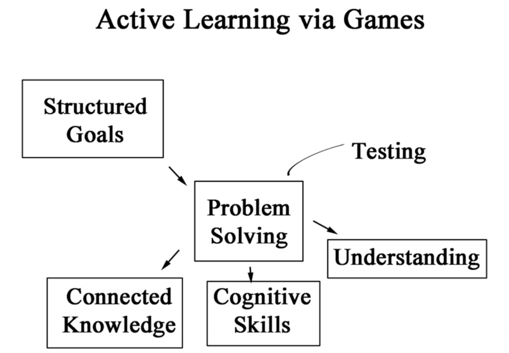


Figure 2.. *Schematic of activity-based learning.*

Recently, game technology and game-like approaches have attracted serious interest from researchers and authors in virtual heritage.[3] At the 2009 meeting of Computer Applications in Archaeology,[4] five non-game game applications explicitly used technology developed in the game industry, two game-like applications, and two actual games. The two games are *Gates of Horus* and *Outbreak.* The 2009 conference on Virtual Systems and Multimedia (VSMM) hosted a workshop on games for virtual heritage. Educational games for virtual heritage are also emerging in the virtual reality, edutainment, and educational literature. A good example is *Virtual Mandan Village.*[5] Of the three games games we just named, *Outbreak* most closely fits the definition of both a game and a learning exercise. In it, the user tries to stop a plague by manipulating key factors of a plague simulation, attempting to achieve an optimal outcome. In the other two, the narrative structure is centered on explicit investigation through question-and-answer and look-and-find activities.

Authenticity is central to virtual heritage applications, but achieving it raises complex issues of representation, scholarship, and audience/user perception.[6] These issues issues are especially interesting in educational games for virtual heritage.[7] The game narrative must be fully integrated and sensible within the content itself, an approach which Hokanson calls *authentic scenario learning*.[8] The rewards and challenges must be part of the content, and not some visual sugar or meaningless allocation of "points" that focuses the student on an irrelevant process.

Accordingly, the narrative structure of *Gates of Horus* comports with the Egyptians' emphasis on scholarship as the best way to engage with the world, which to them was a blend of physical, cultural, and mythic. More literally, scholarship was also the primary way to rise in the social order and literally be allowed into the more important areas of most temples. Jacobson and Holden[9] address issues of authenticity for the Virtual Egyptian Temple. The goal of the game is literally to learn more about the temple, and the reward at each stage is to see more of it. The game offers no points, no prizes, and no competition, but rewards students' study with more materials to study. We believe a game-like narrative structure, properly used, is an effective way to motivate students to engage mentally and emotionally with the material.

We constructed *Gates of Horus* for middle school students (ages 11 to 13) as part of a larger learning study. The first step was to see whether students could learn anything from it. Even though students generally learn something from almost any presentation, the design challenges for educational structure (curriculum), virtual reality, interface control, content, and narrative are such that success is never guaranteed. Therefore our caution was appropriate.

[1] Erik Champion, "Heritage Role Playing—History as an Interactive Digital Game," paper presented at the annual Australian Workshop on Interactive Entertainment, Sydney, Australia, February 13, 2004.

[2] D. Gaither and C. Redfield, "Survey of Electronic Games that Teach," paper presented at the Society for Information Technology and Teacher Education International Conference Chesapeake, VA, USA, 2006; M. J. Dondlinger, "Educational Video Game Design; A Review of the Literature," *Applied Educational Technology* 4 (1) (2007): 21–31.

[3] Erik Champion, "Otherness of Place: Game-based Interaction and Learning in Virtual Heritage Projects," *International Journal of Heritage Studies* 14 (3) (2008): 210–228.

[4] Computer Applications in Archaeology, 2009; www.caa 2009.org.

[5] Guy Hokanson et al., "Studying Native American Culture in an Immersive Virtual Environment," paper presented at IEEE International Conference on Advanced Learning Technologies, Santander, Cantabria, Spain, July 1, 2009.

[6] Bernard Frischer et al., "From CVR to CVRO: The Past, Present, and Future of Cultural Virtual Reality," VAST Euroconference, Arezzo 24–25 November 2000 (Archaeopress: Oxford, 2002) 7–18.

[7] Erik Champion, "Otherness of Place: Game-based Interaction and Learning in Virtual Heritage Projects," *International Journal of Heritage Studies* 14 (3) (2008):210–222.

[8] Guy Hokanson et al. (p. 138n5).

[9] Jeffrey Jacobson and Lynn Holden, "The Virtual Egyptian Temple," paper presented at the World Conference on Educational Media, Hypermedia and Telecommunications, Montreal, Canada, June–July, 2005. http://publicvr.org/downloads/Jacobson2005e.pdf.

We developed a fairly standard multiple-choice and short-answer test, which 20 students took after playing *Gates of Horus*. We compared their test results against those of another 20 students who took the test without having played the game. Tests of this type measure only the lower levels of learning (facts and simple concepts) and do not capture the higher-level (synthesis) learning that game-like Virtual Reality can support. However, we were confident that a standard classroom-like test was adequate for this stage, and we had other tests to explore deeper learning.[1] In a three-part post-test, students who played the game learned well compared to the no-game control group, scoring better in all three parts of the post-test ($p < 0.0016$, $p < 0.0044$, and $p \approx 0$). Importantly, they played the game with rapt attention for 45 to 60 minutes, and most reported that they enjoyed the experience.

This study adds to the small but growing number of successful examples of learning games for virtual heritage. In this paper, we describe the game and its narrative and surface-level content in detail so that the reader can better understand the student's experience. We also go into great detail on the post-test and the evaluation process and present our reasoning for the statistical analysis of the results in the context of the overall valuation. Apart from justifying our approach, we also hope to provide readers with a working example that they can use for their own evaluation studies.

2 THE GAME

The Virtual Egyptian Temple[2] provides most of the content and structure for *Gates of Horus*. The temple has no real-world analog, although it is constructed mostly from elements of the Temples of Horus at Edfu and at Medinet Habu. Its purpose is to embody the key features of the typical Egyptian temple of the New Kingdom period in a way that an untrained audience can understand. The temple consists of four major areas, the exterior (Pylon), the Courtyard, the Hypostyle Hall, and the inner Sanctuary, arranged in that order and separated by gateways. Compared to a real temple, the virtual Egyptian Temple model is simple, having only enough detail to represent the key features required (fig. 1). For example, there is only one of each of the four types of areas, while an actual temple might have had several Courtyards and Hypostyle Halls. Similarly, the hieroglyphics are larger than they would be in an actual temple to make them more legible, and there is a copy of the high priest in each of the major areas, functioning as a pedagogical agent. However, the scale and proportions of the spaces are correct, hieroglyphics make the appropriate statements, murals and statuary are in their proper locations, and so on.

In the game, the student navigates the temple using a standard three-button mouse. The view is "first person," meaning that the monitor is like a window looking into the virtual world, somewhat as though the physical viewer were actually in the virtual space. The student can rotate the view by moving the mouse and move forward and backward by pressing the left and right mouse buttons. When the student presses the middle button (the mouse wheel), the game switches to "selection mode," in which the mouse controls the cursor, shown in figure 1. The student can select an active object by moving the cursor over it in the current view and pressing the mouse wheel.The targeting is three-dimensional in the sense that the student can select the same object from many directions. For example, figure 3 shows a hawk statue in front of the temple as seen from two directions, with the targeting cursor over it. In this image the cursor is green, indicating that it is over an active object. To return to navigation mode, the student can press the mouse wheel again.

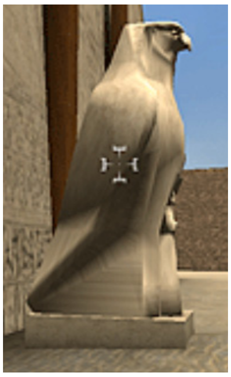


Figure 3. *Hawk statue and the cursor.*

Each time the student clicks on an "active" feature of the temple, such as the hawk, the priest explains its nature and meaning. When the student clicks on the priest, he asks the student a question based on what he has already explained. To progress from one area of the temple to the next, the student must answer all of the priest's questions for that area. The questions are based entirely on what the priest has to say about that area's active features. When the student correctly answers all of these questions, the Gateway to the next area of the Temple opens, and the student explores that area and learns about it in the same way. The student "wins" the game when s/he answers all of the questions from the priest in the inner Sanctuary, causing the divine image of the God to speak and bring the blessings of heaven to the land of Egypt.

Definitions:

1. The temple has four major *areas*, the Pylon (exterior), Courtyard, Hypostyle Hall, and Sanctuary.

[1] Jeffrey Jacobson, *Ancient Architecture in Virtual Reality; Does Immersion Really Aid Learning?* (Ph.D. diss., University of Pittsburgh, 2008). http://planetjeff.net/Jacobson2008.pdf

[2] Jacobson and Holden, p. 138n9 above.

2. An *active object* is a part of the temple itself, a statue, or some other thing on which the student can click to elicit a response from the game.
3. An *explanation* is a short voice recording (usually 10 to 25 seconds) that explains the meaning of some active object. The explanation names the object, introduces its basic meaning, and often describes where it fits into the overall themes of the temple and Egyptian religious life.

Basic Rules:

1. When the student first enters an area, s/he triggers an introductory recording, *only once*, by (virtually) walking into the area associated with it. Entering the area again will not trigger the introductory recording. However, the student may trigger the introductory recording at any time by clicking on the ground just inside the entrance to the area. That part is indicated by the smoke ring effect.
2. When the student clicks on an active object, the student hears a short voice recording which explains the object's meaning. The priest does not move, but the voice is thematically his.
3. Each time the student clicks on the priest, s/he hears a question associated with features s/he has *already selected* or with the introductory recording. For example, suppose a student enters an area, hears the introductory recording, clicks on two features in an area, hears each explanation, and finally clicks on the priest to elicit a question. The priest randomly selects a question regarding only those two features or the introductory recording.
4. The student may answer a question by using the right mouse button, with a single click for "No" or two clicks for "Yes."
5. Once a student has answered a question correctly, the priest will never ask it again.
6. If the student has correctly answered all questions associated with previously selected objects *and* the introductory recording, the priest prompts the student to click on some other feature.
7. When all questions for all objects in an area are answered correctly, the priest asks a set of "goal" questions. When the student has answered those correctly, the priest congratulates the user, and the gate to the next area opens.
8. The student enters the next area and answers all the questions there, in exactly the same manner described above. This happens a total of four times, once each for the Pylon, Courtyard, Hypostyle Hall, and Sanctuary.
9. When the student successfully answers all questions for the Sanctuary, the divine image of the temple god will "speak" in a recorded congratulation for winning the game.

Rules on Question Order:

10. When clicked, the priest always tries (randomly) to ask a question associated with the same object as the previous question, except where it conflicts with the next rule.
11. The priest never asks the same question twice in succession unless it is the last available question. This condition occurs when the student has correctly answered all but one of the questions associated with an area of the temple—the questions based on the introductory speech for the area, questions regarding all objects that student has previously selected in the area, and the final set of questions for the area.
12. If the student gives a *second* wrong answer for an object *since the time that object's recording played*, the student's viewpoint automatically moves to center on that feature, and the student hears the explanation again.
13.

Example

The student enters the courtyard through the main gate of the temple. When s/he (virtually) crosses the threshold s/he hears the voice of the priest give the introductory speech for that area: "The Courtyard is an open and undivided space, made for large religious celebrations and rituals. Everyone comes to these events dressed in the same simple garments. They do this to show how all people are equal and humble before the gods."

Next, the student goes further into the courtyard, and clicks on the highlighted mural, which shows Pharaoh making offerings to the enthroned god. S/he hears, "On behalf of all Egyptians, the King gives thanks by offering 'every good thing' to the god, who is their creator. In return, he blesses the King, the land of Egypt and its entire people with life and prosperity forever."

The student could click on other active objects, but instead chooses to click on the priest. The game could now ask the student questions regarding the mural or the introductory speech. It randomly selects the set of two questions regarding the mural. Of those it randomly selects and asks the student this question: "Do the gods give the King something in return for his offerings?" The student answers incorrectly with a "no" by single-clicking the right mouse button, and the priest says "incorrect" or a phrase with similar meaning. The student clicks on the priest again, so the priest asks the other question associated with the mural "Does Pharaoh represent the people's interest before the gods?" The student answers correctly with a "yes" by double-clicking the right mouse button. The priest congratulates the student and will never ask this question again. If the student clicks on the priest again, he asks the first question again: "Do the gods give the King something in return for his offerings?" If the student gives the correct answer "yes," the priest congratulates the student with a phrase such as "You are right!" and never ask the question again. However, if the student answers "no" a second time, the priest indicates the answer is wrong, and the view automatically focuses on the mural, filling the screen. The student hears the explanation of the mural again, after which s/he is free to navigate, click on more objects, or click on the priest.

In the study, students showed much difference in how they chose to click on active objects and answer questions from the priest. Some clicked on all the objects, listened to all of the explanations, and then attempted to answer all of the priest's questions in one session. Other students preferred to click on the priest immediately after hearing each explanation, while most students pursued a middle strategy.

Gates of Horus is available for download at http://publicvr.org. The code is open source and freeware and is based on two other freeware packages, CaveUT and VRGL.[1] They all depend upon a commercial game, UT2004 by Epic Games, which is currently out of print but still widely available at low cost. This implementation of UT2004 is only a prototype, but code is stable enough for anyone who wants to work with it, and it comes with logging capability if you want to build your own study around it. Changing the content requires editing the actual code, but that is a simple operation a student programmer could handle.

3 EVALUATION

Gates of Horus is a prototype learning game designed for middle school students (ages 11 to 13). After building a working prototype, the next step was to evaluate whether students could learn anything from it. Also, this study was a foundational part of a much larger research project[2] where we examined the learning effects of different display types during game play. At this stage, we are not comparing it to other learning methods, but merely seeing if it works. Here, we describe the testing process, summarize the data, and analyze the results.

We conducted all testing at the Earth Theater of the Carnegie Museum of Natural History in Pittsburgh (CMNH).[3] The theater features a partial dome display and is a venue for educational films and interactive tours of educational virtual environments—including the Virtual Egyptian Temple. The facility also has several workstations (PCs), recording equipment, and software to support data gathering for educational research. The Earth Theater's role and resources as an educational research facility were made possible by support from the Carnegie Museum's own education program, PublicVR, and the University of Pittsburgh.[4]

Through the CMNH's education program, we recruited middle school students (ages 11 to 13) from area schools, civic organizations, and individual families to participate in our learning study. We did our best to make sure the population was gender-balanced and as diverse as possible. This enhanced the *external validity* of our study by showing that our experimental results apply to a wider range of students.

Experimental Design

Students arrived at the Earth Theatre for testing either individually or in small classes, always attended by parents or teachers. Before testing each student, we randomly assigned him or her to an experimental group. After correcting for those who had to drop out for personal reasons and those whose data were not usable for technical reasons, we had data for 20 students in what we will call the Desktop group and 20 in the Control group. Students in the Desktop group played *Gates of Horus* on a standard PC and then took the knowledge post-test to see what they learned. Students in the Control group took the post-test first. We did select a third group, but that has no bearing on this paper.

Randomly assigning subjects to test groups is important to balance differences between the two groups, increasing the chance that the difference in the groups' test scores is meaningful, i.e., in this case, really because of *Gates of Horus* and not some other factor. For example, we would not want all members of one test group to be 11 while all the others are 13. Through random assignment, all ages in our selected age range are evenly distributed within the test groups.

All members of the Desktop group played *Gates of Horus* and took a knowledge post-test afterward. Members of the Control group simply took the test without playing the game. Our experimental hypothesis is:

> Students who played *Gates of Horus* to completion will demonstrate superior knowledge of more of the facts and concepts around the Virtual Egyptian Temple than those who have not. Formally stated: students in the Desktop group will have statistically higher test scores on the post-test than those in the Control group.

We needed to test the Control group this way, for several reasons. First, we could not give a meaningful knowledge pre-test to the Desktop group, because the test itself would have given them too much information about the temple and what we wanted them to learn from it. Essentially, the Control group's post-test takes the place of a pre-test for the Desktop group. More importantly, comparing pre-test and post-test scores for a single group of students who went through a single experience (such as playing the game) is not adequate for scientific research. There is always the possibility that something else may account for the increase in test scores—such as taking the pre-test.

Finally, ancient Egypt is part of the standard school curriculum for middle school students in our area. Giving the Control group the test before having them

[1] Jeffrey Jacobson and Michael Lewis, "Game Engine Virtual Reality With CaveUT," *IEEE Computer* 38 (4) (2005): 79–82. http://publicvr.org/IndexDownloads/Jacobson2005i.html.

[2] Jeffrey Jacobson, p. 139n1 above.

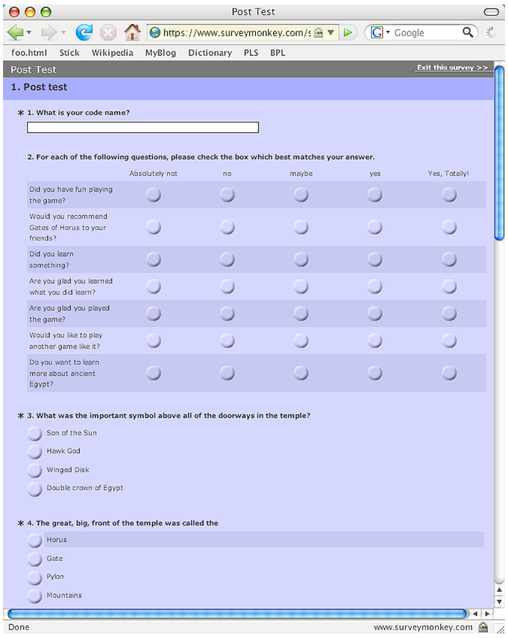
[3] Carnegie Museum of Natural History, 4400 Forbes Avenue, Pittsburgh, PA 15213. www.carnegiemnh.org/.

[4] University of Pittsburgh, School of Information Sciences, 135 North Bellefield Avenue, Pittsburgh, PA 15260. www.ischool.pitt.edu.

play *Gates of Horus* gave us a good idea of how well students in this group could do on the post-test by using prior knowledge. It made the comparison in our experimental hypothesis more interesting. After they took the test, we let them play the game, so they wouldn't feel left out of the fun. We also had a recruiting agreement with their schools and parents to provide an educational experience.

Data-Gathering Method

We implemented the post-test as an online form that students could fill out on any of the desktop computers we set up for that purpose. Figure 4 shows a sample page. The form was a private webpage through a survey data-gathering service (Survey Monkey 2008) which hosted all our questionnaires. The service automatically recorded all students' answers and made them available to us as both raw and summary data via our secure online account with the service. The post-test consists of affective, multiple-choice, and short-answer questions. In this section, we will discuss what these questions are and how we gathered the data.



Post Test

1. Post test

✱ 1. What is your code name?

2. For each of the following questions, please check the box which best matches your answer.

	Absolutely not	no	maybe	yes	Yes, Totally!
Did you have fun playing the game?	○	○	○	○	○
Would you recommend Gates of Horus to your friends?	○	○	○	○	○
Did you learn something?	○	○	○	○	○
Are you glad you learned what you did learn?	○	○	○	○	○
Are you glad you played the game?	○	○	○	○	○
Would you like to play another game like it?	○	○	○	○	○
Do you want to learn more about ancient Egypt?	○	○	○	○	○

✱ 3. What was the important symbol above all of the doorways in the temple?

○ Son of the Sun
○ Hawk God
○ Winged Disk
○ Double crown of Egypt

✱ 4. The great, big, front of the temple was called the

○ Horus
○ Gate
○ Pylon
○ Mountains

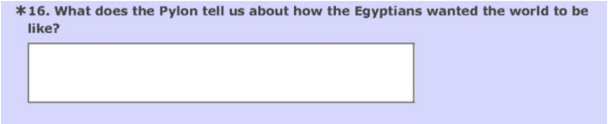
Figure 4. *First page of the post-test showing sample affective and multiple-choice questions.*

Figure 4 shows all of the affective questions, which query attitudes about the game from students in the Desktop group. Students in the Control group skipped these. We will report the totals in the Results section.

The multiple-choice question at the bottom of figure 4 is an example of a question that the computer can score automatically. Questions of this type are very easy to evaluate, but they are able to measure only low-level factual learning. Of the multiple-choice questions, there were nine that required exactly one answer from a choice of several. The student received one point for a correct answer, or no credit at all. Another three multiple-choice questions asked the student to "check all that apply." The student received a fraction of a point for every correct answer, based on the number of correct answers. For example, if a question of this type had three right choices and two wrong choices, the student would receive 1/3 of a point for each correct answer selected. We did not penalize for checking wrong answers—fortunately the students did not know this!

Short Answer Questions

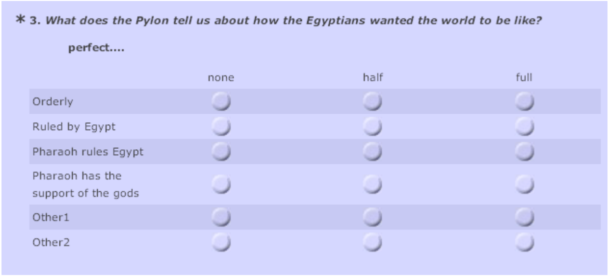
Figure 5 shows a sample short-answer question. The student is not limited to the size of the box in typing the answer because it will scroll if a paragraph is too long for it. However, because of box size, time constraints, and convention, answers rarely exceeded 60 words and often only used a few. These questions require a little more thought and integration of ideas from the student, which is good for detecting conceptual learning, but they require manual grading by qualified evaluators. To use grader time as efficiently as possible, we created an individual grading form for each student's short-answer questions from the post-test. Figure 6 shows part of the grading form for one particular student. The first sentence is the original question. The second sentence is the student's answer. The next line allows for the grader to give full, half, or no credit based on whether the student's answer shows understanding that the Egyptians want the world to be an orderly place. The grader does this by checking one of the three radio buttons in the same line. The software will allow the grader to check only one. In this case, it appears that the student should receive full credit on the first concept, but no credit on the others.



✱16. What does the Pylon tell us about how the Egyptians wanted the world to be like?

Figure 5. *Sample short-answer question.*

The online grading forms were hosted at the same survey service that hosted the post-test itself. All of our graders received training on the content and purpose of the game, the intended meaning of the process questions, and the meaning of the questions on their evaluation form. Importantly, the graders did not know which tests were associated with which students (they never even met the students) or which tests were associated with which group.



* 3. What does the Pylon tell us about how the Egyptians wanted the world to be like?

perfect....

	none	half	full
Orderly	○	○	○
Ruled by Egypt	○	○	○
Pharaoh rules Egypt	○	○	○
Pharaoh has the support of the gods	○	○	○
Other1	○	○	○
Other2	○	○	○

Figure 6. *Original question, the student's answer, and scoring options for the grader.*

It would be much easier to simply have one grader do all of the evaluations, but this could be problematic. Valuable information could be lost if the grader misunderstood one or more of the post-test questions or how to evaluate them. S/he might also be too strict or too lenient in the overall grading, which may nullify differences between students in the overall response to questions. A single grader would have personal knowledge and skills which might influence the final result. That is why we had four graders each evaluate all short-answer questions for all students. They were one high school teacher, one middle school teacher, an art historian (not an Egyptologist), and a tour guide for the virtual Egyptian Temple shown in the Earth Theater. With four graders' scores for each concept in each student's answer to each question on the post-test, we were able to combine those scores to produce a more stable overall measure.

It would be tempting to simply average those four scores, but that could be highly problematic. For example, if two of the graders for one concept gave the student no credit and the other two assigned full credit, the average would be half-credit. However, such a high level of disagreement among the evaluators makes the average score meaningless. We should not use data from that question at all! Conditions like this can be caused by some inadequacy in the written test, the evaluation form, or the evaluators' training, ambiguity in the student's answer, or some other factor.

Reconciling scores from multiple evaluators is a branch of statistics called Interrater Reliability Analysis. In our study, we used the Fleiss Kappa algorithm[1] to combine evaluators' scores. We had to eliminate data from six of the 20 concepts the evaluators graded. However, most of those six questions yielded very few total points across all students, so they were not contributing much to the analysis anyway. This happens when a concept or question was too difficult, the game did not convey it well, or there was some other problem. Nevertheless, the remaining 14 rated concepts provided more than enough data to demonstrate learning for students who played the game. We used a simple average of the data for the 14 surviving concepts.

Rater Impressions

The last short-answer question is special, because it is so open-ended. It asks "Tell us one thing you learned from playing with the Temple." The graders score 3 general aspects of the student's answer, "importance", "relevance", and "generality" on a three-level scale, "low", "average", and "high." In the data analysis we treat "low" as no credit, "average" as half-credit, and "high" as full-credit. The data for the evaluation of all three aspects survived interrater reliability analysis.

Finally, we asked the graders four questions on their overall impression of the student's performance. The responses to two of the questions did not pass interrater reliability analysis, indicating substantial disagreement. The two that did were "Student is making interesting connections between facts/concepts showing a higher level of learning" and "Student is doing a good job of reciting the facts of the temple." The graders responded to these statements on a five-level scale, from "strongly disagree" to "strongly agree." Interrater reliability analysis seems especially appropriate to this type of evaluation.

3 RESULTS

As we expected, students in the Control group did remarkably well on the standard multiple-choice questions, scoring an average of 11.8 points out of a possible 20. We attribute this to a combination of prior knowledge acquired in school and elsewhere and to test-taking skills. Students in the Desktop group, however, did better, averaging 14.55 points. Comparing the two sets of scores with a standard two-tailed T-Test yielded:

$P < 0.0016$

This shows that there is less than a two-tenths of a percent chance that the difference between the averages is a random result. This means that students who played the game almost certainly learned from it. In most research literature, $P < 0.05$ is considered adequate for most tested assertions (hypotheses). We used the "two-tailed" version of the T-Test, which tests for both possibilities: that the Desktop group did better OR that the Control group did better. Obviously, the latter did not happen here, but it could have. It is often tempting to use the one-tailed version, because it usually yields a lower P value, but legitimate situations where it is appropriate are relatively rare.

As we described in the previous section, students answered five short-answer questions. Out of each student's five answers, graders evaluated the student's comprehension of twenty concepts. The data for 14 of those concepts survived interrater reliability analysis. With a perfect possible score of 14, the Control group averaged 2.5461 and the Desktop group averaged 4.4625. (This part of the test was difficult and the

[1] Joseph Fleiss et al., *Statistical Methods for Rates and Proportions* (Hoboken, NJ: Wiley, 2003).

graders often assigned half-credit.) The T-Test revealed a fairly strong result:

P < 0.0044

Thus, there is approximately a 99.5% chance that the Desktop group genuinely scored higher than the Control group and that this is not just a random result.

The next seven questions asked the graders to rate aspects of student performance from 0 to 1. The data for five of the ratings survived interrater reliability analysis. Out of 5 possible total points, Control students averaged 1.0054 and Desktop students averaged 2.9856. Comparing the two lists of student grades with a T-Test,

P < 0.00000086

which is essentially zero. While questions like these have a degree of subjectivity, the strong consensus between the four graders means that it carries some credibility. Interrater reliability analysis reveals that consensus. We believe that this type of rating benefits most from interrater reliability analysis.

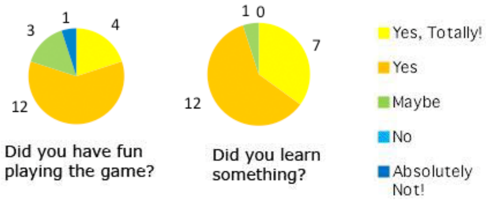


Figure 7: *Results from two of the affective questions.*

Finally, fig. 7 shows a tally of the results from the two affective questions showing the strongest result. Each number represents how many students in the Desktop group expressed the opinion represented by the nearest pie slice. This is not a scientific measure, because it is not compared to anything, but it is encouraging. Students often told us that they like the game "better than schoolwork," which makes sense because many of them were recruited through their schools. Probably the best indicator that the students were focused on the learning activity (the game) was that nearly all test subjects played it with rapt attention for up to an hour.

4 DISCUSSION

The evaluation demonstrates that our virtual heritage game is a genuine way to learn, and not just a toy. Our other goal was to demonstrate that evaluating similar virtual heritage applications is straightforward and not difficult. We used a fairly standard multiple-choice and short-answer quiz, because it is easy to administer and widely accepted.

We emphasize that our evaluation does *not* say whether the game is any *better* than some other learning activity—only that the game worked. This is an interesting area we intend to research. We believe that every mode of teaching has its own advantages and disadvantages. Comparative learning studies can show us which mode is best given the situation and the learning goals. Understanding this helps us use different learning modes *in concert* to build an effective overall curriculum. For example, *Gates of Horus* makes a good companion to the physical collection at the Carnegie Museums, their Virtual Eygptian Temple tour, and a booklet written for the virtual temple. Each venue has its own strengths, and they reinforce each other.

ACKNOWLEDGEMENTS

Gates of Horus and the learning studies based on it depend on the Virtual Egyptian Temple, CaveUT, and VRGL. Each of these is a separate project with a long list of credits—too long to list here. To see them, go to http://publicvr.org and look under "Projects." For *Gates of Horus* and the evaluation study, we first thank the following institutions: (1) PublicVR for providing the funds and open source software to develop the game, testing software, and other testing materials; (2) The Carnegie Museum of Natural History (CMNH) for providing the venue, the test population, payment for Kerry Handron's time, and miscellaneous resources for the game and the study; (3) The University of Pittsburgh Department of Information Sciences for making their Virtual Theater available for software testing and general support for Jeffrey Jacobson's work on this project; (4) Carnegie Mellon University's Studio for Creative Inquiry for their role in developing the Virtual Egyptian Temple, and the Fine Arts department for Lowry Burgess's time advising on the project; (5) Epic Games for their general support for academic research using the game Unreal Tournament 2004 and its software core, the Unreal Engine; (6) Psyonix, Inc. (contractors) for the software programming; and (7) Wrecking Crew, Inc. (contractors) for the voice recordings. Next, we would like to thank the following individuals: (1) Michael Lewis (University of Pittsburgh), who generally supported the work though his role as Jeffrey Jacobson's dissertation advisor; (2) Lowry Burgess for his insights into the dramatic structure of the game and issues of learning and perception; (3) William Winn, Kurt Van Lehn, Peter Brusilovsky, and Stephen Hirtle for their advice on the educational research design, cognition, and adaptivity; (4) Mack C. Shelley for help in selecting the appropriate statistical methods used to analyze the results; (5) Jane Vadnal for her help in the production of materials for the experiments; and (6) Kathy Bruckner, Adrienne Baker, Benjamin Getkin, Jane Vadnal, and Mary Smith for their excellent work in evaluating the students' performance in the post-tests, as well as in test subject recruitement and preparation of test materials.

BIBLIOGRAPHY

Bloom, B. S. *Taxonomy of Educational Objectives*. Chicago: New York; Longman Inc, 1956.

Brusilovsky, Peter, and C. Peylo. "Adaptive and Intelligent Web-based Educational Systems." *International Journal of Artificial Intelligence in Education* 13 (2003): 156–169.

Carnegie Museum of Natural History, 4400 Forbes Avenue, Pittsburgh, PA 15213. www.carnegiemnh.org/.

Champion, Erik. "Otherness of Place: Game-based Interaction and Learning in Virtual Heritage Projects," *International Journal of Heritage Studies* 14 (3) (2008): 210–228.

Champion, Erik. "Heritage Role Playing—History as an Interactive Digital Game," paper presented at the annual Australian Workshop on Interactive Entertainment, Sydney, Australia, February 13, 2004.

Dondlinger, M. J. "Educational Video Game Design; A Review of the Literature," *Applied Educational Technology* 4 (1) (2007): 21–31.

Fleiss, Joseph, et al. *Statistical Methods for Rates and Proportions*. Hoboken, NJ: Wiley, 2003.

Frischer, Bernard, et al. "From CVR to CVRO: The Past, Present, and Future of Cultural Virtual Reality," VAST Euroconference, Arezzo, 24–25 November 2000 (Oxford: Archaeopress, 2002) 7–18.

Gaither, D., and C. Redfield. "Survey of Electronic Games that Teach," Paper presented at the Society for Information Technology and Teacher Education International Conference, Chesapeake, VA, USA, 2006.

Hokanson, Guy et al. "Studying Native American Culture in an Immersive Virtual Environment," paper presented at IEEE International Conference on Advanced Learning Technologies, Santander, Cantabria, Spain, July 1, 2009.

Jacobson, Jeffrey. *Ancient Archiecture in Virtual Reality; Does Immersion Really Aid Learning?* Ph.D. diss., University of Pittsburgh, 2008. http://planetjeff.net/Jacobson2008.pdf.

Jacobson, Jeffrey, and Lynn Holden. "The Virtual Egyptian Temple," paper presented at the World Conference on Educational Media, Hypermedia and Telecommunications, Montreal, Canada, June–July, 2005. http://publicvr.org/downloads/Jacobson2005e.pdf.

Jacobson, Jeffrey, and Michael Lewis. "Game Engine Virtual Reality With CaveUT," *IEEE Computer* 38 (4) (2005): 79–82. http://publicvr.org/IndexDownloads/Jacobson2005i.html.

Jonassen, D. H. "Constructivism and Computer-mediated Communication in Distance Education," *American Journal of Distance Education* 9 (2) (1995): 7–26.

Kee, Kevin. "Outbreak: Best Practices and Potential for the Development of Games for Archaeology and History," paper presented at Computer Applications to Archaeology, Williamsburg, VA, USA, March 22, 2009). PublicVR: http://publicvr.org.

University of Pittsburgh, School of Information Sciences. 135 North Bellefield Avenue, Pittsburgh, PA 15260. www.ischool.pitt.edu.

Winn, William. "Learning in Artificial Environments: Embodiment, Embeddedness and Dynamic Adaptation," *Cognition and Learning* 1 (2003): 87–114.