Using Design Experiments to Investigate Long-Term Program Success

Steven McGee
Center for Educational Technologies®
Wheeling Jesuit University

Dimiter M. Dimitrov
Kent State University

Jennifer Kirby & Steven K. Croft
Center for Educational Technologies®
Wheeling Jesuit University

http://www.cet.edu/research

ABSTRACT

The purpose of this study was to track teachers using *Astronomy Village®* over two school years. Prior research indicated that teachers were successful with the students when they enjoyed extensive support from the project developers. What would happen when the project support was no longer available? Would their students be as successful? Would the teachers continue to implement the program in ways consistent with the developers' intentions? This study used a hybrid experimental design that combined quasi-experimentation and design experiments. The experimental designs were supported by statistical analyses using the Linear Logistic Model for Change. The results indicate that teachers made significant adaptations to their implementation. These implementations were all consistent with the intent of the curriculum designers. An analysis of the learning outcomes indicates that students in the second implementation year achieved greater learning outcomes than students in the first implementation year. Researchers will continue to monitor the progress of these teachers during a third implementation year.

INTRODUCTION

The long-term success of any educational program depends on the extent to which teachers can implement the program without direct support from the program developers. A successful summative evaluation should not be the stopping point for a project. It is important to monitor the program after the developers no longer provide direct support to see if the participating teachers can continue to be successful. The summative evaluation of the National Science Foundation-funded Astronomy Village®: Investigating the Solar System™ indicated that students who used the program not only significantly improved in their understanding of complex solar system concepts, they also improved in their ability to engage in inquiry on the solar system (McGee et. al., 2001). In this study we were interested in the extent to which Astronomy Village teachers could continue to achieve significant student learning outcomes when they no longer enjoyed extensive support from the program developers.

Theoretical Framework

The design experiment approach, a recent advancement in educational research, is a powerful means to conduct ongoing research and evaluation of educational multimedia (Brown, 1992; McGee & Howard, 1998). Using this approach, teachers implement a program and evaluate the impact of the program on student learning. By reflecting on student performance, teachers can identify areas of weakness and make adjustments for the next implementation of the program. A new design experiment cycle begins as teachers implement the program using the new adjustments and then once again evaluate student performance.

The design experiment approach stands in contrast to randomized experiments in which researchers attempt to isolate the effects on a given phenomenon in order to generate a causal explanation. In randomized experiments researchers can manipulate only one variable at a time to determine the effect of that one factor. In contrast, the primary goal of design experiments is to create classroom conditions that will lead to increased learning outcomes. It is often necessary to manipulate several variables at the same time, making it difficult to determine the effects of any one variable. Through the design experiment approach teachers are free to make principled changes to their instruction in an attempt to improve instruction. Qualitative analysis techniques allow researchers to characterize the nature of the changes from one year to the next and from one classroom to the next.

In addition to the comparison from one year to the next, which is a hallmark of the design experiment approach, we have integrated a quasi-experiment design approach. Each participating teacher recruited a matched no treatment control group. The control students were administered the same pre- and posttest measures at the same times as the *Astronomy Village*® students. With the quasi-experimental approach we are able to remove any non-treatment effects within a given implementation, such as maturation or history effects. Once the non-treatment effects are removed, it is reasonable to conclude that any remaining improvement is because of the implementation of *Astronomy Village* by the teacher.

METHOD

Astronomy Village®: Investigating the Solar System™

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). 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 see 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. Students complete the investigation by hosting a virtual press conference in front of a virtual press corps that asks questions about the investigation students just completed. 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.

Teachers using *Astronomy Village*® have adopted one of two basic approaches. In the first students complete the core research project and then complete each of the focused investigations related to the core research project. In the case of Search for Life, there are four focused investigations. In the second approach students complete the core research project and then complete just one focused investigation. The teacher ensures that each of the focused investigations has at least one project team working on it. The students then host a press conference for their peers so that all of the students have the opportunity to learn the content in each of the modules.

Design Experiment

Of the seven teachers who conducted the Search for Life core research project during the summative evaluation (1999/2000 school year), we identified three teachers who also conducted the Search for Life core research project during the 2000/2001 school year (post-summative evaluation). Two of the teachers (NK and CK) served as *Astronomy Village* teachers during the summative evaluation. NK teaches middle school science in the western United States. CK teaches middle school science in the eastern United States. The third teacher (MD) served as an alternative treatment teacher during the summative evaluation (see Dimitrov, McGee, & Howard, 2002, for details on the summative evaluation). MD also teaches middle school science in the western United States. She is part of a team-teaching environment.

We were not able to include all of the summative evaluation teachers in this study. Some of them did not conduct the same core research modules in both years. Other teachers stopped using *Astronomy Village* because their teaching assignment changed to courses that were no longer appropriate for *Astronomy Village.* The remaining teachers continued to use *Astronomy Village* but chose not to participate in this study.

The data from the summative evaluation provides the baseline for this study. During the summative evaluation we asked teachers to spend approximately four weeks having students conduct as many of the investigations as they could during that time period. In most cases the students completed the core research project and 2-3 focused investigations. All of the students conducted the same modules. In the case of MD, who served as an alternative treatment group, the students completed the background research activities for both Search for Life and Mission to Pluto, but not any of the data analysis activities.

In the postsummative school year all three teachers made significant changes to the manner in which their students conducted the investigations. These approaches are summarized in Table 1. Both NK and CK had students complete only the core research investigation and one focused investigation. In both cases the students were expected to learn all of the content in the entire Search for Life core research area. The students took responsibility for learning the content in their focused investigation and teaching that content to the other students in the class. The teachers ensured that each of the focused investigations was conducted by at least one group.

MD conducted the investigations in a manner similar to the summative evaluation. The students conducted the core research investigation as a class. Then they proceeded to complete all four of the focused investigations in sequence. One significant difference between MD's implementation and the summative evaluation is that she had access to four student teachers. Each of the student teachers was assigned to one focused investigation. The students worked in small teams and rotated through each of the stations managed by All three of the teachers devoted a similar amount of instructional time to *Astronomy Village.* For the purposes of the subsequent analyses, the three teachers were considered as a group. Although the qualitative descriptions indicate that there were important differences in the context of use, the sample size of three teachers is not large enough to allow analyses that would tease apart the influence of these differences upon learning outcomes.

Table 1: Postsummative Implementation (2000/2001 School Year)

Teachers	Gender	Instructional Time	Structure
NK	F	6 wks	Students completed one focused investigation and presented it to the whole class.
CK	F	5 wks	Students completed one focused investigation and presented it to the whole class.
MD	F	5 wks	Students completed all four focused investigations.

Assessment Instrument

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 *Astronomy Village*® 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 Villag*e.

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, in press; 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. 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. This study focused on the Search for Life complex content and problem-solving subscales.

RESULTS

Our analyses combined the design experiment and quasi-experimental approaches. In the quasi-experimental approach we compared the pretest to posttest improvements within a given school year to a no treatment control group. Once it was established that there was an effect because of the implementation over and above effects because of maturation, we compared the learning outcomes from one year to the next.

We used the Linear Logistic Model for Change (LLMC) to analyze pretest to posttest changes in learning outcomes. Benefits of LLMC include information about the magnitude of the changes on a ratio scale and separation of changes because of treatment from changes from natural trends across time points of measurements (e.g., pretest and posttest). The theoretical framework of the LLMC is not presented here because of its relative complexity and prerequisites of psychometric background for the reader (see, e.g., Fischer, 1995; Fischer & Ponocny-Seliger, 1998). Provided are basic concepts and interpretations of LLMC results reported in this study.

In the item response theory context, the term ability connotes a latent trait that underlies the student's performance on a test (e.g., Hambleton, Swaminathan, & Rogers, 1991). The ability score of a student relates to the probability for this student to answer correctly any test item. The units of the ability scale, called "logits," typically range from -4 to +4. The results of the analyses are subdivided into <u>treatment effect</u> and <u>trend effect</u> that measure ability changes because of treatment and natural trends, respectively. The trend effect accounts for factors such as natural biological maturation and cognitive development between pretest and posttest measurements. The ratio of two effects indicates <u>how many times</u> one of the effects is greater (or smaller) than the other effect. In this study the LLMC calculations were performed using the computer program LPCM-WIN 1.0 (Fisher & Ponocny-Seliger, 1998).

The LLCM results in Table 2 show that there was a small, but statistically significant, trend effect (0.114, $p < .05$) on the content understanding scale. Thus, a small change in the students' ability on content understanding may be attributed to factors other than the treatments associated with the two implementation years. After controlling for these trend effects, there are still substantial treatment effects for both implementation years on both content understanding and problem solving. Comparing the differences in treatment effects from one implementation year to the next, we see significantly higher learning gains in the second year. On the content subscale students in the second implementation year improved 1.38 times more than students in the first year (1.371/0.993). On the problem-solving subscale students in the second year improved 1.62 times more than students in the first implementation year (0.996/0.616 = 1.62).

	Content understanding		Problem solving	
Implementation Year	**Treatment Effect**	**Trend Effect**	**Treatment Effect**	**Trend Effect**
1999/2000	0.993** (0.072)	0.114* (0.051)	0.616** (0.052)	0.003 (0.030)
2000/2001	1.371** (0.088)	0.114* (0.051)	0.996** (0.054)	0.003 (0.030)

Table 2: Pretest to Posttest Treatment and Trend Effects for the 1999/2000 and 2000/2001 Implementation Years (* $p < .05$. ** $p < .01$)

Note. The standard errors of the effect estimates are given in parentheses. All treatment and trend effects are on the same ratio scale (in logits). The ratio of any two effects indicates how many times one effect is greater (or smaller) than the other effect.

CONCLUSION

The primary goal for this study was to determine the extent to which the *Astronomy Village®* teachers could continue to achieve significant student learning outcomes when they no longer enjoyed extensive support from the program developers. We used a hybrid approach to the investigation. The LLMC analysis supports both the quasi-experimental design and the design experiment. In quasi-experimental design the LLMC analyses separated the effects because of trends from effects because of the treatment. From these analyses we were able to conclude that the teachers achieved significant learning outcomes during each of the implementation years. In the design experiment the LLMC analyses model the learning outcomes on a ratio scale that allows for comparisons from one year to the next. From these analyses we conclude that not only did the teachers continue to achieve significant learning outcomes, their students' learning outcomes were greater in the 2000/2001 school year than they were in the 1999/2000 school year.

Typical of the design experiment approach, these teachers changed a number of factors between the two implementation years. There were different students. The students spent more time on the investigations. The basic curriculum structure was altered. The same test was used from one year to the next, so the teachers were more familiar with the specific questions asked on the test. (Prior research has indicated that as teachers become familiar with a test, their teaching gradually emphasizes the content on that specific test.) It is our belief that all of these factors played a role in the improved learning outcomes. The design experiment approach does not allow for analyses that tease apart the effects due to any one factor.

In subsequent years we will continue to collect data on student learning outcomes from these teachers as well as other teachers who continue to use the software from one year to the next. Will the teachers continue to achieve increases in learning outcomes at the same rate as they did from 1999/2000 to 2000/2001? Or will the classroom implementations stabilize such that teachers achieve the same level of learning outcomes from one year to the next?

As the number of teachers involved in this research increases, it will be possible to extend the quasi-experimental design approach to tease apart the effects because of some of the factors that vary from one teacher to the next. Finally, in subsequent years it will be important to expand the pool of assessment items and to vary them from one year to the next. This will ensure that teachers are teaching not to the specific content of the test, but to the underlying concepts required to successfully answer the questions on the test.

This research on the implementation of a specific program is in its nascent stage. The critical assumption underlying this research is that long-term reform involves much more than the design of curriculum materials. Not enough is known about how innovations evolve over the course of time in the face of strong market forces. The track record for innovations in science education reform has been one of early success, followed by gradual obsolescence as the original designers eventually fade away from the project (McGee, 1996). A long-term perspective will provide developers and reformers with a better understanding of how to achieve long-term success for new innovations.

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.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences, 2*(2), 141-178.

Dimitrov, D. M., McGee, S., & Howard, B. C. (2002). Changes in students' science ability produced by multimedia learning environments: Application of the Linear Logistic Model for Change. *School Science and Mathematics Journal*, *102*(1), 15-22.

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., & Ponocny-Seliger, E. (1998). *Structural Rasch modeling. Handbook of the usage of LPCM-WIN 1.0*. Netherlands: ProGamma.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE.

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*. Paper presented at the meeting of the American Educational Research Association, Seattle, WA. [Online] http://www.cet.edu/research/papers.html

McGee, S. (1996). *Designing curricula based on science communities of practice.* Unpublished doctoral dissertation, Northwestern University, Evanston, IL. [Online] http://www.cet.edu/research/papers.html

McGee, S. & Howard, B. C. (1998). Evaluating multimedia in the context of use. *Journal of Universal Computer Science*, *4*(3), 273-291.

McGee, S., & Howard, B. C. (1999, April). Generalizing activity structures from high school to middle school science. Symposium presented at the meeting of the American Educational Research Association, Montreal, Canada. [Online] http://www.cet.edu/research/papers.html

McGee, S., Howard, B. C., Dimitrov, D. M., Hong, N. S., & Shia, R. (2001, April).

Addressing the complexities of evaluating interdisciplinary multimedia learning environments. Presented at the annual meeting of the American Educational Research Association, Seattle, WA. [Online] http://www.cet.edu/research/papers.html

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*.