Addressing the Complexities of Evaluating Interdisciplinary Multimedia Learning Environments

Steven McGee[1], Bruce C. Howard[1], Dimiter M. Dimitrov[2],
Namsoo S. Hong[1], & Regina Shia[1]

[1]Center for Educational Technologies™
Wheeling Jesuit University

[2]Kent State University

AERA 2001
Division C6: Technology Research

1 Introduction

In this paper we report results of the summative evaluation of *Astronomy Village®: Investigating the Solar System*™. Funded by the National Science Foundation, *Astronomy Village* is designed to teach students fundamental concepts in life, earth, and physical science by having them investigate cutting-edge questions related to the solar system. There are two fundamental evaluation problems that are a consequence of the interdisciplinary nature of *Astronomy Village*. First, it was difficult to determine what would be an adequate comparison group since there were no equivalent materials that taught the same topics in an alternative format. Second, since students would be studying diverse topics in-depth, it was difficult to find sufficient numbers of items that covered the nontraditional topics to the depth necessary for an evaluation instrument. We discuss the strategies that were developed to address these problems and then present the results of the evaluation.

1.1 Astronomy Village®

Through *Astronomy Village* students are transported to a virtual village in Hawaii where they investigate one of two Core Research Topics: what the surface of Pluto might look like when the first NASA mission arrives in 2015, or the search for life in the solar system (McGee & Howard, 1999). The program is designed such that a virtual mentor guides students in completing multiple investigation cycles that mirror the phases of scientific inquiry. In the first investigation cycle, students are introduced to the core research question concerning either the surface of Pluto or the core requirements for life. During the exploration phase of the investigation, students are exposed to the types of data they will be using in the investigation to prepare them for future analyses. In the background research phase, students read library articles and listen to lectures to help them understand key background concepts. During the data collection and analysis phases, students use the results of their analysis to draw conclusions about the research question. This core investigation cycle lasts about one week. Students then follow the same sequence of phases as they did in the core investigation when they undertake a focused investigation on a narrower topic. For example, students may investigate whether icy volcanoes could exist on Pluto by examining the surfaces of icy moons in the

Solar System. Or students may examine temperature/pressure relationships on a variety of planets and moons to determine where the conditions are right to support liquid water. Students complete the investigation by hosting a virtual press conference in front of a virtual press corps that asks the students questions about the investigation they just completed.

1.2 Comparison Group Problem

For the evaluation of *Astronomy Village*, we were most interested in how technology could be effectively used to teach interdisciplinary content ideas related to the surface of Pluto and the search for life. Therefore, we were not interested in comparisons between technology use and no technology use. Instead, we were interested in a comparison between alternative uses of technology for teaching interdisciplinary content. *Astronomy Village* supports an inquiry-based approach to solar system content. Through a content analysis of the software, we determined that the unique strength of the software lay in the use of image analysis to answer important research questions. We focused on a comparison between access to image analysis activities and no access to image analysis activities. Therefore, comparison students studied the same content using the same technology as *Astronomy Village* students, but were denied access to the image analysis capabilities. We deemed the comparison students as the alternative treatment group.

The interdisciplinary nature of the investigations meant that students would be studying diverse topics in-depth. For example in the Search for Life Research Project, students study photosynthesis, chemosynthesis, water erosion, planetary evolution, wavelengths of light, pressure/temperature relationships, and the lower bounds to the size of life. In the Mission to Pluto Core Research Project, students study volcanism, cratering, and plate tectonics on other planets. There were no existing set of materials that taught these topics in an interdisciplinary way. Therefore, we were faced with creating curriculum materials that could serve as a comparison to *Astronomy Village.*

The alternative treatment group was provided with the collection of background materials in *Astronomy Village* that supports each investigation: lectures, library articles, and hands-on activities. These materials were presented through a Web site specifically designed for the comparison group. Given a focus on content-related activities only, the alternative treatment group was able to cover all of the topics in *Astronomy Village* over a four-week period. This is reminiscent of traditional approaches to science education that focus on covering a wide array of content at a superficial level. In contrast, the students in the Ast*ronomy Village* treatment group participated over an equivalent four-week period but covered only the topics related to one of the two Core Research Topics. *Astronomy Village* students engaged in both the content-related activities as well as the inquiry-oriented image analysis activities. This evaluation design held technology as a constant and fostered a comparison between traditional breadth approaches and the depth approaches as recommended by the *National Science Education Standards* (NRC, 1996).

1.3 Appropriate Assessment Problem

Traditional assessment measures break complex content into constituent, simpler content ideas. Assessment items related to the simple content are developed. The collection of simple content ideas represents the universe of assessable ideas. The assumption is that the assessment of understanding constituent, simple concepts can represent the complex concept. The benefit of testing simple conceptual ideas is that it is easier to maintain high levels of reliability. Given the complexity of the ideas presented in *Astronomy Village*, we did not feel it would be feasible to assess students' understanding of the simple ideas that comprised the complex ideas. There were too many to cover and we felt it would not be adequate to sample from the domain of simple conceptual ideas. Instead, we chose to measure understanding at the complex level of understanding. Although this approach sacrificed our ability to achieve high levels of reliability, we gain in our ability to interpret the results of the evaluation in light of the actual content in *Astronomy Village*.

There were three guiding principles for the design of the assessment instrument. First, the assessment instrument should reflect important thinking and problem solving skills from the discipline of planetary science (Hickey, Wolfe, & Kindfield, 1999; Sheppard, 2000). In both the *Astronomy Village* treatment group and the alternative treatment group, students investigated authentic questions, such as whether liquid water exists in the solar system, that require important thinking and problem solving skills from the discipline of planetary science. Therefore, we achieved this principle by designing assessment tasks that reflect the thinking and problem solving that is targeted in *Astronomy Village*.

The second guiding principle was measuring the extent to which students transfer their thinking and problem solving skills into new contexts (Bransford, Brown, & Cocking, 1999). This principle reflects the philosophy that a critical aspect of education is whether learning transfers (Sheppard, 2000). When there is no specific transfer situation, the assessment becomes the transfer situation (Hickey, Wolfe, & Kindfield, 1999). *Astronomy Village* supported transfer by having students investigate critical processes and features on a variety of planets and moons. For the assessment instrument, students had to transfer their understanding to hypothetical planets and moons.

The third guiding principle was ease of administration and scoring for the target population. In prior research at the high school level, we have had success measuring complex problem solving and argumentation abilities using an extended response format (Shin, Jonassen, & McGee, 1998; Hong, McGee & Howard, 2001). However, at the middle school level, there was concern that the extended response format would be a better reflection of students' writing abilities than their problem solving abilities. In addition, the extended response format was too labor intensive to score within the budget limitations of the project. We therefore chose to use a machine-readable multiple-choice format. Taking into account the three guiding principles collectively, we felt confident in developing an assessment instrument that would measure important learning outcomes in a cost effective manner.

We identified the key complex content ideas that were presented in each of the nine investigations within *Astronomy Village* along with the key problem-solving skills related to drawing conclusions from data and inferring planetary processes from analyzing images of surface features. We contracted with item writers to develop the assessment items related to the underlying concepts within the investigations (see the Appendix for example items). The resulting instrument has four subscales: Search for Life complex content, Search for Life problem solving, Mission to Pluto complex content, and Mission to Pluto problem solving.

Our hypothesis for the evaluation is that students in the alternative group would perform well on the content subscales related to each core area but not as well on the problem-solving scales. Whereas, the *Astronomy Village* groups would perform well on both the content and problem-solving subscales of the core area they studied, but not do as well on the core area that they did not study.

2 Method

2.1 Participants

There were a total of 940 students from schools around the United States participating in either the *Astronomy Village* treatment group or the alternative treatment group. The gender breakdown was 38% female, 40% male, and 22% did not indicate. The ethnic breakdown was 48% Caucasian, 9% Asian/Pacific Islander, 6% Hispanic, 3% African American, 2% Other, and 32% did not indicate. There were 543 students that served as a no treatment control group. We did not collect demographic data from the no treatment control group. The grade level breakdown was 8% fifth grade, 18% sixth grade, 25% seventh grade, 26% eighth grade, 17% ninth through twelfth grade, and 6% did not indicate.

2.2 Procedure

Teachers were recruited to participate in the evaluation of *Astronomy Village* through an application process. We received around 50 applications and selected 12 teachers to participate. Those teachers were trained during a summer workshop on how to use the software and how to administer the assessment questionnaires. Half of the teachers decided to have their students investigate the Search for Life Core Research Project. The other half decided to have their students investigate Mission to Pluto. The teachers were instructed to have their students spend approximately four weeks completing as many of the focused investigations under the Core Research Project as possible.

In response to the request for proposals for the alternative treatment group teachers, we received and accepted five teachers who met the criteria for participation. They were trained during a second summer workshop on how to use the Web site to teach the core concepts as well as data collection procedures. In addition to the *Astronomy Village* treatment group and the alternative treatment group, we asked each participating teacher to recruit another teacher at their school to serve as a no treatment control group.

In addition to simulating traditional instruction through access to traditional materials, we also wanted to simulate the pacing of traditional instruction, where topics are covered quickly in order to move on to the next topic. Therefore we asked the alternative group to cover as much of both topics as they could in the same four-week timeframe.

3 Results

Traditionally, changes in variables such as achievement and attitude have been measured by differences between pretest and posttest scores. This approach has serious drawbacks. First, difference scores are less reliable than the scores entering the difference (see, e.g., Allen & Yen, 1979, p. 208). Second, raw scores are test dependent (i.e., depend on the difficulty of the test). Third, raw scores do not adequately represent response patterns since different response patterns on test items may lead to the same raw score. Fourth, pretest to posttest raw scores differences do not separate changes due to experimental treatments from changes due to natural trends such as maturation and experience. Finally, raw score differences do not provide information about magnitudes of changes on a ratio scale--in other words, it is not known how many times a given change effect is larger than another change effect. Item response theory models such as Linear Logistic Models for Change (LLMC) eliminate the problems with the classical pretest-posttet score differences and provide additional information for validation and interpretations of the results (see, e.g., Fischer, 1995). The theoretical framework of the LLMC is not given here because of its relative complexity and prerequisites of psychometric background for the reader (see, e.g., Fischer, 1995; Fischer & Ponocny-Seliger, 1998).

Two important concepts for the interpretation of LLMC reported in this study are the *treatment effect* and the *trend effect*. The treatment effect measures changes in students' ability due to the treatment. The trend effect measures changes in students' natural trends independent of the treatment. In item response theory, the term *ability* does not imply a general ability, but instead implies the theoretical ability that underlies a student's performance on a given test (see, e.g., Hambleton, Swaminathan, & Rogers, 1991, p. 77). The ability score of a student determines the probability for this student to answer correctly any given test item. The units of the ability scale, called "logits", typically range from -4 to +4. They represent the natural logarithms of odds for success on the test items. For example, if a student succeeds on 75% and fails on 25% of the test items, the odds ratio for the test is 3/1 = 3. Thus, the ability score of this student in logits is the natural logarithm of 3, which is 1.10 (i.e., about one unit above zero on the logit scale). Both treatment effects and trend effects are presented in logits with the LLMC for measurement of change in student's ability.

In order to compare the *Astronomy Village* treatment group to the alternative treatment group, we collapsed the four subscales into two subscales related to content and problem solving. The two treatment groups, *Astronomy Village* and alternative treatment groups, were compared against the same control group in the LLMC design for the calculation of treatment effects and trend effects on each of the two scales: content understanding (22 items) and problem solving (40 items). The LLMC calculations were performed using the computer program LPCM-WIN 1.0 (Fischer & Ponocny-Seliger, 1998). Cronbach's alpha for the reliability of measurements was adequate for the purposes of these comparisons, with the values of .80 and .97 for the content understanding scale and the problem solving scale, respectively. The results in Table 1 show that there were no statistically significant trend effects on either of the two scales for either of the treatment groups. This finding is not a surprise given the short time period of 4 weeks between the pretest and the posttest. Indeed, one can hardly expect changes in students' science ability due to trend effects such as natural maturation and cognitive development within a short period of time. On the other hand, there was a statistically significant effect due to treatment for each of the two treatment groups on each of the two scales, content understanding and problem solving. Figure 1 shows a bar graph of the treatment effects by subscale and treatment group.

The largest treatment effect, 0.647 (in logits), was associated with the alternative treatment group on the content understanding scale. Compared to the treatment effect of the *Astronomy Village* treatment group on the same scale, 0.511 (in logits), the ratio of the two effects shows that the students from the alternative group have 1.27 times better odds for success on the content understanding items than the *Astronomy Village* students. This inference is legitimate since the logit scale on which effects are measured with the LLMC is a ratio (proportional) scale. One can also say that the odds for success on content understanding are 27% better in favor of the alternative treatment group. Alternatively, the *Astronomy Village* students did slightly better on the problem solving scale than the alternative treatment group students with a treatment effect of 0.479 (in logits) versus 0.455 (in logits). In other words, the *Astronomy Village* students have about 5.3% better chances (in odds) for success on problem solving items than the alternative treatment group students.

The comparison across scales shows that both the *Astronomy Village* and the alternative treatment group students gained more in content understanding than in problem solving. The difference between the treatment effects in content understanding versus problem solving was more sizable for the alternative treatment group (0.647 versus 0.455, in logits) than for the *Astronomy Village* group (0.511 versus 0.479, in logits). Prior research has demonstrated the importance of well-organized content understanding for success at problem solving (Shin, Jonassen, & McGee, in press; Hong, McGee, & Howard, 2001). This would suggest that if students are applying new content understanding to solving problems, the ratio of the content treatment effect to problem solving treatment effect would be near one to one. A ratio greater than one would suggest that students are learning new content that is inert and not accessible during problem solving. The ratio for the alternative treatment group was 1.42, suggesting a high degree of inert knowledge. Alternatively, the ratio for the *Astronomy Village* treatment group was 1.07, suggesting that by learning content in the context of problem solving, the *Astronomy Village* students were learning content in such a way that is was accessible during problem solving.

4 Conclusion

Overall, we were very excited about the large treatment effect sizes for both the *Astronomy Village* treatment group and the alternative treatment group. This indicates that students are learning a great deal from the use of *Astronomy Village.* Although, it may appear discouraging that the treatment effects for the alternative treatment group and the *Astronomy Village* treatment group were very similar, it is important to keep in mind that the materials the alternative treatment group used were in themselves innovations. There does not exist any curriculum materials that teach the concepts in *Astronomy Village* using an interdisciplinary approach. The alternative treatment group materials still maintained the context of the overarching questions of the surface of Pluto and the search for life. If students were using traditional content materials presented with no context, we would expect to see the content performance to be very similar, but we would expect there to be an even greater disparity on the problem solving effect in comparison to the *Astronomy Village* treatment group.

The results of this research are consistent with prior discussions about the discrepancy between the stated goals for education, namely scientific inquiry, and the superficial nature of prevalent curriculum approaches, namely lecture and cookbook labs. This has been characterized in TIMSS reports as a mile wide and inch deep curriculum. Discussions dating back to the Sputnik-era of science education reform argue that "less may be more," meaning that a focus on inquiry may mean less content covered, but more is actually learned (Morrison, 1963). Our research points out that there is not necessarily a tradeoff in the quantity learned, but a qualitative difference in what is learned. The evidence indicates that the content learned by the alternative treatment group, which simulated traditional instruction, may be inert and not readily available for problem solving. Whereas, the content learned by the *Astronomy Village* treatment group may be more useful for problem solving.

The strength of the strategies we used to overcome the comparison group problem and the appropriate assessment problem lies in the formative information that is provided to teachers and policy makers. Our evaluation points to specific outcomes that are matched to the approach represented by the treatment group. This information can be used by teachers and policy makers to make adjustments in how they implement *Astronomy Village* with future groups of students so that there is a better match between the stated objectives and the learning outcomes. In the long run, the test of innovations such as *Astronomy Village* is not necessarily how well students in a given year performed relative to a comparison group, but rather how well do teachers make adjustments from one year to the next so that student performance continues to improve over time.

5 References

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Ed.). (1999). *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Research Council.

Fischer, G. H.(1995). Linear logistic models for change. In G. H. Fischer and I.W. Molenaar (Eds.). *Rasch models. Foundations, recent developments, and applications* (pp. 157-180). New York: Springer-Verlag.

Fischer, G. H., & Pnonocny-Seliger, E. (1998). *Structural Rasch modeling. Handbook of the usage of LPCM-WIN 1.0.* ProGamma, Netherlands.

Hickey, D. T., Wolfe, E. W., & Kindfield, A. C. H. (1999). Assessing learning in a technology-supported genetics environment: Evidential and systemic validity issues. *Educational Assessment, 6(3),* 155-196.

Hong, N.S., McGee, S., & Howard, B.C. (2001, April). Essential components for solving various problems in multimedia learning environments. Presented at the annual meeting of the American Educational Research Association, Seattle, WA.

McGee, S., & Howard, B. H. (1999, April). Generalizing activity structures from high school to middle school science. In S. McGee (Chair), Changing the game: Activity structures for science education reform. Symposium presented at the annual meeting of the American Educational Research Association. Montreal, Canada.

Morrison, P. (1963). Less may be more. *American Journal of Physics, 31,* 441-457.

Sheppard, L.. (2000). The role of assessment in a learning culture. *Educational Researcher, 29(7),* 4-14.

Shin, N., Jonassen, H.D., & McGee, S. (in press). Predictors of well-structured and ill-structured problem solving in an astronomy simulation. *Journal of Research in Science Teaching*.

6 Appendix

Complex Content

If a planet has a thick atmosphere, you would expect to see

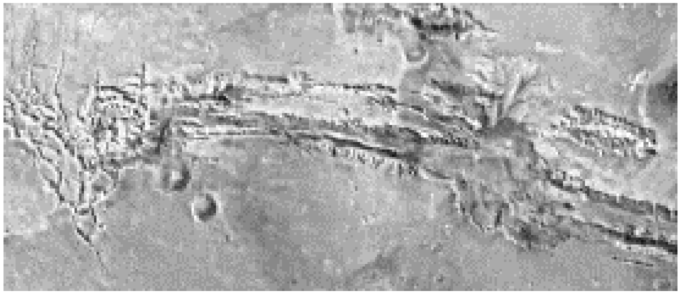
a) distinct craters such as those seen on the Moon or Mercury.
b) few craters such as those seen on Earth or Venus.
c) broad, shallow craters.
d) broad, deep craters.
e) narrow, deep craters.

Why do scientists believe that if life forms exist elsewhere in our solar system, the life forms will likely be very small?

a) Most living things on Earth can only be seen with a microscope.
b) Nature gives smaller organisms better protection than larger ones.
c) Smaller organisms are less complex than larger ones and have fewer energy requirements.
d) If organisms were large, such as a large animal, they would have probably been discovered by now.

Problem Solving

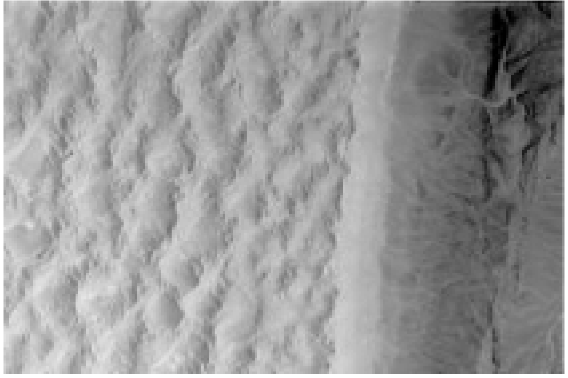
Image 1: Surface of Edina



Edina is the eighth planet in a solar system orbiting an imaginary star. The diameter of Edina is about 1/3 that of Earth. Image 1 of Edina was taken by a satellite orbiting the planet. The canyon structure in the image suggests that

a) In the <u>past</u>, Edina's climate has been warmer and wetter.
b) Samples of rocks from Edina would contain radioactive elements.
c) Edina was formed through accretion.
c) Edina must <u>currently</u> have water.

Image 2: Pagano



Which of the following statements is most likely true about the imaginary planet Pagano (based on Image 2)?

a) The surface of Pagano is composed mainly of sedimentary rocks.
b) Pagano has a cold, inactive core.
c) Pagano has winds on the surface that have eroded the rocks.
d) Pagano has many desert areas but not many mountainous areas.

Table 1: Treatment and trend effects from pretest to posttest by treatments and scales

	Content		Problem Solving	
Treatment Group	**Treatment**	**Trend**	**Treatment**	**Trend**
Astronomy Village	0.511**	0.105	0.479**	0.123
Alternative	0.647**	0.105	0.455**	0.124

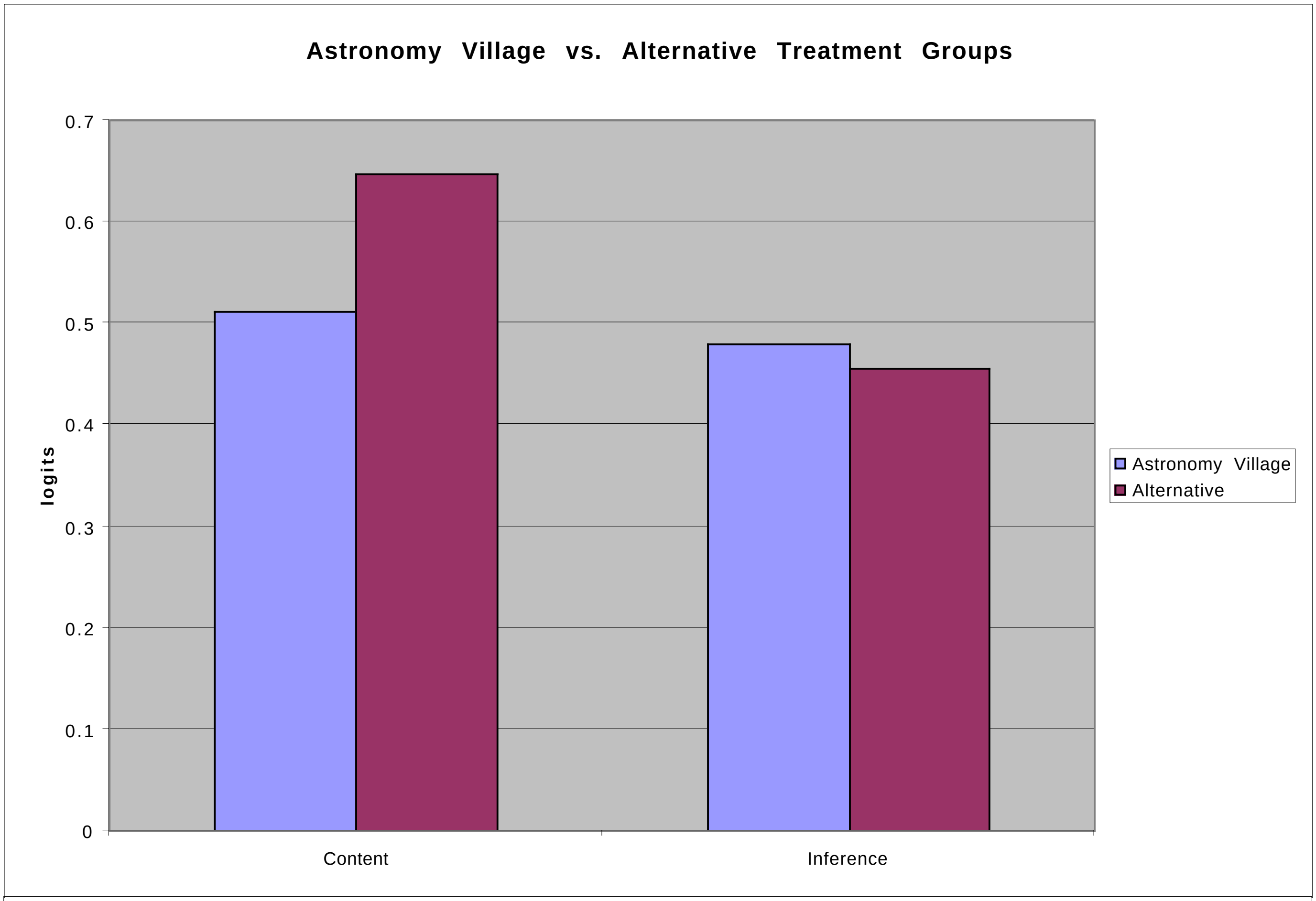


Figure 1: Comparison of pretest to posttest treatment effects for the content and problem-solving subscales by treatment group.