Comparing Short and Long-term Term Learning Effects between Stereoscopic and 2-Dimensional Film Showings at a Planetarium

C. A. Price, H.-S. Lee, E. Kasal, M. SubbaRao, J. Aguilera

Introduction

Over 90 million guests visit science museums, planetariums, and other science centers in the world annually (Association of Science-Technology Centers, 2013). Learning opportunities provided at informal, free-choice learning environments in these science centers account for nearly half of the average American adult's understanding of science (Falk, Storksdiek, & Dierking, 2007). Films are an essential part of the science center experience (Association of Science-Technology Centers, 2008), which have used them to educate the public for decades (Fraser, Heimlich, Jacobsen, Yocco, Sickler, et al., 2012). Recently, rapid technological advancement has changed how the films are presented to audiences, mainly due to the digital revolution and a rapid decline in the cost of visualization hardware (Lantz, 2011; Wyatt, 2005). This has helped science centers to respond to changes in society calling for more technology and interactivity (Blooloop, 2014; Jacobson, Wisne, MacGillivray & West, 2014).

Stereoscopy (a.k.a. "3D") is the technique used to create a physiological sense of depth in an image. It has been used in science education for over a century and a half (Wheatstone, 1852; Holmes, 1861) and first began appearing in a major fashion in museums starting in the 1970s. While the use of stereoscopy is not new to the general public (Parker, 1983), stereoscopy's popularity tends to ebb and flow over time (Dukes & Bruton, 2008; Gurevitch & Ross, 2013), but right now is at a peak due to advances in affordable and commercial uses of the technology. Partially because of this, audiences are becoming familiar with stereoscopic film (Ji & Lee, 2014) and the novelty effect is wearing off (Mitchell, 2012), affecting its educational potential through diminished motivation and interest (Price, Lee & Malatesta, 2014).

Today's rising generation knows only of a world where stereoscopic movies are commonplace. Yet, there is little empirical research about how stereoscopic film (Janicke & Ellis, 2013) affects science learning. We have found that currently available research are based mostly on self-reported responses from the audience from a handful of program evaluations. With so many people getting their science education through science centers outside of formal education, it is important to investigate how stereoscopy can impact science learning of the general public. In this study, we use randomized field trials of stereoscopic film showings alongside conventional two-dimensional showings to investigate the short-term and long-term-term learning effects associated with spatially demanding astronomical concept. Our research questions are:

(1) How does stereoscopy influence short and long-term term learning by adults who view a short film related to galaxy formation, shape, and modeling in a planetarium?
(2) How are these learning effects related to individual differences in demographic background and spatial ability?

To answer these questions we designed a field trial involving randomly assigned treatment and control conditions with pre-test, post-test and delayed post-test measures. First, we developed a film about galaxy morphology, using best practices from the spatial and multimedia cognition literature. We created two versions with identical content – the only difference is one was shown stereoscopically and the other was shown in traditional, 2D format. Then we showed the films to adult guests at a major urban science center in the United States. Participants' test performances were analyzed using repeated measures ANCOVAs with demographic and spatial ability covariates in order to control for individual differences that might be present between the treatment group who watched the stereoscopic galaxy film and the control group who watched the 2-dimensional galaxy film (. Findings were interpreted through the lenses of cognitive load and the Limited Capacity Model of Mediated Message Processing.

Implications are presented for science visualizers and researchers.

We begin with a literature review of stereoscopy in film and informal education settings. Then we describe the stereoscopic design principles that were manifested in creating the galaxy film. We then describe our research process, instruments, and analysis procedures. Next, we summarize results and interpret them through our theoretical framework. Implications are drawn for the use of stereoscopy in educational films along-term with new research directions and the advantages of using mobile technology in an informal research setting. Finally, we describe limitations of this study and present a conclusion.

Literature Review

Most visualizations of three dimensional objects in science education are displayed using flat, "2D" technology on televisions, computer displays and movie screens. These visualizations are limited in how they present a sense of spatial depth, sometimes leading to misconceptions and confusion in learners (Hubona, Wheeler, Shirah & Brandt, 1999; Barab, Hay & Duffy, 2000; Hansen, Barnett, MaKinster & Keating, 2004). For example, when determining a route from North America to Asia, a standard Mercator Projection map would suggest a straight line across the Pacific would be the shortest route, when a three dimensional globe would make it more apparent that the shortest route is across the North Pole. Stereoscopy can address this limitation by providing a physiologically based sense of depth. While a wide range of stereoscopic technologies exist, but they all work in the same basic way – by supplying a slightly different version of the same image to each eye. Objects within each image are laterally separated by a horizontal distance related to their distance to the viewer (closer objects are displaced further than distant objects). This creates a binocular disparity (a.k.a. parallax) that mirrors how our vision system perceives depth in real life. This provides stereoscopic visualizations with the ability to provide a sense of depth that is much more authentic than what can be inferred from flat visualizations.

Stereoscopy and Film

Stereoscopy has been a visualization technique used in photography and film as far back as 1838 (Zone, 2014). Almost a decade after Wheatstone first suggested using stereoscopy in education (Wheatstone, 1852), Oliver Wendell Holmes referred to its sense of immersion as "…hypnotism… a dream-like exaltation of the faculties." He added: "It is not a toy… it is a divine gift nominally placed in our hands by science," (Holmes, 1861, p. 31). Advance in time about a century and a half and optimism has given way to pessimism. Pulitzer-prize winning film critic Roger Ebert famously said, "3D will never work," not least of which because he felt the audience would become mentally tired from watching lengthy 3D movies (Ebert, 2011).

The viewer's perception of a stereoscopic film is significantly different than that of a traditional, 2D film in many ways. Most obviously, the volume of space the film is "filling" is much higher since the space is front of and/or behind the physical screen is now part of the scene. Depending on the technology, the film may appear to jump out to the viewer or recede back into the screen. This volume is highly dynamic throughout a film (Sun & Holliman, 2009) causing a viewer to mentally treat it as a round canvas with indistinct edges, as opposed to a traditional planar view perpendicular to themselves (Liu, 2013). This is one of the reasons stereoscopy is often considered an immersive technology, that potentially can be used to enhance situated learning (Dede, 2009).

There is little empirical research about how stereoscopic films impact learning. Our search of ERIC and Google Scholar using a combination of keywords taken from these three groups: 1. "stereoscopy", "3D", "stereo" 2. "education", "science education" 3. "film", "movies" and "video", revealed none. However, we found significant literature on *affective* impact of stereoscopy used in film (Rooney & Hennessy, 2013), TV (Janicke & Ellis, 2013) and video games (Schild, LaViola & Masuch, 2012). In each case, the direction of its effect on participants' affective responses such as emotional arousal, engagement, interest and persuasion was mixed partly due to factors beyond the presentation format difference, such as how it was implemented and what subject matter the film was showing.

Science museums and other informal science centers have long used stereoscopic films as an educational tool (Parker, 1983). Most recently, it is used in presentations with IMAX (Counts, 2009), digital planetarium (Fluke and Bourke, 2005; Lantz, 2011) and digital flat screen displays (Johnson, Leigh, Morin & Van Keken, 2006). There have been calls by industry leaders to use stereoscopy as part of a shift towards more digital, immersive exhibitions that are more mobile and less resource intensive (Jacobson, Wisne, MacGillivray & West, 2014) along with calls for more research into their effectiveness (Steinbach, 2011). In a review of the literature related to films shown on giant screens instead of relatively small computer screens for example, Fraser, et al. (2012) only found results from-self reported, non-experimental audience surveys (evaluations) or interviews of focus groups, which led to their calling for more controlled research designs to identify added benefits of stereoscopic films over conventional two-dimensional films. In one such evaluation based on self-reported surveys, the audience claimed to learn more when watching a 2D film while reporting higher enjoyment of a stereoscopic film (Apley, et al. 2008). Outside of usage in museums and science centers such as large screen films, stereoscopy has been often used in the training that involves virtual simulations (Nash, Edwards, Thompson & Barfield, 2000), for pilots (Menendez & Bernard, 2000) and medical doctors (van Beurden, IJsselsteijn, & Juola, 2012). Studying stereoscopy in informal settings is challenging due to the nature of the setting where the learning tasks place. Audiences are frequently under time restrictions and have considerable external distractions. As films are often an important revenue for most science centers, it is challenging to adjust a film showing schedule to accommodate experimental study designs with comprehensive assessments.

Galaxy Film: Goals, Theoretical Basis and Stereoscopic FilmDesign Principles

The Galaxy Film Content and Learning Goals

We developed a film for a planetarium entitled *What Does the Milky Way Look Like?* (hereafter: galaxy film) addresses the shape of the Milky Way galaxy and the challenges with determining that shape when looking at the Milky Way galaxy from within[1]. It was produced in high definition and consists of a combination of photography and computer graphics/simulations with narration. The learning goals of the film were to teach the shape of the Milky Way and how it is related to the distribution of stars as seen in the night sky. It includes a number of challenging spatial concepts such as imagining the shape of an object when looking at it from within and how to distinguish the shapes of different semi-transparent objects that lie within the same line-of-sight. The film is 7 minutes 50 seconds in length.

The main learning goal of the film was to depict how astronomers go from two dimensional observations to constructing 3D models, using the structure of the Milky Way as a case study. The film begins with a view of the Milky Way as seen on a dark, clear night. It then gives examples of how ancient cultures interpreted its shape differently. It then covers various models for the shape of the Milky Way as proposed by historic astronomers. Some drawings made by those astronomers are superimposed over the night sky, using movement and stereoscopy to show how a seemingly flat sky can be interpreted in three dimensions. The next part of the film discusses how the discovery of other galaxies provided outside perspectives of how galaxies appear, giving astronomers more insight into possible shapes of our own. Then the film shows how not only perspective, but dust also makes it difficult to see our entire galaxy while introducing some tools astronomers use to see through dust, such as infrared imaging. Finally, links the current shape of the galaxy to its history and how it will change in the future as the Milky Way interacts with other, nearby galaxies.

Theoretical Basis: Limited Capacity Model of Mediated Message Processing

The Limited Capacity Model of Mediated Message Processing (LC4MP) is a psychological theory that describes how limits on the human cognitive ability affect the ability to process and recall messages. LC4MP is commonly used framework for media studies in education (Harris & Sanborn, 2013; Fisch, 2014). It describes a process whereby cognition, motivation, media, dynamic human behavior and communication all impact how a message is received (Lang, 2000). The process incorporates three phases. The first is *encoding*. Lang (2006) described it as "…making a mental representation of a stimulus by encoding important aspects of a message". In stereoscopy, the added sense of depth elevates the importance of spatial properties in a visual message. The second stage is *storage*, which is placement of the message in short-term memory and is strongly affected by motivation (Lang, Sanders-Jackson, Wang & Rubenking, 2013). In this regard, stereoscopy has been consistently shown to affect emotional and engagement aspects of viewers (Bokander, 1966; Koh, Tan, Tan, Fang, Fong, et al., 2010). The final stage is *retrieval*, which is based on free call (without cues) of the message. Stereoscopic films have been shown to enhance recall of objects that are prominently displayed while harming recall of objects that are not (Terlutter, Diehl, Koinig & Waiguny, 2013), suggesting that the encoding process is critical for deciding which aspects of a message are recalled or not. Overall, for a message to be received it has to be received into short-term memory, stored in long-term-term memory and retrieved – all within an environment of limited cognitive resources. According to this theory, stereoscopy could have its biggest impact at the encoding stage by increasing the importance placed on the spatial properties of the message – thereby increasing short and long-term learning.

[1] It can be viewed in 2D and stereo, in English and in Spanish, at https://www.youtube.com/user/AdlerSVL and is available for use under a Creative Commons Attribution Share Alike 3.0 License.

Stereoscopic Film Design Principles

Prior studies have shown that stereoscopy can increase cognitive load (Kooi and Toet 2004; Okuyama 1999). To help mitigate that, we assembled four *Stereoscopic Film Design Principles* (SFDPs) based on a review of literature. These SFDPs are theoretically grounded in the spatial cognition and cognitive load theory literature and were interpreted by the creators of the film. The following is a summary of the SFDPs along-term with examples of how they were implemented in the galaxy film. All principles could affect audience interpretation of stereoscopic films. The last two principles also apply to two-dimensional films, but we think have added importance in a stereoscopic film.

Stereoscopic Film Design Principle 1: Gradually Adjust Viewers to the Presentation Medium

It takes time for a viewer to become used to watching a stereoscopic film. Yet, many stereoscopic films, especially those created with an entertainment element, begin with a striking stereoscopic shot meant to shock the viewer and grab their attention. It usually includes an object that appears to “float” in space very close to the viewer. This can create extraneous cognitive load and physical discomfort while also taking the viewer out of the narrative of the film (Miller & Beaton, 1991). Studies have suggested that viewers of stereoscopic films first mentally translate the stereo images to 2D, because they are simply used to watching 2D films (Miller, et al., 1991; Surdick, Davis, King & Hodges,1994; Price & Lee, 2010; Ting, Tan, West, Squelch & Foster, 2011). It takes time for their minds to adapt. Kosslyn, et al. (1994) said, “one actively anticipates what one will see when one makes a [mental image] movement (page number).” Prior experience with the viewing technology plays an important role in formulating viewers' expectation on what comes next. To address this, our first principle calls for easing the viewer into stereoscopic imagery. In our film, this principle was implemented in three ways. First, the opening credits involve a mostly flat and static shot with only the film’s title in stereo – and even then the title is not occluded with any object; occlusion can cause visual fatigue (Devernay & Beardsley, 2011). Second, it involves pictures of stars and hills – something the viewer will most likely have seen in real life (as opposed to starting with a computer generated effect or a photograph of something they have no real life experience with). Finally, first fully stereoscopic shot does not occur until after 40 seconds of 2D imagery of mountains and the night sky. We felt this eased the viewer’s transition from the expectations of watching a 2D film to the reality of watching a stereoscopic film.

Stereoscopic Film Design Principle 2: Stereo Discipline

Not everything needs to be in stereo. Stereoscopy should be used only when the added sense of depth will help the viewer understand what they are viewing (Vandeland, 2013). This is usually with spatially complex items (Hansen, et al., 2004; Price, et al., 2014). Tasks requiring integration of spatial dimensions benefit mostly from stereoscopic views, whereas tasks requiring focused attention on one or two dimensions benefit from 2D views (Wickens, Merwin & Lin, 1994). This principle not only applies to content, but also technical implementation of stereoscopy. When stereo is used, the disparity should be as limited as needed to present the sense of depth required. Benefits of stereoscopy plateau at an intraocular distance assumption of 3cm (Rosenberg, 1993), if you go any further you risk creating mental discomfort in the viewer or causing visually induced motion sickness (VIMS) (Takada, Murakami, Matsuura, Takada, Iwase & Miyao, 2011). An example of how we applied this principle is in a scene where we showed how past scientists noted the distribution of stars across the sky. The scene is based on an ancient drawing of the stars around the Earth. We began the scene in 2D, showing a photograph of the drawing. Then we transitioned to a digital model of the drawing, but that too was in 2D. Only then did we transition yet again to a stereoscopic view of the model before virtually flying through it. This multistep transition sequence took forty seconds. In general, when deciding whether to use stereo, default to 2D unless it is expressly required. And when it is required, use as little stereo disparity as needed.

Stereoscopic Film Design Principle 3: Utilize Pictorial Depth Cues

Pictorial depth cues may help lessen the mental burden of processing the physiological stereoscopic depth cue. Pictorial depth cues (ex: shadows on objects – see Mehrabi, Peek, Wuensche & Lutteroth, 2013 for a categorization of common depth cues) are built into 2D representations to supply a sense of depth. Whether the use of multiple depth cues at the same time is additive or not is likely influenced by the environment and learning task at hand (Reinhardt, 1990). However, research evidence is strong enough to suggest that the use of a few pictorial depth cues, within specific implementation guidelines, can help support the physiological depth cue of stereoscopy (Reichelt, Häussler, Fütterer, Leister, 2010). Effective depth cues should have a relatively minor presence in images (Surdick, et al., 1997) and they should not be used to convey other information (Miller, et al.,1991) and they need to be consistent (Devernay, et al. 2011). Three types of cues may be especially useful with stereoscopy. First, motion can be used to imply depth (Wickens, 1989; Sollenberger & Milgram, 1993; Ware & Franck, 1996). For example, partial motion along-term the 3D line of sight can be used to imply a sense of depth (Wallach and O'Connell, 1953). Second, shadow can be used for positioning, orientation and depth (Wanger, Ferwerda & Greenberg,1992; Tory, Kirkpatrick , Atkins & Moller, 2006), but only one fixed light source should be used (Pani, Jeffres, Shippey & Schwartz, 1996; Hubona, 1999). Finally, a frame of reference can be used to convey depth. A picture of the Moon by itself may seem flat compared to a picture of the Moon rising above a mountainous horizon and/or behind clouds. Visual tools such as mesh frames (Lieu & Wickens, 1992), and drop lines (Brooks, 1992; Barfield, 1995) can be used, especially when linked to a ground plane (Surdick, Davis, King, Corso, Shapiro, et al., 1994) and when exocentric (McCormick & Wickens, 1995). One example of a pictorial depth cue in our film is the use of a virtual camera moving around a simulation of the Milky Way galaxy – using motion to increase the sense of depth.

Stereoscopic Film Design Principle 4: Minimize Other Sources of Cognitive Load

Most viewers of stereoscopic films are also in the presence of other distractions that create cognitive load. This is especially important in areas of informal learning, such as museums, where other activities may be going on in the same viewing space as the film. Also, immersive environments usually have other cognitive demands other than the visual display. Thus, one needs to diminish unnecessary cognitive load created by the visualization. One way to do so would be to adhere to the Cognitive Theory of Multimedia Learning (CTML) (Mayer, 2005; Mayer, 2008). Based on cognitive load theory, CTML describes a series of guiding principles that should be used to minimize simultaneous demands on the same cognitive channels. For example, the Temporal Contiguity Principle states that when narration is used, it should be presented at the same time as the visuals to which it is linked (Moreno & Mayer, 1999). Also, the Specific Redundancy Principle suggests that written text should not be used if it duplicates spoken text with pictures (Van Merriënboer & Sweller, 2010). And there is the Modality Principle, which suggests that spoken rather than written text should be used alongside animation (Low & Sweller, 2005). Our film adheres to each of those principles. For example, in adherence to the Modality Principle, we only audio narration in the film. Text is only used for the title scene and closing credits.

Methods

Research Setting

The research took place at the Adler Planetarium located in Chicago, Illinois, USA. The planetarium is a popular cultural attraction in the region with about 500,000 visitors per year. Specializing in visualizations, the planetarium has four public theaters. One of the theaters is housed within the Space Visualization Laboratory (SVL - Figure 1) – a publicly accessible space where visualization experts develop and test new films while interacting with guests. This study took place in the SVL, using its 90" screen and polarized stereoscopic projectors. The space was chosen to preserve validity of the study. As a public part of the planetarium's floor space that has daily presentations, it has much of the same distractions and characteristics as other museum theaters. Throughout the day, staff and signage around the planetarium recruited guests to participate in the study. Originally, recruitment proved difficult but became easier once research staff became more comfortable approaching guests. After the first few weeks, most guests who declined participation only did so due to technical reasons (other time commitments and non-English speakers being the main reasons). As of six months after data collection ended, less than half of the gift cards offered as incentives were redeemed, suggesting that participation was driven by interest as well as incentives.

Visitors were not informed about the film presentation format during recruitment. Adults were targeted, but families with children were also recruited with the understanding that the children would not be used for data collection (they watched the films alongside the parents or played with other exhibits in the space while parents took the tests). Data were collected mostly on weekdays from mid June to mid August. On average, there were between four and five participants per film session.

A session began as participants entered the SVL. They sat down in chairs in front of the screen and were given iPads (Figure 2) running a software app customized for the project.[2] The iPad was chosen as the platform to minimize the time spent on data collection and to retain guest interest. In our experience, those are two of the largest barriers to guest recruitment. It is also closely aligned with the rest of the SVL space, making the research experience part of the guest experience. In fact, the independent evaluation of the project found that participants were more interested in learning about the research process than in the film content (Borland, 2014). The app began with a screen that asked them for basic demographic information. After they filled out the information, the app displayed a pause screen. Once all participants had answered the questions, the research assistant told everyone to continue past the pause screen and begin the pre-test. When the pre-test was completed, a second pause screen was displayed (this one associated with a five minute pre-programmed delay to eliminate the temptation to immediately take the post-test). The film was begun once everyone had completed the pre-test. The presentation format of the screening (2D or stereo) was randomized at the daily level at the start of the project. Because of technical limitations, it was not possible to change the film or format between sessions on the same day. When the film was completed, the participants were told to move beyond the pause screen and take the post-test. Total session time, from entering the demographic information to finishing the post-test, took an average of 20 minutes. Each adult was compensated with a free film ticket or a gift card to an on-site store. About six months later, they were sent an e-mail inviting them to take a delayed post-test that had been placed online (except for five participants who did not provide an e-mail address when taking the pre-test). Those who took the delayed post-test were compensated with an Amazon.com gift card.

Overall, 498 adults took the pre- and post-tests and 123 also took the delayed post-test. Their ages ranged 18 - 82 years (mean = 35). They self reported as 52% female and 48% male. About 8% claimed to have difficulty viewing stereoscopic content and 92% reported no known prior difficulties. About 6% reported a high level of knowledge of astronomy, 39% reported a medium level and 55% reported a low level.

Instruments

The pre--test consisted of three sections: a background section, a spatial cognition section, and a content knowledge section (hereafter: knowledge section). The post-test consisted of the same knowledge sections and a drawing task. The background section asked for age, gender, whether they had difficulty looking at 3D images and their overall knowledge about astronomy. The spatial cognition section contained five items from the Purdue Spatial Visualization-Rotation (PSVT) Test (Bodner & Guay, 1997) to distinguish participants according to their spatial abilities. The PSVT measures mental visualization ability, which is related to the learning goals of the film and has been often used with other stereoscopic studies. The five items in the spatial cognition section were chosen to represent a broad range of difficulty after piloting the assessment with 45 adult visitors prior to the study. The knowledge section of the test included five items, four of which addressed the content of the films related to the shape, composition, appearance, and evolution of the Milky Way Galaxy (Table 1). The other item was a confidence rating item related to identifying the shape of the Milky Way galaxy from a photograph. These items were in the multiple choice format, with the options condensed from responses to open-ended versions of the questions included in the same pilot study where we tested the PSVT items. The first item was formatted as a multiple choice image selection. The second item measured their confidence in their prior answer. The third, fourth and fifth items were all multiple choice items with three, four and three choices respectively. Each item was matched with one of the four learning goals of the film, except for the confidence item. For the last section of the test, the guest used their fingers and the on-screen touch keyboard to draw and label an object. A stylus was available for use upon request. The software recorded their answers and how long-term they spent on each item, in seconds. Visitors spent an average of 346 seconds (SD = 110 seconds) on the pre-test and 130 seconds (SD = 55 seconds) on the post-test. Accelerometer data was also recorded to look for differences in how they held the devices. No differences were found between stereo and 2D groups and thus the accelerometer data are not used in this paper. The delayed post-test included the knowledge section plus two additional items asking the guest to rate their enjoyment of the film and whether they recall seeing it in 2D or in stereo. Data were stored in a MySQL database and then analyzed using PASW Statistics 18 (a.k.a. SPSS).

[2] The software is available at https://github.com/aaronp808/Two-Eyes--3D under a GNU Affero General Public License.

Scoring

Answers to the PSVT items were coded as correct (score 1) or incorrect (score 0). Each visitor was assigned a single spatial score based on the number of correct answers, ranging from 0 to 5. Answers to the knowledge questions were also coded as correct (score 1) or incorrect (score 0). For each of the two tests, a score of the total number of correct knowledge items was assigned (range of 0-4), omitting the confidence item score.

The drawings were coded by two pairs of independent researchers using a common rubric (Table 2). The rubric was based on the presence of four structural features of the Milky Way galaxy discussed in the film: the nucleus of the galaxy, the central bar of stars across the nucleus, the spiral shape of the galaxy and the Sun's location in the galaxy. For the first three features, the coders simply looked for the presence or lack thereof of the property. For the last feature, if the Sun was present the coders also analyzed at its location as correct or not. After the first pair of coders recorded their scores, the drawings for which they did not agree then were evaluated by a second pair of researchers using the same rubric. The drawings that still did not have an agreement from this second pair had the scores for both compared from all four coders. If there was a majority agreement, that code was adopted. The second pair of researchers then negotiated a final code for the few drawings without majority agreement. A final drawing score for each test was created by adding up the scores on all the rubrics of the test (range of 0-5). IRR ranged from $\alpha = .64$ to $\alpha = .94$ (Table 3).

Analysis

The first stage of our analysis was a comparison of the background characteristics between the 2D and stereoscopic presentation groups for the Galaxy film. We used chi square tests for categorical demographic variables such as gender and difficulties with stereoscopy and independent samples t- tests for continuous demographic variables such as spatial cognition ability scores and age to test whether the two groups were statistically similar. Next, we conducted non-parametric Wilcoxon signed ranks tests on individual knowledge items to investigate performance differences between the pre- and post-test scores. We created total individual performance scores on pre- and post-tests by summing up the four knowledge items (excluding the confidence rating item score). We applied repeated measures ANCOVAs on the total pre-test and post-test performance scores to look for differences linked to the presentation format (2D vs. stereoscopic). Covariates were also entered to control for age, gender, difficulty with stereoscopic visualization, prior knowledge and spatial abilities in order to examine the differences in performance changes from pre to posttests (short-term learning) between the two groups. Third, we compared differences in the drawing scores between the presentation format using ANCOVAs with age, gender, difficulty with stereoscopic visualizations, prior knowledge, and spatial abilities as covariates. Finally, we applied repeated-measures ANCOVAs to compare total performance scores on the post-test and delayed post-test for long-term learning. We did not compare pre-post-delayed post test scores due to the fact that there was a significant reduction in the total participants who completed all three tests (n=123) from those who completed pre- and post-tests (n=498).

Background characteristics of the two groups were similar. Chi-square tests indicate no significant differences between the 2D and stereo groups in terms of gender (48% male in the 2D group vs. 50% male in the stereo group, $p = .77$), difficulty with stereoscopic visualizations (7% having difficulty with stereo in the 2D group vs. 10% having difficulty with stereo in the stereo group, $p = .18$), and prior knowledge (53% no knowledge vs. 50% some knowledge vs. 7% high knowledge in the 2D group as compared to 52% no knowledge vs. 42% some knowledge vs. 6% high knowledge in the stereo group, $p = .86$). Also, ANOVAs indicate no significant difference between spatial cognition scores (2D M = 2.98 vs. stereo M = 3.16, $t(497) = 1.43$, $p = .15$. However, the average age of the 2D group was slightly older (M = 36.67, SD = 15.58) than that of the stereoscopic group (M = 33.68, SD = 14.92), $t(497) = 4.75$, $p < .05$.

Performance Change on Individual Knowledge Items between Pre- and Post-Test

Table 4 shows percentages of participants providing correct answers at the pretest and the posttest in the 2D and the stereo groups. According to Wilcoxon analysis results, both 2D and stereo presentation groups showed significant improvement across all four knowledge items and they showed increase confidence in their answers. On the pre-test, the order of item difficulties measured by the percentage of correct answers from most difficult to easiest was: Galaxy-Shape, Galaxy-Engulf, Galaxy-Spheres and Galaxy-Stars. On the post-test, the item difficulties were in the order of Galaxy-Engulf, Galaxy-Spheres, Galaxy-Shape and Galaxy-Stars. This order was the same in both the 2D and stereo groups. The most improved item was Galaxy-Shape, which changed from the most difficult pre-test item to the second easiest post-test item, showed a 41% increase in the 2D group and 35% increase in the stereo group. The average increases of all other items was 16% for the 2D group and 18% for the stereo group. In addition, the confidence item associated with photo identification of the Milky Way Galaxy significantly increased in both groups. For the 2D group, participants choosing very confident increased from 28% to 75% while those choosing very confident increased from 36% to 75% for the stereo group.

In terms of completion time on the knowledge items, the 2D group took an average of 112 seconds with a standard deviation of 39 seconds in the pretest while the stereo group took an average of 103 seconds with a standard deviation of 36 seconds. This difference was statistically significant, $t(496) = 2.54$, $p < .05$. For the posttest on the knowledge items, the 2D group spent an average of 64 seconds (*SD* = 42 seconds) and the stereo group took 58 seconds (*SD* = 25seconds. This difference was not statistically significantly different, $t(496) = 1.80$, $p = .07$.For the drawing task, there was no significant difference between the two groups in terms of completion time, 2D M = 68 seconds (SD = 45 seconds) vs. stereo M = 70 seconds (SD = 46 seconds), t(496) = .64, p = .53.

Short-term Learning Gains

Repeated measures ANCOVAs were applied to examine whether significant learning differences existed between 2D and stereoscopic presentation types while controlling for age, gender, difficulty with stereoscopic visualizations and spatial ability. Participants' self-reported age information was used to generate a categorical variable: (age 18-24 – assigned as 0), medium (age 25-44 – assigned as 1) and older (age 45+ - assigned as 2). A dichotomous categorical variable for gender was created with males assigned a 0 and females assigned a 1. Similarly, a difficulty with stereoscopic visualizations variable was created and assigned a 0 for no and a 1 for yes. An astronomy knowledge categorical variable was created with three categories and values: low (0), medium (1) and high (2). Finally, spatial ability was converted into a dichotomous variable: those who got 0-2 answers correct as a low spatial group and assigned a 0 and those who got 3-5 answers correct as a high spatial group and assigned a 1. The independent variable for the ANCOVAs was presentation type (2D and stereo) and the dependent variable was the number of correct items on the knowledge section of the test, excluding the confidence item.

Table 5 shows descriptive statistics on pre/post test performances by groups defined by the presentation type variable as well as covariates. Table 6 lists the significance of various effects obtained from the repeated measures ANCOVA The increase in performance on the post-test from the pre-test persisted across age, gender and difficulty with stereoscopic visualizations (Main time effect in Table 6). Table 5 also shows effect size measured as Cohen's d, i.e. the pre/post test performance difference in pooled SD units. The overall effect size was .92, which was a large impact.. This performance improvement was found in both 2D and stereo groups. No significant interaction effect between time (pre-test and post-test) and presentation type indicate that both groups improved similarly from pre-test to post-test. In addition, ANCOVA results indicate significant differences by prior knowledge and spatial abilities. According to Table 5, performances of participants with high prior knowledge were generally higher than those of participants with low prior knowledge. Similarly, performances of participants with high spatial abilities were generally higher than those of participants with low spatial abilities. There was an interaction effect between time and spatial ability and age. That is, those with higher spatial ability and higher age showed greater increase between the pre- and post-tests. There were no other significant interaction effects related to gender, difficulty, and prior knowledge, indicating that all groups defined by these characteristics benefited similarly from the galaxy film.

Long-term term learning: From Post-Test to Delayed Post-Test

To see if the change detected in the post-test persisted over time, the post-test and delayed post-test scores were compared using descriptive statistics and a similar repeated measures ANCOVA. There was a drop in scores among both groups on the delayed post-test (Table 7). However, the drop was greater for the 2D group than the stereo group as confirmed by the significant time X presentation type effect in Table 8. This difference remained significant even after controlling for background variables, $F(1,118)=4.01$, $p < .05$. As before, prior knowledge and spatial ability were also significantly related to the overall knowledge scores, but there were no significant interaction effects with time.

Guests reported more enjoyment out of the stereo film than the 2D film. On a scale of 0-5, the mean score of the 2D film was 4.0 (SD = 1.04) and the mean score of the stereo film was 4.39 (SD = 1.02).

Though the stereo group's average enjoyment score was higher than the 2D group's score, this difference was not statistically significant according to an independent samples t-test, *t(1020= 1.94, p = .055.* When asked to remember what type of presentation they saw, 11 of the 48 guests (23%) who saw it in 2D incorrectly remembered it as being in stereo while only four of the 56 guests (7%) who saw it in stereo incorrectly remembered it as being in 2D. A chi-square test found this difference significant, $\mathcal{X}$= 86. 7, *p* < .001. On the delayed post-test, confidence identifying the shape of the Milky Way galaxy reported by both groups rose back to their pre-test level. The 2D group averaged a mean confidence score of 1.23 (SD = .66) and the stereo group averaged a 1.29 (SD = .63). The difference between the post-test and delayed post-test scores between the two groups was not statistically significant according to a repeated-measures ANCOVA.

Drawing Task

There was no difference on drawing task scores between those who watched the film in 2D or in stereo (Table 9 for descriptive statistics and Table 10 for within and between subjects effects). The mean score on the rubric of the 2D group was 2.53 (*SD* = 1.44) (on a scale of 0-5) and the stereoscopic group's mean was 2.58 (*SD* = 1.47). The scores remained virtually identical even when controlling for age, gender, difficulty with 3D, prior astronomy knowledge and spatial ability (Table 9). Age, prior astronomy knowledge and spatial ability were all significantly related to drawing task scores.

Discussion

Despite the ubiquitous presence of stereoscopic films in science centers, the learning benefits associated with stereoscopic films on the general audience are understudied. In this study, we overcame shortcomings of studying learning in informal settings by (1) implementing a quasi-experimental research design involving treatment (stereoscopic film presentation) and control (2D film presentation) groups, (2) using iPads to collect learning data, (3) collecting demographic information to control for learning differences related to pre-existing differences, (4) studying short and long term learning effects, and (5) using various measures of learning such as multiple choice knowledge items and open-ended drawing tasks.

Guests performed better on the knowledge section under both conditions for both the initial pre-test and post-test, suggesting that the film achieved its educational goal. However, we also observed that most guests lost what they learned from the film six months later. With that established, we can turn to our core research question: how was short and long-term term learning affected by the presentation type of the film? There was no difference in changes in performance between the two groups from pre-test to post-test. Even the relative order of item performance was the same between the two groups in the pre-test as well as in the post-test. While gender and prior knowledge of astronomy showed no effect on learning gains, spatial ability and age were both positively related. We also utilized a drawing task, which requires generally more complex thought than a simple multiple-choice test (Barrazza, 1999; MacPhail and Kinchin, 2004) and can overcome issues such as linguistic barriers to test performance (Navarrete & Gutske, 1996) while allowing for more rich analysis (Maneja-Zaragoza, Linde & Juncà, 2013). We analyzed our drawing results using four rubrics related to several spatial features of the Milky Way galaxy to look for a wide variety of differences. However, we found no differences between the two groups at the post-test. Combined with the lack of differences in the knowledge section of the test, the results are pretty consistent in that presentation type had no effect on short-term learning.

Our prior work in this same lab, using different tasks but with the same equipment and in the same environment (Price & Lee, 2010), also found no benefit of stereoscopy on performance accuracy. But it did increase the amount of time it took guests to perform the requested tasks. We interpreted this as being due to an increase in cognitive load, something other stereoscopic studies have also reported. In this study, we also found no increase accuracy. However, this time we found no difference in the amount of time it took them to complete the tasks. This might be related to the fact that in our prior study subjects were manipulating stereoscopic vs. 2D objects as part of performance tasks while in this current study subjects were answering multiple choice questions presented in the same 2D format. In addition, this could be a sign that the stereoscopic design principles were working and reducing cognitive load. In general, the stereo group needed significantly less time to complete the pretest and the posttest than the 2D group. Anecdotally most guests seemed to enjoy the experience and were inquisitive afterward – showing no signs of mental fatigue. We also found no difference between the groups completion time for the drawing task. So in terms of the LC4MP, we believe this is evidence that the environment was conducive to successful encoding and storage at this stage of the experiment.

However, the story with long-term term learning was different - the stereoscopic presentation group retained more of their learning gains than the 2D group. In fact, as seen through their test scores, the stereoscopic group retained almost all their learning gains (86% retained) versus the 2D group (28%). There was no interaction effect with age, gender, prior astronomical knowledge or spatial ability. The latter result is most interesting as spatial ability *was* related to the short-term learning gains. It suggests that spatial thinking was more in demand on the immediate test than on the delayed test, which is further evidence that the immediate test was using a different type of cognition than the delayed test.

At its core, the LC4MP model describes a multi-stage process for learning and recall in an environment where cognitive resources are limited. Stereoscopy has the potential to enhance long-term term learning because it can emphasize spatially salient parts of a message, thereby increasing resources allocated to the first stage of learning in this model - encoding. It also affects the second stage, storage, by increasing motivation, something we see in our delayed post-test results showing guests who watched the film in stereo reported slightly elevated enjoyment out of the film compared to those who saw the 2D version. Ji, Tanca & Janicke (2013) found that audiences do not always report more enjoyment from stereoscopic films, however their films were not educational and they analyzed their results through the lens of LC4MP's impact on emotional arousal and emotional fatigue, which would not be as important in an educational film as opposed to a strictly narrative film (Ji, et al. 2014). Finally, our results show that recall, the final LC4MP stage, is better for the stereoscopic group – which would be the expected outcome if encoding and storage were more productive in that group. That we do not see a difference at the storage level (the first post-test) is not a surprise because that test was cued in that it took place immediately after the film and only about ten minutes after they finished the pre-test. On the other hand, the delayed post-test was without cues as the guests were recruited with no prior knowledge that a delayed post-test would be part of the study and it took place far removed from the museum environment. The length of time between the treatment and delayed post-test, six months,   is important. A priori, one may not expect much to be recalled half a year later after watching a 7-minute film shown in a museum (a distraction heavy environment). That is, the treatment dosage was very small compared to the length of time we waited to do the post-test. Nevertheless, our results are a testament to the resiliency of such short-term learning gains. We are not aware of other research that has studied long-term term learning associated with educational films in an informal setting. More research should be conducted on this, including delayed tests of staggered length, to see how the recall process switches from short-term to long-term term storage when dealing with educational media associated with advanced concepts and high cognitive demand.

Our findings are aligned with many other recent studies suggesting that stereoscopy does not provide an advantage to immediate (short-term) learning over 2D representations (Ting, et al., 2011; Joseph, 2011; Mukai, Yamagishi, Hirayama, Tsuruoka, Yamamoto, 2011; Price, et al., 2014;). However, this study shows that stereoscopy can have an impact on long-term term learning. The film was designed carefully to limit extraneous cognitive load, which may have allowed the stereoscopic effect to enhance the salience of the spatial components of the film. This could have implications for science visualizers and media producers who are looking to enhance long-term term learning of spatially challenging topics. Follow up research on the importance of limiting cognitive load in stereoscopic film is important. The tendency of the edutainment industry is to increase cognitive load to produce an affective result from the audience. While that is important, more research will show whether it is harming learning and what may be done to maintain both affective and learning gains in film. Finally, while the NRC has called for more longitudinal studies of informal media (Feder, et al., 2009), they also present the need for cumulative longitudinal studies that look at how repeated exposures to the same, or similar, media affect learning over time. A study of multiple stereoscopic film presentations over time would help measure its impact on those who watch science media repeatedly, such as on stereoscopic televisions at home.

Finally, the use of a mobile device (in this case the iPad) in the study helped us overcome many of the recruitment and assessment limitations inherent in informal learning spaces, which have traditionally used "older forms of analysis" (Feder, et al., 2009). The iPads helped keep the interest of the guests and create an experience that was more aligned with the high-tech and highly visual atmosphere of the Planetarium. Guests often got really into the assessments – especially the drawing task. Of course, the individual tasks could have been done on paper, however the iPad greatly simplified the process and limited the amount of time guests spent in the session. It is our belief, based on years of working with informal audiences, is that an assessment on a clipboard would have harmed the validity of the study by altering the mindset of the guests such that they felt like they were in a lab setting as opposed to watching a film on the museum floor. The freedom providing by the guests' interest in the procedure is what allowed us to use a RCT design. Aligning assessment strategies with their environments is critical for research in informal spaces.

Limitations

All studies in active, real-world settings like museums and planetariums are subject to limitations caused by the background characteristics of guest population of such science centers, the active physical environment, time limitations, etc. All together they limit the ability to generalize findings to formal educational settings. Also, it is possible that only those more interested in the subject responded to the delayed post-test, thus creating an audience of guests who have more prior knowledge and interest. However, we offered a considerable incentive for participation and achieved our response quota within 24 hours of sending the recruitment e-mail, which stated that the study would end after 125 responses. This suggests that the incentive was a strong reason for participation. The authors have considerable experience recruiting museum guests via e-mail to take online surveys and had never seen such a rapid response. Also, we found no difference between the reported prior astronomy knowledge between the pre-test and delayed post-test groups. Also, the knowledge section is only four items long-term. More items can improve the accuracy of capturing participant performances. The test was, of course, constrained in scope by the limited duration we could keep the guests. But it does cover a wide range of items, each aligned with one of the four conceptual learning goals of the film.

Conclusion

We tested the effects of stereoscopy on short and long-term term learning with an astronomical film designed to limit potential cognitive load associated with stereoscopy. Results show that stereoscopy had no effect on short-term learning but a substantial positive effect on long-term term learning. We believe the added salience that stereoscopy provides to spatial elements through its elevation of the sense of depth increases the chance of those elements to be persistently encoded and later recalled. However, that benefit does not appear in the short-term, perhaps due to other situational elements competing for the same cognitive resources on an immediate post-test. Stereoscopy may be a valuable tool for teaching about spatial elements in scientific visualizations, if the film is designed to limit extraneous cognitive load and the recall of those elements is done in an uncued environment.

Acknowledgements

The project was funded by National Science Foundation award DRL-1114645 and supported by the American Association of Variable Star Observers under the direction of Dr. Arne Henden. We thank Dr. Eric Chaisson, Dr. Janice Gobert, Dr. Maria Roussou, Dr. Holly A. Taylor and Ryan Wyatt for their advice on this project and manuscript. We also thank Clockwork Active Media Systems.

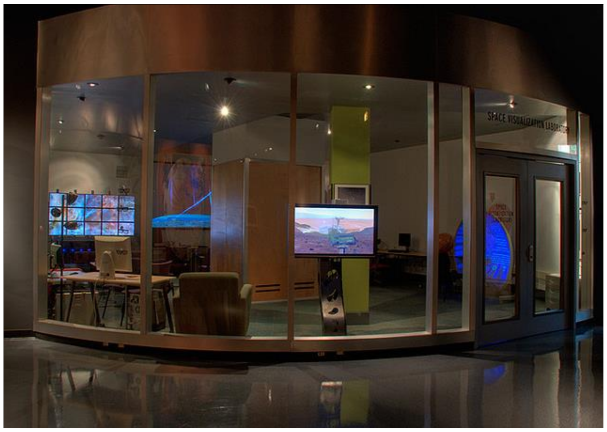

Figure 1. The Space Visualization Laboratory.

Figure 2. Guests taking the pre-test.

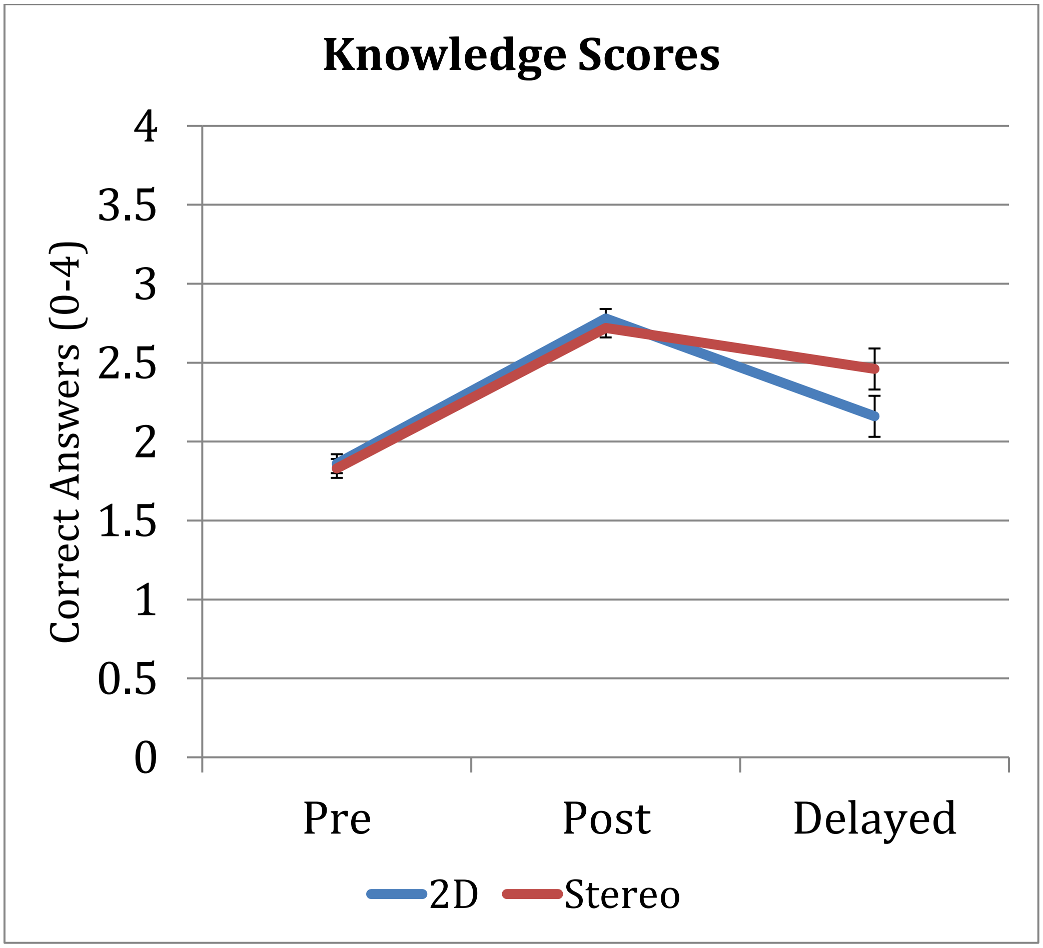

Figure 3. Scores on the knowledge section of the test.