Expert Evaluation of Student Presentations to Assess Learning in an Interactive Digital Dome; Opening the Gates of Horus

Abstract accepted for the 2012 annual meeting of the
American Educational Research Association, Vancouver, CA

Jeffery Jacobson, PublicVR, Boston, USA
jeff@publicvr.org

The educational potential of virtual reality (VR) and augmented reality (AR) can be realized only with the development of theory to guide design. That in turn requires realistic and effective means to measure when, where, and how immersive technologies and techniques are best used. The topic is large and poorly defined, so we will focus on the effect of visual immersion and the use of knowledge domain experts to evaluate student performance. *We believe that if the students are asked to demonstrate their newly gained knowledge in a meaningful way, domain experts will be effective in evaluating student learning.*

Empirical research on the educational use of visual immersion is remarkably rare, despite its much touted promise and significant resources spent on it, but some work has been done. Experiments by Salzman, Dede, and Loftin (1998) and by Jacobson (2011) show that an egocentric (inside) view of some *thing* (a magnetic field, virtual classroom, Egyptian Temple) can measurably assist learning in certain topics. This requires that some key aspect of the knowledge domain is expressible as a three-dimensional structure *and* that the egocentric view of the structure must improve the student's ability to access information. While not directly focused on this question, studies by Winn (2001) and Bailenson (2008) indirectly support our thesis.

Here, we describe a follow-on analysis of data from the study in Jacobson (2011), employing expert evaluators. The results confirm our original experimental result and refine our understanding of it.

In our study, we used an instructional game, *Gates of Horus*, developed by PublicVR. The structure of the game and its effectiveness for learning is reported in Jacobson (2009). In the game, the student gains access to progressively more hidden areas of a virtual Egyptian temple by answering questions in a dialogue with the virtual Egyptian priest. Temples of this kind were a highly developed art form designed to inform viewers on many levels. The student learns about the temple and Egyptian society through facts and statements, which are illustrated in the structure and artwork of the temple. In our learning study reported in Jacobson (2011), some students played the game on a standard desktop computer (control group), while others used an all-digital dome (treatment group).

Among other tests, each student recorded his/her own guided tour of the Temple, which was independently scored by three evaluators, unaware of which students were in the dome (treatment) or desktop computer (control) group. Each evaluator watched each video, using a checklist to assign points for key facts and ideas as the student mentioned them. Students who used the dome recited more facts in their presentations ($P < 0.05$). We theorize that the students with the egocentric view (1) learned more facts, (2) had a better visual mental map of the temple and were able to mentally "walk" the structure to recall those facts, (3) had a deeper understanding of the knowledge overall, or (4) experienced some combination of these three factors.

This result was encouraging, but unsatisfying, because an open-ended fifteen-minute exposition on the temple contained so much information on the student's understanding and presentation style. Tantalizingly, graders' responses to a single subjective question for each student ("Is the student doing a good job of reciting the facts of the temple?") showed a stronger sensitivity to the experimental treatment ($P < 0.01$) than the aggregate score of our carefully designed and scored checklists they used in the formal evaluation ($P < 0.05$). Our graders did know the temple and the game itself well enough to assess the accuracy and completeness of the facts presented by the students. So, could the graders' educated intuition be more accurate than their responses to detailed questionnaires?

Lynn Holden was our content expert for both the temple and the Gates of Horus, at the time. To further investigate this effect, we gave him a questionnaire with broader and deeper questions, such as "When in one space, does the student refer to objects in the other spaces?" We hoped that his expertise would give him deeper insight. Results based on his judgments also showed that the dome group did better than the control group, with high significance ($P < 0.0004$)(Jacobson, 2008, p180). To verify this result, we had a more senior Egyptologist, Dr. Robyn Gillam, score the same videos in the same way, with much the same main result.

The next step was to combine Holden's and Gillam's judgments, subject to Interrater Reliability Analysis. For each question on their questionnaire, each expert Egyptologist gave a rating for each of the 61 student videos. For each question, this gave us two lists of values, one from each expert, to average together. But we will only do this if the two lists are in sufficient agreement. We decided that calculating a simple correlation coefficient was appropriate, because we cared only whether each expert (unknowingly) gave higher scores more often to students in the dome group. We did not care whether one expert was consistently more lenient or whether both experts gave the same grades to individual students. (Fliess's Kappa, which we used for the non-expert graders, is sensitive to both of these factors.) We decided that the correlation between scores awarded by the two experts for each question had to be correlated at $R = 0.5$ or higher. Otherwise, we judged their opinions (data) to be too greatly in disagreement to be meaningful.

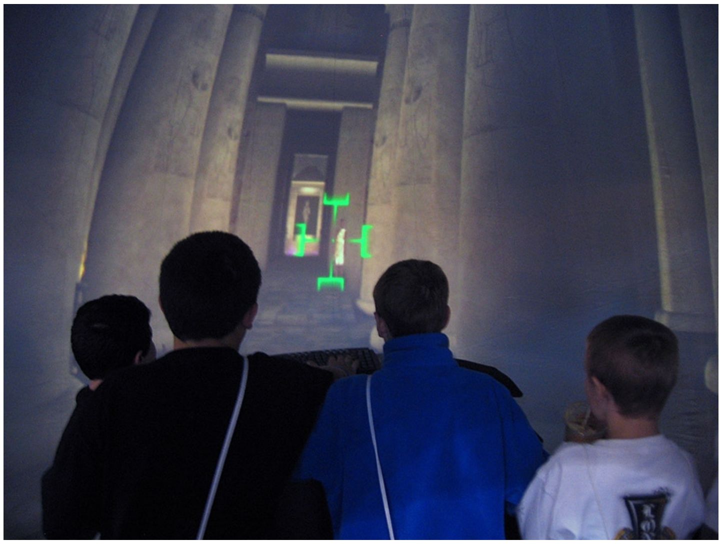
The expert evaluators' results were correlated over R=0.5 for five questions. However, we had to reject the results for one of the questions because Holden awarded a non-zero score to only 7 students (out of a total 61) for that question. There is nothing wrong with this judgment on his part, but it meant that the correlation with Gillam's observations for that question was probably random, or at least based on very little evidence. We were therefore left with valid data for four questions, which we could safely average together for further analysis.

We used a two-tailed uneven samples T-TEST (one-way ANOVA) to compare the means of the data for the dome (treatment) and desktop computer (control) groups. The results are shown below. As with the non-expert evaluators, the Overall Impression question produced the same result as the average of results from all the other questions, including the ones that did not show significant difference between groups or failed the Interrater reliability test. In fact, the averaged scores for the "overall impression" question and all other questions are correlated at r=0.90, which is nearly perfect.

	Correlation between rater results.	TTEST means, treatment vs. control
Does the student seem to know the facts that s/he is talking about?	r=0.6833	p = 0.0071
Is the student doing a good job integrating the visual and verbal knowledge they got from the game?	r=0.5895	p = 0.0001
Is the student connecting the facts with the visuals?	r=0.5031	P = 0.0038
Average of ALL data except the "Overall" question.	r=0.6452	P = 0.0004
What is your overall impression of the student's performance?	r=0.5708	P = 0.0002

We conclude that students who played the game in the dome tended to have a better synthesis or mental map of the information. This comports with our supposition that an Egyptian Temple is meant to be seen and "read" from the inside, as a whole piece. Students in the dome may also have gained a mechanical advantage for accessing the information.

We also conclude that the component questions are useful mainly for the insight they give us into the anatomy of students' performance. A domain expert can provide a reliable overall score of student performance by answering a single overall question, if the student has an opportunity to display his or her knowledge in a meaningful way. By asking fewer questions of more expert evaluators we can lower our research cost and complexity. More importantly, their judgment is probably a more sensitive instrument, given the much greater levels of significance we found in our study compared to analysis by non-experts.



One student plays the game, Gates of Horus, using a cursor to select features of the virtual Egyptian Temple.

Bailenson, J. N., Yee, N., Blascovich, J., Beall, A.C., Lundblad, N., and Jin, M. (2008). The Use of Immersive Virtual Reality in the Learning Sciences: Digital Transformations of Teachers, Students, and Social Context. *The Journal of the Learning Sciences*, 17, 102-141 ISSN: 1050-8406 print / 1532-7809 online.

Jacobson, J. (2011). The Effect of Visual Immersion in an Educational Game; Gates of Horus, *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, special issue on educational applications, Spring, 2011, IGI Global.

Jacobson, J., Handron, K., and Holden, L. (2009). Narrative and Content Combine in a Learning Game for Virtual Heritage. *Computer Applications in Archaeology*, Williamsburg, VA.

Jacobson, J. (2008). Ancient Architecture in Virtual Reality; Does Visual Immersion Really Aid Learning? *Dissertation*, School of Information Sciences, University of Pittsburgh

Salzman, M. C., Dede, C., Loftin, R. B., and Ash, K. (1998). Using VR's Frames of Reference in Mastering Abstract Information. *Third International Conference on Learning Sciences*, Charlottesville, VA.

Winn, W. (2003). Learning in Artificial Environments: Embodiment, Embeddedness and Dynamic Adaptation. *Technology, Instruction, Cognition and Learning*, 1, p. 87-114.

Winn, W., Hoffman, H., Hollander, A., Osberg, K., Rose, H., and Char, P. (1997). The effect of student construction of virtual environments on the performance of high- and low-ability students*. Paper presented at Annual Meeting of the American Educational Research Association*, Chicago, IL.