# Evaluating Scientific Visualizations

**Robert Garfinkle**
Exhibit Developer at the Science Museum of Minnesota

**Vivian Johnson, Ph.D.**
. Science Coordinator, Physics Department, Augsburg College.

## Introduction

"Seasons" is a computer-based model of the natural processes that cause the seasons on Earth. The project makes sophisticated visualizations of scientific phenomena, usually reserved for scientists, accessible to the museum visitor. "Seasons" is permanently installed in the Experiment Gallery at the Science Museum of Minnesota.

Having nearly completed the development process for this exhibit, we have found that such scientific visualizations may benefit from different evaluative tools and practices than other computer-based interactives. This paper will describe our formative evaluation process to date. The ways in which the formative evaluation has failed are the most instructive, and the paper will look closely at why such failures occurred. Some implications for future development projects and the use of appropriate evaluation methods will be discussed.

## Project Description

"Atmospheric Explorations" is a collaborative project of Augsburg College and the Science Museum of Minnesota, designed to produce three exhibits that demonstrate weather-related physical processes. The first of these exhibits, "Seasons," is the subject of this paper. This exhibit seeks to give the museum visitor a highly interactive, stimulating experience that reinforces the concept that seasons on Earth are caused by variations in solar energy brought about by the tilt of the Earth's axis. It does this by allowing visitors to design their own "seasons" by manipulating the tilt of the Earth's axis and the shape of its orbit. A numerical model is then executed and the visitor is shown a visualization of temperatures on the Earth's surface and other graphics that reinforce the cause and effect relationships. It is the visualization of a scientific model that is at the heart of the exhibit, as well as its most visually appealing feature.

The visitor is invited to explore the possible Earth seasons as they please, making whatever changes in tilt and orbit shape that interest them. While the exhibit first presents the visitor with tilt and shape values that are appropriate to Earth, the visitor is then free make changes, such as tilting the Earth on its side, or elongating the Earth's orbit past that of Mars. There is no limit on the number of model runs that the visitor can make, and a "logbook" utility is provided for systematic comparison of different runs if the visitor desires. A guided tour of the exhibit is available, as are scientific explanation screens (some of which are interactive) and a description of the energy budget model.

The exhibit makes use of both animation and sound, and relies most heavily on scientific visualization. The primary example of this is a representation of the Earth (with continental outlines to show its spin) in orbit about the Sun. The visitor can select to have the surface of the Earth colored according to temperature, the amount of incoming daily solar radiation, or the hours of daylight (all as computed by the model). The combined effect of seeing the spinning Earth orbiting the Sun, its surface color changing as the temperatures change throughout the year, is often mesmerizing to visitors. A complete year of seasonal changes in temperature is displayed in about 20 seconds.

## Evaluation Plan for Seasons

At the beginning of the "Atmospheric Explorations" project, we conducted focus groups with children and teachers to understand what they understood about weather-related phenomena, and what they found interesting. We discovered that most children are confused about basic facts concerning the seasons: what their origin is, how to define them, etc. We knew from this that it would be challenging to design an exhibit about seasons.

Against this backdrop, we decided to evaluate "Seasons" at every stage possible. The purpose of this strategy was to minimize the amount of time-consuming re-programming work that would be needed, and to keep testing our concepts against the misunderstandings we knew were prevalent. The more we could learn early on, we reasoned, the less coding and re-coding we would need to do later. We chose evaluation methods that are commonly employed in the development of hypermedia and multimedia programs. Formative evaluation was planned in three basic stages:

1. *paper tests*—Using paper mockups of screens, testers would be walked through the exhibit to evaluate basic navigation and logical flow;
2. *screen displays*—Using static screens from the exhibit, testers would walk through the exhibit to further evaluate its interface design, navigation, and conceptual design;
3. *observation and interview*—Once a working version was completed, users would be systematically observed using the exhibit, and then interviewed about their experiences. In addition, basic

timing and group composition data would be gathered to assess exhibit attraction and holding power.

# Formative Evaluation Results

## Paper Tests

Once the development team settled on a basic flow for the program and a general layout for the screens, mocked-up screens were created on paper. These were then shown to a few museum visitors, and to museum staff and youth working in the museum, none of whom had any involvement in this project.

Purposes:

- To determine whether users understood how to navigate and interact with the exhibit; and
- To evaluate the representations on the screen.

Findings: The users who viewed the screens knew how to move through the exhibit, but they had little idea what the exhibit was about. The difficulty for the visitors of imagining what the screen would look like when animated and how the exhibit would operate was so daunting that we cut short the test.

Interpretation of Findings: It was the project teams belief that the paper screens themselves were not representing the exhibit well enough to regard any of the feedback as useful. The fact that nothing was moving on the screens severely hampered peoples ability to understand the exhibit. As a result, the paper tests gave us little useful information. The development team proceeded to work on the next stage without any real feedback from outside the project team.

## Static Screen Tests

The next stage was using static computer screen displays intended to look like screens from the completed exhibit. Since color is such an important way of representing the scientific model in Seasons, we thought the full-color screen displays would give us more information than the paper tests. There was no working model at this point, but the navigation and visual design elements were in place. The users for this test were similar to those taking the paper tests, although they were mostly different people: staff members, youth workers at the museum, and a few museum visitors.

Purposes:

- To determine whether users understood how to navigate and interact with the exhibit;
- To evaluate the screen design, especially the densely-packed main screen; and
- To collect preliminary feedback on the potential quality of the user experience with the working model.

Findings: Again, we found that the users who viewed the screens knew how to navigate the exhibit, but most still struggled with the basic concepts around the exhibit. They seemed to understand many of the parts, but missed the whole. Visitors did not understand what the program might do when it was fully functional. There were surprisingly few comments about the density of information on the main screen

Interpretation of Findings: It was the project teams belief that the static screen tests were, like the paper tests, highly confounded by the lack of movement and dynamism on the screens. The project team was concerned about lack of understanding. However, the test gave us no useful information about the degree of the problem, much less how to fix it.

The finding that density of information was not a problem surprised us. It gave us some license to retain the richness of the main screen, while still trying to thin out the number of buttons and choices to be made and thus call more attention to the critical elements

## Observation and Interviews

Once we had a working, viable model on the museum floor, we began structured observations and interviews. We developed and tested an observation protocol that allowed us to carefully watch and record what visitors did. We observed about 35 people, and interviewed about 28 of them directly afterwards. In addition, we timed people at the exhibit.

Purposes:

- To determine whether users understood how to navigate within and interact with the exhibit;
- To collect preliminary feedback on the potential quality of the user experience with the exhibit;
- To see how long visitors spend with the exhibit; and
- To see if users understood the conceptual framework of the exhibit.

Findings: We found that most of the 35 people interviewed understood the key ideas about the exhibit: that it was about the seasons and temperature change. A minority mentioned the difficulty level of the program; a number expressed that they wished they knew more about the topic before encountering the exhibit. From the timing part of the study we found that the average time spent at the exhibit was 4 minutes, 15 seconds, and the median time spent was 3 minutes, 45 seconds. Sixty-two percent of the 45 people observed spent better than 3 minutes, and 27 percent spent more than 5 minutes.

The exhibit was a highly social one. Visitors who spent extended periods of time at the exhibit were interacting with one another a great deal, pointing to the screen and tracing the movement of the Earth around the Sun.

We also found that the screen present when a visitor began using the exhibit had some effect on the length and quality of their interaction. The screen that worked best for first interactions was the main visualization screen. Its elements of color and motion proved attractive and stimulating. In contrast, people who started at the introduction screen, which was designed as the starting point of the exhibit, tended to leave the exhibit early.

Interpretation of Findings: We learned many things about the exhibit. Some questions were answered, more were sharpened, and many new questions were raised. Certainly, though, we gained a deeper understanding of how the exhibit is actually used by visitors. We learned that we need to explore further ways to ensure that more people understand the role of the Earths axis tilt in causing seasons. We found out that the exhibit leaves some people feeling inadequate to using it; we will certainly look for ways to change that. We learned any number of things about specific parts of the exhibit, too numerous to detail here.

The observation and interviews of the working model provided a rich set of data to inform the development process. There was no doubt, unlike the formative evaluation done with the static images, that this feedback was relevant and informative.

## Commentary

We have not found any useful formative evaluation tools appropriate for a pre-working model of a scientific visualization. All but the most generic of navigation and screen design questions are difficult to answer with the evaluation methods we tried. Evaluation conducted on the pre-working model is almost completely confounded by the inability to visualize the scientific process being demonstrated. Exhibits based on scientific visualizations may benefit from different evaluative approaches than those used for hypermedia or other information-based systems.

Looking back, it appears to us that there is an important distinction to be made between evaluation of this exhibit before and after the scientific model underlying it was implemented Building exhibits based on scientific visualizations is not a steady development process; the exhibit is qualitatively different once it has a working model.

Before the model is running we can only show static images, and the process relationships we are attempting to illustrate are hidden. Once the model is running, users can see the effects of their manipulations and can begin to grasp the scientific concepts being portrayed.

It is clear to us that the evaluation work done on the pre-working model was only marginally useful. But another question remains: why are such early formative evaluation techniques useful in hypermedia and multimedia projects, but not here?

The exhibit differs from most hypermedia and multimedia in that its presentation relies heavily on motion and conveying a sense of spatial abstraction. The paper studies and static screens proved ineffective because the gist of this exhibit is so intimately bound up in watching the abstraction and mapping it into the visitor's physical reality. Visitors had no chance to understand it unless they saw the abstraction. Flat paper screens didn't convey it—they were poorer than a textbook at conveying the abstraction. Static screens lacked motion and therefore both the temporal and spatial elements needed for understanding. Only the running exhibit could begin to supply the visitor with these factors. Of course, the extent to which the visitors understood the exhibit was also strongly influenced by the environment, the introductory screens, etc., but those elements also exist for other types of exhibits.

The implication of this finding is that developing a scientific visualization for a museum setting requires a difficult choice. Evaluation may be useless or confusing if the prototype is incomplete (i.e., some features are missing or not all visualization elements are in place). On the other hand, the advantages of formative evaluation tend to diminish the longer one waits to begin. Prototype features tend to become entrenched with the passage of time, and as computer code grows more complex it is less easily changed. The danger for development teams is that so much work will be done before the all-important feedback can be gathered from visitors.

## Implications for the Future

Seasons is the first of three exhibits making use of scientific visualizations being developed by Augsburg College and the Science Museum of Minnesota. Our experience with Seasons has caused us to alter both our formative evaluation practices as well as our exhibit development practices. Below are some ideas of changes we will make.

## Changes in Formative Evaluation Practices

**The more iterations of formative evaluation we have time and resources to do, the better**. One measure of successful evaluations is the quality of the conversations it engenders in the development team. Good evaluation work should generate more questions. For example, we noticed that visitors scanned quickly through the set of buttons that plotted different kinds of data on the moving globe: temperature, amount of daylight hours, Sun energy on the Earth, etc. Visitors would click through this set of buttons, but would rarely point at the globe as it changed, or stop at it. This observation generated a lively discussion among the project team. We are about to study that more closely, using think-aloud interviews with users. Good evaluation engenders more evaluation.

**Evaluation cycles need to be quick**. This is actually a corollary of the above idea. This first phase of formative evaluation has taken 2 months. In the meantime, the development teams thinking has progressed, and threaten to overrun the evaluation process. It would be much more valuable if we could do evaluations more quickly.

**Use different evaluation methods to answer different questions**. While this is obvious, it takes some discipline to think creatively about each question asked. Once the bugs have been worked out of a protocol, it is hard to toss it aside. For example, the observation tools we used were changing throughout this first phase. They are finally in good shape now. However, it looks like think-aloud interviews would be the best way to answer new questions that need more in-depth exploration. We hate to start over developing a new tool, but being flexible and creative is the trick in formative evaluation.

Using new methods includes taking advantage of technologies. For example, we are considering using the computer to record mouse movements and screen selections, which would give us a wealth of data. This could supplement the rich data we are getting from observing peoples social interactions, facial expressions, and physical actions.

## Changes in Exhibit Development Practices

**Develop the exhibits using a rapid prototyping model**. For the next two exhibits we are going to try to develop a working model as soon as possible. This means not stopping for any formative evaluation, and also not spending lots of time as a development team trying to guess how it should be. For example, we will not stop to discuss screen design and navigation issues before building the model; those things will all change as a result of observation of and feedback from visitors. The lead developer and programmer will hammer out a model, and then the rest of the team, including the visitors, will get involved.

**Be willing to start over**. As noted in the section above, the advantages of formative evaluation tend to diminish the longer one waits to begin. If you start by writing lots of code before you have the first visitor feedback, you risk wasting valuable time. Such a process can be successful only if the development team is willing to toss that code in the trash and start over. As we work through the next two exhibits, we will have to be prepared to start over if necessary.

**Do more and quicker formative evaluation cycles**. The evaluation process must keep pace with the development process. For the next two exhibits we will plan and execute evaluation more quickly, and do evaluation more often.

**Acknowledgments** The authors wish to thank Drs. David E. Venne and William H. Jasperson of the Center for Atmospheric and Space Sciences, Augsburg College, for many helpful discussions. This paper was prepared under Grant ESI-9353480 from the National Science Foundation.