Journal of Universal Computer Science, vol. 4, no. 3 (1998), 273-291
submitted: 10/12/98, accepted: 28/12/98, appeared: 28/3/98 © Springer Pub. Co.

Evaluating Educational Multimedia in the Context of Use

Steven McGee, Ph.D.
NASA Classroom of the Future, Wheeling Jesuit University
United States
mcgee@cotf.edu

Bruce Howard, Ed.D.
NASA Classroom of the Future, Wheeling Jesuit University
United States
howard@cotf.edu

Abstract: Researchers at the NASA Classroom of the Future have been using the design experiment framework to conduct evaluations of multimedia curricula. This method stands in contrast to more traditional, controlled experimental methods of evaluating curricular reforms. The methodology presented here is integrated with Walter Doyle's [1983] notion of using academic tasks to describe how classroom activities impact student learning. We will report the results from a design experiment with a multimedia program developed at the NASA Classroom of the Future, and we will examine the methodologies that were used in the evaluation.

Key Words: Multimedia, Evaluation, Curriculum, Design Experiment, Methodology, Computer Uses in Education, Science Education

Categories: K.3.1, K.3.2, K.3.m

1 Introduction

Traditionally, the evaluation of curricula has been conducted using the „horse race" method. In this method, evaluators challenge a traditional curriculum „horse" to a race using an experimental curriculum „horse." The horses are placed in the starting gate by giving students a pretest that covers the learning objectives. The teachers in each condition run the horses, and a posttest serves as the finish line. Whichever curriculum horse achieves the greatest student-learning gains from pretest to posttest is declared the winner. If the experimental curriculum wins enough races across a variety of contexts, then reformers proclaim it superior to the traditional curriculum. Although the horse race method has been widely used, there are two fundamental assumptions associated with the this method of curriculum evaluation that have recently been called into question.

The first assumption that has been questioned regarding the horse race method is the feasibility of creating comparable experimental conditions and isolating the variable of experimental interest [Brown 1992][Collins 1990]. Both notions are derived from laboratory-based research and both translate poorly to an actual classroom setting. In a laboratory, researchers randomly assign subjects to either treatment or control groups to ensure that subjects in each group are comparable. Subjects in each condition are treated the same, with the exception of the variable of experimental interest. By ensuring comparability and isolating the variable of experimental interest, researchers can be confident that the superiority of one group over the other is attributable only to the treatment. The horse race method of evaluation uses principles of quasi-experimental research [see Cook and Campbell 1979] to apply experimental constraints to classroom settings.

When moving from the laboratory to the classroom, however, it is not feasible to meet the stringent requirements of ensuring comparability between groups or to isolate the variable of experimental interest. It is virtually impossible to create comparable groups using entire classrooms as a group, since each classroom is a unique learning context that is affected by factors such as state and local education policies, and the classroom teacher. Quasi-experimental researchers overcome this limitation by assuming that the entirety of instruction within each group was implemented identically, thus concealing differences in how teachers implemented the experimental curriculum. In addition, it is difficult to isolate only the variable of experimental interest, since reformers must often enact changes to multiple variables at once in order for the reform itself to be successful. Thus, the frenetic environment of real classrooms makes the laboratory style, horse race method of evaluation unrealistic for testing treatment effectiveness in real classrooms [Brown 1992].

The second assumption that has been questioned is that reform will occur through gradual acceptance of curricula deemed „superior" by the horse race method. The horse race evaluation method has been used repeatedly over the last century to promote reform. In science education reform, the goal has consistently been to help students engage in scientific inquiry and cease being passive receptors of information [McGee 1996]. In every era of reform, there have been pilot projects that have successfully engaged students in scientific inquiry. In those pilot projects, the horse race evaluation method has been used to demonstrate that students learned more and developed healthier attitudes toward science than did students in traditional science classrooms [e.g., Collings 1923][Bredderman 1983][Shymansky, Kyle, and Alport 1983]. However, contrary to the assumptions of the reformers, an investigation of today's science classroom indicates that teachers have not uniformly adopted curriculum materials that engage students in scientific inquiry [Bybee and DeBoer 1994], even though these materials were judged superior across a variety of horse races.

The horse race method overlooks an important reality: that teachers always need to adapt the materials they are given [Cuban 1993]. To accommodate the often conflicting demands of school boards, administrators, parents, and students, teachers must make decisions about what features of the reform materials to preserve and what features to adapt. However, the horse race method does not provide information about which features are most critical to the success of the implementation. Teachers are left with making an educated guess, about how to adapt the reform materials. Teachers can easily make the wrong adaptations, by neglecting or distorting critical elements of the innovation. Therefore, the horse race method fails to provide long-term support that would help teachers make prudent adaptations.

In order to use multimedia as a tool for supporting educational reform, it will be important that the evaluation of multimedia provide richer insights into performance than those of the horse race method. At the NASA Classroom of the Future (COTF) program, researchers have been pursuing ways of evaluating educational multimedia. The strength of their methodology lies in its ability to produce *reliable* evidence of superiority despite the frenetic environments of disparate classroom settings. Their evaluative techniques result in „implementation frameworks,“ which prescribe guidelines on how to implement the multimedia within classroom settings.

2 Conceptual Framework for COTF Evaluation Models

In order to successfully implement multimedia software that has been judged effective by the horse race evaluation method, teachers need to consider differences between their schools and the schools that participated in the evaluation. For example, if schools in the evaluation had a wealth of technology available, students could use the program individually. However, if a teacher’s school owns only a fraction of the technology used in the evaluation, they would have to make adjustments to implement the new multimedia software. Teachers would have to consider whether it would be as effective to place students at computers in pairs. Dozens of such considerations are inevitable during actual implementation. Without tools for comparing model implementations with their own, teachers may unknowingly alter a critical reform element when adapting software, and expected benefits may be attenuated. The once „superior“ new curriculum may be reduced to mediocrity and relegated to the bookshelf.

It is crucial that teachers be empowered with ways to successfully adapt multimedia, since nonuniformity of implementation across real classrooms is inevitable. Sections 2.1 and 2.2 describe two alternative evaluation models that underpin the evaluation methods being developed at COTF.

2.1 Academic Tasks: Solving the Problem of Multiple Curricular Levels

It is a well-documented observation that there are three levels of curricula and that the topics addressed within each differ significantly [Cuban 1993][Schmidt et al. 1996]. The curriculum as intended represents the content and processes outlined in official curriculum documents, such as the *National Science Educational Standards* [National Research Council 1996], as well as curriculum materials that implement the recommendations of the official documents [see Tab. 1]. The curriculum as implemented represents teachers’ and students’ experiences with the curriculum. Teachers adapt curriculum materials based on personal content goals, pedagogical beliefs, and the constraints imposed by their local schools [Schmidt et al. 1996]. In addition, students bring their own goals to the classroom and have varying experiences with the curriculum materials, resulting in learning outcomes that are unique to each student. The curriculum as measured represents the content and processes that appear on tests for measuring the amount of learning gain. The curriculum as measured is distinct from the other two levels because assessment materials contain only questions about content and processes for which it is possible to ask valid and reliable questions [Crocker and Algina 1986]. By focusing evaluation efforts on the curriculum as intended and measured, the horse race evaluation method fails to reveal the diverse ways in which teachers can implement the curriculum.

Curriculum Level	**Focus of Evaluation**	**Other Research in This Area**
1. Curriculum as Intended	Examine software itself, and its learning objectives	• „Official“ Curriculum [Cuban 1993] • „Intended“ Curriculum [TIMSS- Schmidt et al. 1996]
2. Curriculum as Implemented	Examine classroom use, without experimental constraints	• „Taught“ and „Learned“ Curricula [Cuban 1993] • „Implemented“ Curriculum [TIMSS- Schmidt et al. 1996]
3. Curriculum as Measured	Examine measures used to assess desired learning objectives	• „Assessed“ Curriculum [Cuban 1993] • „Attained“ Curriculum [TIMSS- Schmidt et al. 1996]

Table 1: How levels of curricula may affect evaluation

Examining curricula from a variety of levels is not new. It has been used in evaluations previously, albeit with various terminology [see Tab. 1 Column 3]. For example, [Cuban 1993] argues that curriculum reform has not been sustainable because reformers have failed to recognize that multiple curriculum levels exist. He refers to these levels as the „official,“ „taught,“ „learned,“ and „assessed.“ The Third International Mathematics and Science Study (TIMSS) also recognizes the existence of multiple curriculum levels. In the design of international assessment instruments, the TIMSS staff capture information about all levels of the curriculum so that assessment results can be interpreted in light of students experiences with the curriculum as implemented [Schmidt et al. 1996]. They refer to these levels as „intended,“ „implemented,“ and „attained.“

We concur with these researchers in believing that each level of the curriculum must be considered before conclusions are drawn about overall program effectiveness. At the level of the curriculum as intended, evaluators should consider improvements to be made to the software and should examine potential mismatches between expected learning objectives and actual learning outcomes. At the level of the curriculum as implemented, evaluators should examine how the software is used, without experimental constraints, and the degree to which this use mirrors what was designed. At the level of the curriculum as it is measured, evaluators should consider the validity of assessment instruments used to measure desired learning objectives.

Academic task research reveals the diverse ways in which teachers implement the curriculum (level 2). Academic task research was initiated by Walter [Doyle 1983] as a method of investigating the implemented curriculum by breaking down classroom procedures into individual, measurable units called *academic tasks.* Doyle defined academic tasks as composed of (a) the *products* that students are expected to produce, (b) the *resources* that are available to students while fulfilling the task, and (c) the *operations* that students are expected to perform to turn resources into the assigned products. Given the direct influence of academic tasks on student behavior, [Doyle 1983] proposed that focusing on academic tasks, as opposed to focusing on the official curriculum, would provide better explanations for how to support student learning [Doyle and Carter 1984]. He noted, „Tasks influence learners by directing their attention to particular aspects of content and by specifying ways of processing information“ [Doyle 1983, p. 161].

[Stein, Grover, and Henningsen 1996] provide an excellent example of how academic task research can be used to investigate the curriculum as implemented. They examined how mathematical tasks evolved as teachers introduce them to students and how those tasks evolved further as students completed them. Stein et al. analyzed the task goals and task operations that the teachers assigned, and then investigated the conditions under which students were not able to meet the teacher’s expectations. The results indicated that students often had great difficulty in meeting the demands of those tasks that most closely matched the [National Council of Teachers of Mathematics 1989] curriculum standards. In those cases, teachers often modified the task to make it more conventional, and in many cases, students completed the task in a rote fashion. Although more difficult, when students were successful at engaging in mathematical problem solving according to the NCTM standards, it led to greater improvements in learning outcomes [Stein and Lane, in press].

By focusing on the actual tasks that teachers assign, as opposed to the tasks suggested by software and curriculum materials, academic task researchers can get a more accurate account of how instruction influences student learning in classroom contexts. Academic task research also reveals differences among the various levels of the curriculum, whereas the horse race approach conceals differences. Through academic task research, evaluators can analyze the full spectrum of the curriculum in order to identify particular components that are essential for achieving desired learning outcomes.

The COTF model of evaluation includes two facets of academic task research that have been linked to student achievement: task completion and the level of task demand [Hiebert and Wearne 1993][Stein and Lane in press]. That is, we measure the degree to which students are able to successfully accomplish the task goal, and the degree to which such tasks present an appropriate level of cognitive demand. In addition, we report on the development of an indicator for task completion that can be used to monitor implementation of reform.

2.2 Design Experiments as an Alternative to Quasi-experiments

Design experiment research complements academic task research so that evaluators can be relieved of the quasi-experimental requirements of ensuring comparability between groups and isolating the variable of experimental interest. Through academic task research, it is possible to identify variables that are relevant to curriculum implementation. Through design experiment research, teachers can compare the efficacy of multiple variables at once using their own curriculum implementations as comparable groups. This comparison is important because teachers do not have the luxury of comparing their own teaching to that of other teachers. However, it is possible to compare their own improvements from one semester to the next. Thus, design experiments are a way for teachers to systematically test instructional manipulations and revisions, so as to adapt the implemented curriculum in a manner that is consistent with the curriculum designers' reform principles. Design experiments used in conjunction with academic task research can provide the direction that is necessary for sustaining reform at the classroom level.

Instead of comparing a treatment to a control group, teachers in a design experiment compare successive designs of their own curriculum [Brown 1992]. Teachers begin a design experiment by using their knowledge of content and pedagogy to design the best possible instruction for meeting the learning objectives or design goals. The resulting design becomes a working hypothesis that instantiates a teachers' ideas about how learning can take place through educational multimedia [Tanner 1997]. After implementing the instruction, design experiment teachers reflect on how well the outcomes match the design goals and then redesign their procedures by developing new working hypotheses. The next time the instruction is implemented, the teacher can compare how closely the results of the new implementation match the design goals as compared to the results of previous implementations. This design experiment cycle can continue indefinitely.

The success of the design experiment process rests on the soundness of the measures used to determine effectiveness. In the case of educational multimedia, research on academic tasks can provide objective measures of effectiveness. If there is overall improvement on objective measures from one year to the next, then it can be argued that the instructional modifications improved the quality of the instruction.

Throughout the history of reform, there are several examples of successful design experiments. John Dewey provides an early example of what would later come to be called a design experiment [Tanner 1997]. From 1896-1904, Dewey ran the University of Chicago Laboratory School. Dewey's fundamental belief was that students learn best when investigating problems of personal interest [Dewey 1916]. The lab school began as way for he and the lab school teachers to test this idea through successive designs of an elementary school curriculum. Dewey met with the lab school teachers on a weekly basis to help teachers articulate design goals for instruction and to help them reflect on the outcomes of instruction relative to the design goals. Through this iterative process, Dewey and the lab school teachers successfully developed authentic problems that balanced the learning objectives with student interest [Tanner 1997].

[Polman 1997] and [Linn and Muilenburg 1996] provide more recent examples of design experiments. Polman documented the curricular changes of one science teacher over a two-year period as he made a transition from traditional science instruction to a completely project-based approach. His students conducted three to four extended science projects per year. After each project cycle, the teacher modified the instructional supports based on the areas of difficulty for the students. Over time, the student projects more closely resembled the process of scientific inquiry. [Linn and Muilenburg 1996] describe the Computers as Learning Partner (CLP) project. The goal of the project was to teach students to use pragmatic models for investigating the distinction between heat and temperature. Over a ten-year period, the CLP group worked with the same middle school teacher to conduct a design experiment on a set of semester-long activities. At the end of each semester, the teacher assessed the percent of students who could accurately describe the distinction.

In each design experiment example, the teachers and researchers simultaneously manipulated multiple variables related to the curriculum. Working hypotheses were developed explicating how the design would achieve the desired outcome goal. These working hypotheses were tested through the process of implementation. Through multiple iterations of the same curriculum and a set of objective outcome measures, it was possible for the teachers and researchers to monitor whether the curriculum adjustments enhanced the overall quality of the instruction.

The design experiment and academic task research models described above served as the basis for methods used to evaluate a multimedia program developed at COTF called *Astronomy Village: Investigating the Universe.* As part of the design experiment, researchers conducted three studies across three semesters. The design experiment methodology allowed us to manipulate multiple variables related to the implementation, yet still compare the results of each implementation to determine whether the changes to the curriculum were enhancing the instruction. The comparability of the results was established using the academic task research methodology. In the next section we describe the *Astronomy Village* software. The remainder of the paper will describe the methodology we used to evaluate *Astronomy Village*. For a complete description of the results of the evaluation, see [McGee, Howard, and Hong 1998].

3 *Astronomy Village*: The Multimedia Evaluation Target

The NASA Classroom of the Future program (COTF) is a NASA-funded research and development center that specializes in the development and testing of educational multimedia for math, science, and technology education. In March 1996, COTF published a CD-ROM called *Astronomy Village: Investigating the Universe* for use as a curriculum supplement in high school science classrooms. It has been distributed to over 11,000 teachers, educators, and resource centers, and won *Technology and Learning* magazine's Science Software of the Year Award for 1996 [Technology and Learning 1996]. *Astronomy Village* uses the metaphor of living and working at a mountain-top observatory [the village] as the primary interface from which students investigate contemporary problems in astronomy [see Pompea and Blurton 1995][see www.cotf.edu]. Academic activities are designed to promote learning of both astronomical concepts and processes related to scientific inquiry. Students join a „research team“ and choose one of ten „investigations“ to complete. In the Stellar Nursery investigation, for example, students investigate how stars are born. For each investigation, students progress through five phases: background research, data collection, data analysis, data interpretation and presentation of results. For any given phase, there are from three to seven content-related activities to be completed before proceeding to another phase. The primary means of tracking progress through an investigation is the *Research Path Diagram*-- a chart that shows each phase of the investigation and icons to represent activities within each phase [see Fig. 1]. Students also have access to an electronic LogBook for recording their scientific notes and observations.

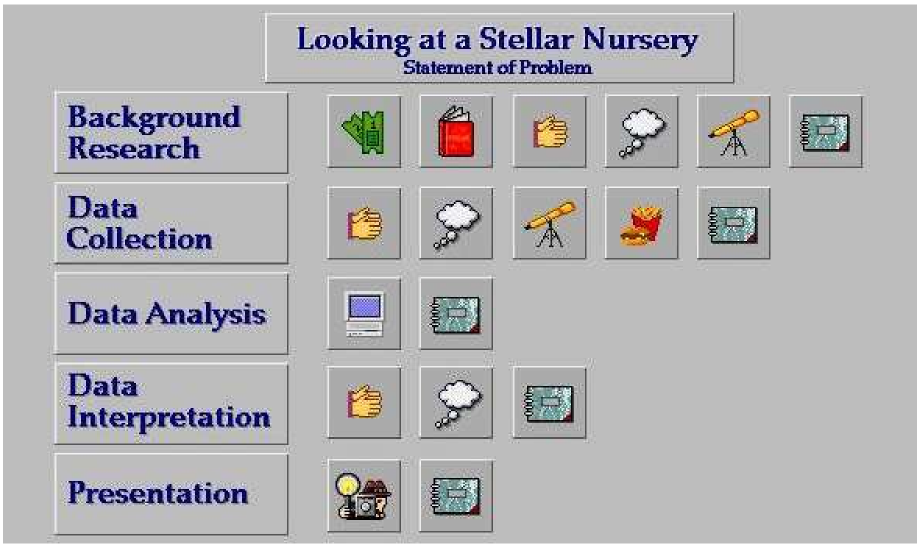

Figure 1: Stellar Nursery research path diagram

4 Method and Data Sources

In conjunction with academic task research techniques, researchers at COTF recently completed a design experiment using *Astronomy Village*. The design experiment spanned three semesters across two school years. In each study of the design experiment, two area schools used the COTF facilities, and students conducted astronomy projects using the software.

4.1 Design Experiment Populations

All three studies of the design experiment involved schools from a rural community with a population of approximately 35,000. The demographics for each study varied [see Table 2]. In the first study, thirteen students from the ninth grade class of a girls' academy (college preparatory) participated. In the second study, nine students from the eighth-grade class of the same academy participated. In the third study, twelve students from the tenth and eleventh grade of a large public high school participated. Students from the third study were from an at-risk population. In each case, students attended class daily for approximately four weeks at the COTF facility in lieu of their science class. Sessions using *Astronomy Village* were co-taught by the students' classroom teacher and the first author.

	Study 1	**Study 2**	**Study 3**
Time Frame	Apr-May 1996	Oct-Nov 1996	Mar-Apr 1997
Number of Students	13	9	12
Gender	All Female	All Female	F = 4; M = 8
Grade Level	9th	8^{th}	10th-11th
Type of Students	college preparatory	college preparatory	at-risk

Table 2: Demographics of Design Experiment Populations

It should be noted that the students in study 2 were younger than students in study 1, and in study 3 the population shifted from college preparatory students in a private school to at-risk youth in a public school setting. Based on previous research [Blank and Gruebel 1995], the population characteristics of each study would predict that the ninth grade students from the college preparatory academy should perform the best with *Astronomy Village*.

4.2 Data Sources

There were three sources of data for this investigation that enabled academic task research to support the design experiment. The first source was a compilation of documents in the students' electronic notebooks. In the first two studies, students used the electronic notebook embedded within *Astronomy Village.* This notebook is called the *LogBook*. In the third study, students used an electronic notebook called the *Collaboratory Notebook* developed as part of the Learning Through Collaborative Visualization Project (CoVis) Project at Northwestern University [Edelson and O'Neill 1994]. The second source of data was videotapes of student interactions while using the software. The third source of data was field notes and classroom observations by the teachers and first author.

For each academic task in *Astronomy Village,* students were expected to produce written responses in their electronic notebooks. Examples of notebook entries include activity summaries, answers to „press conference“ questions, answers to teacher-posed questions, and reflections on „thought experiments.“ Activity summaries consisted of a brief description of the activity, a statement of how the activity related to the main research question, and any new questions that arose from the activity. Press conference questions were posed to students by members of a simulated press corps to which they had to respond with answers from their investigation. Thought experiments were designed to prompt students to integrate path activities. Other collected virtual „data“ included images, snippets of articles they had read, or student reflections.

4.3 Dependent Measures

As indicated from prior research on academic tasks, it is essential that students be able to complete academic tasks in a cognitively demanding way in order for them to learn from the tasks [Hiebert and Wearne 1993][Stein and Lane in press]. The primary dependent measure for each study within the design experiment was an indicator of task completion called the task completion rate. This was an objective measure of the extent to which students were able to complete the task goals for the assigned academic tasks. Task completion rate was defined as the number of activity summaries that students completed divided by the number of academic tasks that the mentor suggested.

The task completion rate was mediated by the level of cognitive task demand associated with each academic task. In utilizing academic tasks, teachers must consider the cognitive demands of the task in the context of learning objectives. In order for learning to take place, there needs to be an appropriate amount of cognitive demand. If an academic task is too demanding, students will encounter a cognitive overload and will have difficulty completing the task [Sweller and Chandler 1994]. If an academic task is not demanding enough, students won’t learn from the task. Through an examination of the task completion rate, it became possible to identify academic tasks that might be too demanding for the students. In those cases where students were systematically not completing a task, the teachers provided instructional supports that would lessen the cognitive overload associated with those tasks. Student notebooks, teacher field notes and video tapes provided the data to make judgments about task demands.

5 Results of A Design Experiment with Astronomy Village

In brief, the design experiment began with instructional procedures in line with the *Astronomy Village* curriculum as it was intended, and examined the effects of the curriculum as it was consequently implemented. Each proceeding study then refined the procedures for more effective outcomes. In the description of the first study, there is a complete discussion of the instructional procedures used and the impact of that instruction on the task completion rate. In the descriptions of the later studies, there is a discussion of how instructional procedures were subsequently modified. The section concludes with a discussion of planned modifications to the *Astronomy Village* software based on results from the design experiment.

5.1 Study 1

In the first study (20 days in April-May 1996), the overall goal for the students was to complete all of the academic tasks as suggested by the virtual mentor. Our purpose in this study was to implement the curriculum as closely as possible to the curriculum as intended by the software designers. Students completed activities related to background research, data collection, data analysis, data interpretation, and presentation of results. As intended by the developers, students began by watching a videotape introduction to the software. Next, they worked within their project teams to complete a tutorial on using the software. After completing the tutorial, students proceeded to the virtual Conference Center in *Astronomy Village*, where the virtual mentors were available to describe the different investigations. The students selected different mentors to hear a description of the research that each mentor was conducting at the Village. The students chose an investigation that interested them and „signed-up" to join an investigation „team." Next, students were given access to the Research Path Diagram, to show them the resources on the *Astronomy Village* CD-ROM that related specifically to their investigation. By following the Research Path Diagram, students should have been able to complete the steps necessary to learn the appropriate concepts and conduct the level of problem solving needed for their investigation. Each recommended activity was defined as one academic task for purposes of this evaluation.

Within the first week of software use, the interplay between task completion and task demand became apparent. Students were expected to record the results of each academic task in their LogBook. However, the software did not provide any templates to help the students create LogBook entries. The students were either copying and pasting entire articles in their LogBook, taking detailed notes, or not writing anything in their LogBook. The teachers felt that none of these strategies were going to be effective for the students to be able to synthesize all of the investigation activities. Therefore, the teachers created an activity summary worksheet that students used to summarize each activity. For each summary, the students were asked to include a brief description of the activity, a statement indicating how the activity fit into their investigation, and any new questions that arose during the activity.

The activity summary template provided an instructional support to lessen the cognitive overload associated with the investigation. Activity summary worksheets made task completion easier, although completion rates were still below teacher expectations. The average task completion rate for this study was 42%. This value indicates that over the four-week period, students completed less than half of the activities that the mentor suggested. Further analysis of the task completion rate within each phase of research (i.e., background research - 55%, data collection - 75%, data analysis - 35%, data interpretation - 20%, and presentation - 27%) revealed that the task completion rate was much lower during later phases of research [see Figure 2]. This analysis of the task completion rate and task demand indicates that the Research Path Diagram and virtual mentors were not sufficient to guide students in complex problem-solving. If students are not able to complete the tasks in the path, they will not achieve the desired learning outcomes.

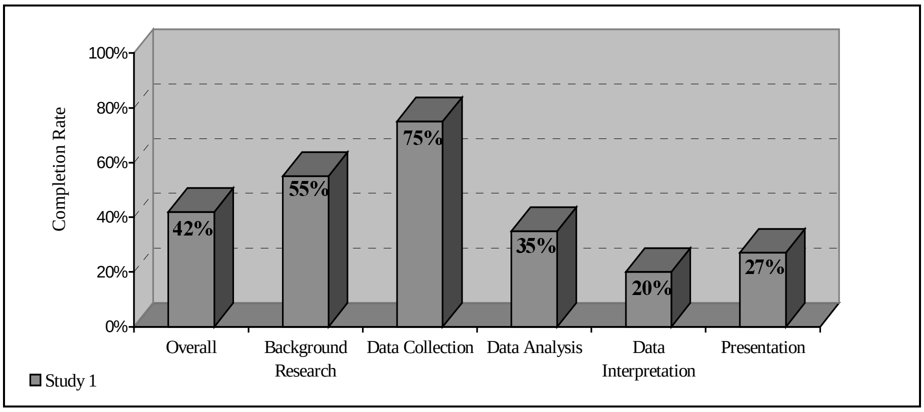

Figure 2: Task Completion Rate by Research Investigation Phase for Study 1

In examining additional data from the four weeks of instruction, there were several noteworthy observations. First, students spent most of their time in the background research phase of *Astronomy Village*, which left little time for the remaining activities. This was most likely due to the high task demand of background research activities, which hindered students' attention to time management issues. That is, students were observed to be engrossed in these early activities, to the exclusion of process-related discussions or questions that indicated that they were tracking individual activities in the context of the larger investigation. Second, students averaged 3.5 class periods to complete the tutorials and select their investigation. Student comments and questions indicated that what was learned during the advance tutorials was not remembered very well, if at all, when the time came to apply such knowledge. It was concluded that such tutorials, if needed at all, should be completed on an as-needed basis, thus lessening cognitive overload during the earlier phases, where demand was highest. In the next study, the teacher and first author used these results to guide a redesign of the task demands to meet the needs of the students.

5.2 Study 2

Based on the apparently poor performance of the students in Study 1, the teachers for Study 2 (20 days in October-November 1996) modified the curriculum slightly from the procedures intended by the designers of the *Astronomy Village* . The overall goal of the modifications was to relieve cognitive overload so that task completion could be improved. The modifications took three forms: target dates for phase completion, more time for task completion, and more contextualized training.

First, students were given target dates for each of the phases of research so that they would have sufficient time to complete later phases of research. This instructional support was intended to lessen cognitive overload during the phase of background research. Second, the teachers eliminated activities that did not seem to benefit the students. That is, students selected a pathway from a list of abstracts prior to beginning the study, and the tutorial was eliminated. These two changes resulted in approximately 17% more time for completion of academic tasks (3.5 instructional days out of 20). Third, training was given in the context of use, which was intended to lessen cognitive overload incurred by needing to remember various software procedures that would be used in later phases of research. Instead, the training was done at the beginning of each phase of research by the teacher demonstrating software features that would be needed.

The average task completion rate for this study was 85% [See Figure 3]. This value indicates that over the four-week period, students completed twice as many activities as the students in the first study even though these students were a year younger. The task completion rate within each phase of research was as follows, background research - 78%, data collection - 100%, data analysis - 83%, data interpretation - 62%, and presentation - 100%. The students in this study had more opportunity to achieve the desired learning outcomes than students in the previous study. Through an objective measure such as the task completion rate, it is possible to begin to investigate the factors that contribute toward students' abilities to complete the assigned academic tasks.

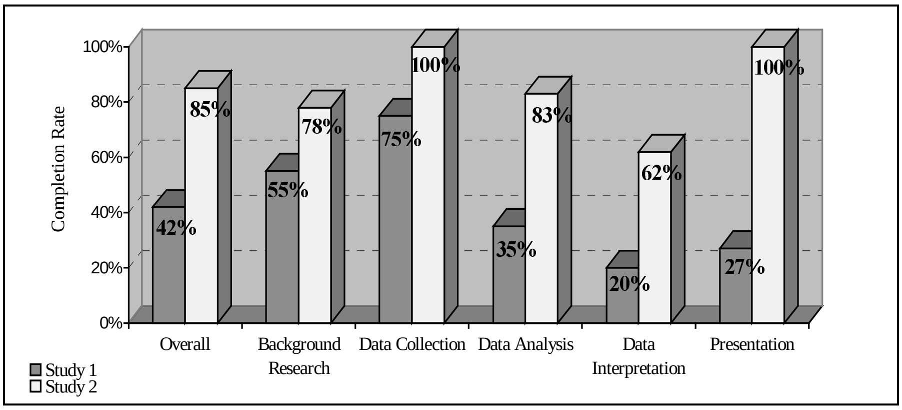

Figure 3: Comparison of task completion rate across Studies 1 and 2

At the end of the second study, researchers at COTF had the opportunity to review the videotapes, logbooks and field notes from both of the studies. Researchers used this analysis to begin to develop indicators of student learning. This review identified two important issues related to students engaging in scientific inquiry. First, within the activity summary, students were capable of developing brief descriptions of the tasks. However, they were not developing good statements of how the activity related to the overall research path. Second, there was no evidence that students were activating prior knowledge about the research question in order to build connections to these new experiences. The emphasis of the third study was on helping students build connections between prior knowledge, academic tasks and the research question.

5.3 Study 3

In Study 3 (19 days in March-April 1997), the researchers and teachers initiated several major modifications. The purpose of all modifications was to provide more structure to support the activation of prior knowledge and provide opportunities for students to complete multiple investigations in four weeks. Rather than have students select the paths to work on, the teacher and first author selected the pathways that students would work on, and all student teams investigated the same pathways.

In order to accomplish the goal of completing investigations more quickly, research phases were revised and truncated into five alternative phases: the motivating question phase, background research, background review, data analysis, and reflection. In the motivating question phase, the teacher posed the main investigation question and the students individually typed responses in their electronic notebooks. Next, the teacher showed the students the data that they would be analyzing and asked them to record observations. These two activities were meant to activate students' prior knowledge and connect it to the activities of the investigation. This approach is similar to Minstrell's benchmark lessons [Bruer 1993]. In the background research phase, the teacher selected the three most relevant articles from the pathways, and students each read one of the three articles and developed an activity summary. In the background review phase, students used their activity summaries to answer questions as a team that would prompt students to integrate across the readings that were done individually. Since each student was an expert on only one of the articles, students would be forced to discuss the readings with each other in order to answer the question. In the data analysis phase, students completed analysis worksheets as a team. And finally, in the reflection phase, students responded to integrative, teacher-posed questions in their notebooks. Using the redesigned pathways, it was possible to fit three investigations paths into the four-week period.

The average task completion rate for Study 3 was 74%. This rate is comparable to the task completion rate from Study 2, even though this group consisted of students who were at-risk. The task completion rate for each phase was as follows, motivating question - 82%, background research - 85%, background review - 62%, reflection - 73% [see Figure 4]. The instruction for this study involved more explicit prompts for students to reflect on their background knowledge and relate that knowledge to the tasks in the research path. These expanded prompts increased the likelihood that students would assimilate the new information from *Astronomy Village* into existing knowledge structures.

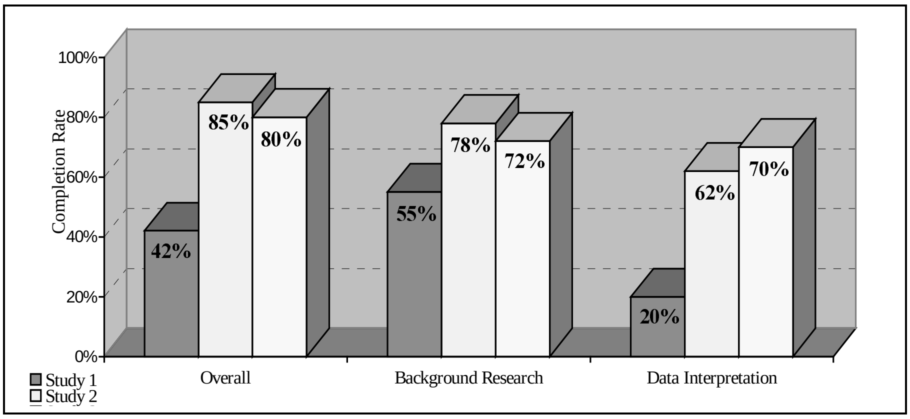

Figure 4: Comparison of task completion rate across all three studies

5.4 Design Changes to Astronomy Village

This design experiment resulted in planned changes to the *Astronomy Village* interface in subsequent revisions to the software. Currently, the logbook pages are unstructured. When a student requests a new page, they are given a blank textbox. This research has indicated that it is difficult for students to relate the content in *Astronomy Village* to their background knowledge. It is also difficult for students to use the logbook as a progress monitoring device. Through the use of writing prompts and questions, the teachers were able to increase students' ability to plan their project as well as reflect on how the content relates to their own experiences. The design experiment context has provided hypotheses for the types of prompts that might work the best. These hypotheses will be tested in more detail through laboratory and quasi-experimental research before being introduced into the software.

Another mechanism that was introduced to increase students' connections between their background knowledge and the content in *Astronomy Village* was the motivating question phase. This new phase allowed students the opportunity to explore the basic question of the path and come to understand why it was a problem. The success of the motivating question phase has resulted in the planned addition of an exploration phase to the research path diagram.

Finally, it was determined that a single tutorial at the beginning of the project was not sufficient for guiding students on how to use the *Astronomy Village* interface. In Studies 2 and 3, the teachers provided modeling and coaching on how to use the interface features at the beginning of each phase of research. This assistance made the instruction more immediate to when students were going to be using that knowledge, and it required students to hold less information in memory at the same time. In subsequent versions, the tutorials will be distributed within each of the Village buildings so that students will learn about the interface for that building upon first entering the room.

6 Conclusion

6.1 Summary

The main findings of this study indicate that students need structure in processing individual activities in the context of a multi-week curriculum centered around researching a singular scientific question. By providing this support, along with deadlines for task completion, students were able to finish more and understand more. In the third study, by scaffolding conceptual development through building links to prior knowledge, the teacher was able to maintain task completion rate while altering the level of task demand and cover more investigations with a more difficult population. Based on these findings, it is recommended that future implementations of this curriculum use the model of building a structure for scaffolding and fading (like activity summaries), activating prior knowledge (like asking motivating questions) and directing coaching of procedural knowledge, as needed for students to complete investigations.

6.2 Benefits of the Design Experiment Method

The results and conclusions of the design experiment would not have been possible using the horse race evaluation method. Through the horse race evaluation method, reformers can conclude whether use of an educational multimedia program will enhance student learning in particular contexts. However, every classroom presents a unique context that may or may not match the evaluatory contexts. Teachers must decide, therefore, how to adapt the multimedia for their classroom situation. The design experiment techniques being developed at the classroom of the future offer promise for providing teachers with the tools for evaluating their instruction.

The design experiment allowed the curriculum as it was intended to be studied so specific recommendations could be made for subsequent software revisions. In addition, the curriculum as it was implemented was examined, and teacher adaptations, such as the need to provide a notebook template, were evaluated in vivo. Further, the design experiment allowed for examination of the curriculum as measured. New measures such as task completion rate and judgments of task demand were explored and found useful. Through the research techniques demonstrated here, many limitations inherent in the horse race method were overcome.

Academic task research provides a uniform structure for describing the activities in which students engage in the classroom. Using the structure, it is possible to identify the goals that are assigned by the teacher and the extent to which the students were able to accomplish the goal. This is the task completion rate. This construct can be used to identify activities that are difficult for students to implement. The teacher can then design instructional supports that will help students to be able to accomplish the tasks.

Reform will only take place when teachers are able to explore the implications of a new curriculum within the context of their own classrooms. The fact that an educational multimedia program was used successfully in a neighboring school district does not mean that it will be effective in a new teacher's classroom. Design experiments are especially powerful if attempts are made based on design principles. Creating partnerships between teachers and researchers is useful for enhancing reform by giving teachers tools for evaluating their own implementations and providing researchers a means to compare implementations across different contexts.

6.3 Future Directions

When designing instruction, teachers need to balance the level of cognitive challenge of an academic task with students' abilities to complete the task. The most cognitively demanding tasks are those that are most difficult for students to complete. With the task completion rate indicator, teachers can monitor the extent to which students are capable of completing tasks and they can identify those parts of an academic task that students are having the most difficulty with. In the case of *Astronomy Village*, the task completion rate indicator served to alert the teacher and researcher to the fact that students were not engaging in efficient planning of their solution, preventing them from completing academic tasks that came later in the project cycle. Having been alerted to the problem, the teacher made adjustments in Study 2 that increased the task completion rate, without negatively affecting the cognitive demand of the task. Future research at COTF will focus on developing efficient indicators for monitoring cognitive task demand.

Just as the results of each individual study resulted in new hypotheses and promoted instructional redesign, the results of the design experiment resulted in working hypotheses for the next experiment. The results of this study will be used in further curriculum evaluations conducted by COTF to design models of appropriate task demands and to develop assessment instruments to examine the curriculum as intended and implemented.

Acknowledgments

This research was supported in part by grants from the National Aeronautics and Space Administration (NCCW-0012) and by the National Science Foundation (ESI-9617857). We would like to thank the teachers and students who participated in the design experiment. We would like to thank Brigitte Gegg for her assistance in conducting the studies and processing the data. We would like to thank John Hornyak and Steven Croft for answering our endless questions about astronomy content. We would like to thank all of our Wheeling Jesuit University student interns who helped in analyzing the data. We would like to thank Dorothy Frew and Pat Carlson for their helpful comments on earlier drafts of the paper.

References

[Blank and Gruebel 1995] Blank, R. K., and Gruebel, D.: „State Indicators of Science and Mathematics Education 1995"; Washington, DC: Council of Chief State School Officers. (1995).

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

[Bredderman 1983] Bredderman, T.; „Effects of activity-based elementary science on student outcomes: A quantitative synthesis"; Review of Educational Research, 53, 4, (1983), 499-518.

[Bruer 1993] Bruer, J. T.: „Schools for thought: A science of learning in the classroom"; Cambridge, MA: The MIT Press (1993).

[Bybee and DeBoer 1994] Bybee, R. W., and DeBoer, G. E.: „Research goals for the science curriculum"; In D. L. Gabel [Ed.], Handbook of Research on Science Teaching; New York: Macmillan Publishing Company (1994)

[Collings 1923] Collings, E.: „An experiment with a project curriculum"; New York: The Macmillan Company (1923).

[Collins 1990] Collins, A.: „Toward a design science of education" [Technical Report #1]; Cambridge, MA: Bolt Beranek and Newman (1990, January).

[Cook and Campbell 1979] Cook, T. D. and Campbell, D. T.: „Quasi-experimentation: Design and analysis issues for field settings"; Boston: Houghton Mifflin Company (1979).

[Crocker and Algina 1986] Crocker, L. and Algina, J.: „Introduction to classical and modern test theory"; Orlando: Harcourt Brace Jovanovich (1986).

[Cuban 1993] Cuban, L.: „The lure of curricular reform and its pitiful history"; Phi Delta Kappan, (1993), October, 182-185.

[Dewey 1916] Dewey, J.: „Democracy and education"; New York: Free Press (1916).

[Doyle 1983] Doyle, W.: „Academic work"; Review of Educational Research, 53, 2, (1983), 159-199.

[Doyle and Carter 1984] Doyle, W., and Carter, K.: „Academic tasks in the classrooms"; Curriculum Inquiry, 14, 3, (1984), 129-149.

[Edelson and O'Neill 1994] Edelson, D. C., and O'Neill, D.K.: „The CoVis Collaboratory Notebook: Supporting collaborative scientific inquiry"; In A. Best [Ed.], Proceedings of the 1994 National Educational Computing Conference (pp. 146-152). Eugene, OR: International Society for Technology in Education in cooperation with the National Education Computing Association, (1994).

[Hiebert and Wearne 1993] Hiebert, J., and Wearne, D.: „Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic“; American Educational Research Journal, 30, 2, (1993), 393-425.

[Linn and Muilenburg 1996] Linn, M., and Muilenburg, L.: „Creating lifelong science learners: What models form a firm foundation?“; Educational Researcher, 25, 5, (1996), 18-24.

[McGee 1996] McGee, S.: „Designing curricula based on science communities of practice“, Dissertation, Northwestern University (1996).

[McGee, Howard, and Hong 1998] McGee, S., Howard, B., and Hong, N.: „Evolution of academic tasks in a design experiment of scientific inquiry“; Paper to be presented at the American Educational Research Association. San Diego, CA (1998).

[National Council of Teachers of Mathematics 1989] National Council of Teachers of Mathematics: „Curriculum and evaluation standards for school mathematics“; Reston, VA: National Council of Teachers of Mathematics (1989).

[National Research Council 1996] National Research Council: „National science education standards“; Washington, DC: National Academy Press (1996).

[Polman 1997] Polman, J.: „Guided science expeditions: The design of a learning environment for project-based science“; Dissertation, Northwestern University (1997).

[Pompea and Blurton 1995] Pompea, S. M., and Blurton, C.: „A walk through the Astronomy Village“; Mercury, (1995 Jan-Feb) 32-33.

[Schmidt et al. 1996] Schmidt, W. H., Jorde, D., Cogan, L. S., Barrier, E., Gonzalo, I., Moser, U., Shimizu, K., Sawada, T., Valverde, G. A., McKnight, C., Prawat, R. S., Wiley, D. E., Raizen, S. A., Britton, E. D., and Wolfe, R. G.: „Characterizing pedagogical flow“; Dordrecht: Kluwer Academic Publishers (1996).

[Shymansky, Kyle, and Alport 1983] Shymansky, J. A., Kyle, W. C., and Alport, J.: „The effects of new science curricula on student performance“; Journal of Research in Science Teaching, 20, 5, (1983), 387-404.

[Stein, Grover, and Henningsen 1996] Stein, M. K., Grover, B. W., and Henningsen, M.: „Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms“; American Educational Research Journal, 33, 2, (1996), 455-488.

[Stein and Lane in press] Stein, M. K., and Lane, S.: „Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project“; Educational Research and Evaluation, in press.

[Sweller and Chandler 1994] Sweller, J. and Chandler, P.: „Why Some Material is Difficult to Learn“; Cognition and Instruction, 12, 3, (1994), 185-233.

[Tanner 1997] Tanner, L. N.: „Dewey's Laboratory School: Lessons for Today“; New York: Teachers College Press, (1997).

[Technology and Learning 1996] Technology and Learning, 17, 3, 1996.