AN ABSTRACT OF THE DISSERTATION OF

Kathryn A. Stofer for the degree of Doctor of Philosophy in Science Education presented on April 29, 2013.
Title: Visualizers, Visualizations, and Visualizees: Differences in Meaning-Making by Scientific Experts and Novices from Global Visualizations of Ocean Data.

Abstract approved: _____
Shawn M. Rowe

Data visualizations designed for academic scientists are not immediately meaningful to everyday scientists. Communicating between a specialized, expert audience and a general, novice public is non-trivial; it requires careful translation. However, more widely available visualization technologies and platforms, including new three-dimensional spherical display systems in schools and informal science education institutions, often use the same visualizations that experts use to communicate amongst themselves, resulting in a public which often fails to make significant meaning from the visualizations.

This dissertation uses a pragmatic, bricolage framework, incorporating cognitivist, social constructivist, and sociocultural perspectives. I used semi-clinical interviews and eye-tracking to investigate academic scientific experts and novices as they attempted to make meaning from global visualizations of ocean data. Stimuli were fifteen visualizations, three topics with five versions of each visualization with different levels of scaffolding to improve communication: no scaffolding; changes to color scale; addition of geographic labels; revision of title and measurement unit; or all three forms.

Laboratory interviews revealed that non-science major novices struggled with decoding almost every part of unscaffolded visualizations, while experts had difficulty only in understanding the time of year and season represented. Novices did not always use supporting elements such as the title and key, could not understand jargon in unscaffolded titles, conflated the meaning of the standard academic science “rainbow” color scale used across multiple topics, and could

not always orient themselves geographically to the visualizations centered on the Pacific Ocean basin. However, their understanding improved on the scaffolded visualizations. Interviews in a public interpretive science center revealed further struggles with meaning-making; scores were lower than either laboratory participant group.

Eye-tracking confirmed the differences between the participant groups at the level of visual search of visualizations, revealing that novices looked at the map portion of the visualizations less comprehensively than experts in the unscaffolded case. However, novice scan paths on the scaffolded visualizations more closely resembled experts'. Fixation durations started out significantly lower on scaffolded visualizations than unscaffolded, suggesting better comprehension of the scaffolded visualizations. Both participant groups' fixation durations decreased over the course of repeated trials in the experiment, suggesting practice improved meaning-making.

The fact that novices could make more academic scientific meaning from visualizations of data if exposed more often to meaningful, scaffolded visualizations in all formal and informal learning and communication settings leads to recommendations for exhibit design, visualization design, and instruction on using visualizations in meaning making about science topics.

Visualizers, Visualizations, and Visualizees: Differences in Meaning-Making by Scientific Experts and Novices from Global Visualizations of Ocean Data

by
Kathryn A. Stofer

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of

the requirements for

the degree of

Doctor of Philosophy

Presented April 29, 2013

Commencement June 2013

Doctor of Philosophy dissertation of Kathryn A. Stofer presented on April 29, 2013.

APPROVED:

Major Professor, representing Science Education

Dean of the College of Education

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Kathryn A. Stofer, Author

ACKNOWLEDGEMENTS

Extensive thanks go to many people who made this project possible and more importantly, finishable. First of all, to my advisor Dr. Shawn Rowe, for doing all it took. Thanks also to my committee, John Falk, Amy Lobben, Jon Palfreman, and Stacey Harper, for challenging and supporting me through the process.

In no particular order: Laura Good, for endless advice and patience as we got through it together. Teresa Wolfe for editing and hockey. Laia Robichaux, for "running tech" for my defense. John Baek, for discussions that led to the idea of combining neuroscience methods with more traditional science education methods. Jennifer Bachman, for serving on my committee for my oral exam, and along with her husband, Greg Kise, offering technical advice and solutions. Mark Farley, for being Mark. The FCL lab group and My SMED PhD cohort, Matt Campbell, Mike Furuto, and Becka Morgan, for listening over the years. Celeste Barthel, Susan O'Brien, and Michelle Mileham for making sure my committee was well-fed for my defense. Paula Dungjen for helping me cross all the t's and dot all the i's. Lauren Atwell, for final copy-editing.

Rob Simmon and Jesse Allen of NASA Goddard Space Flight Center for developing visualizations for me to use. The IRB, especially Candi Loeb, for getting and keeping this project ship-shape. Xuan Che, for helping me from soup to nuts with my statistical analysis. Anthony Hornof, Christy Steffke, Julie Libarkin, and SMI, particularly Albert Crooker and Mark Mento, for eye-tracking advice and troubleshooting.

I am grateful for the funding for this project, which came from the Curtis and Isabella Holt Award from Hatfield Marine Science Center. Funding for me over the years has come from the National Science Foundation, the National Oceanic and Atmospheric Association, the U. S. Department of Education, and the

Howard Hughes Medical Institute. I also thank the students of the Hatfield Student Organization for a travel award and Oregon Sea Grant for their support.

The SST, climatology, and chlorophyll data were obtained from the Physical Oceanography Distributed Active Archive Center (PO.DAAC) at the NASA Jet Propulsion Laboratory, Pasadena, CA. http://podaac.jpl.nasa.gov.

All my friends who didn't abandon me when I abandoned them to get this thing done.

And saving the best for last, my family, without all ya'll I would never be this far in the first place. Thank you all so much. I love you.

TABLE OF CONTENTS

Page

Introduction 1

Rationale 1

Purpose of Study 10

Literature Review 14

Differences Between Academic and Everyday Science Thinking 15

How Experts Create and Use Visualizations 20

Modeling Expert Thinking 32

Bridging Multiple Disciplines 43

Methods 47

Conceptual Framework 47

Research Plan 57

Sampling 58

Visualization Preparation 61

Location Setup 64

Pilot Testing 67

Experimental Procedure 68

Analysis 72

Triangulation of Methods 78

Methodological Limitations 79

TABLE OF CONTENTS (Continued)

Page

Interview Results 82

Participant Characteristics 82

Scoring 86

Analytic Coding – Expected Codes 98

Visualization Element Use 108

New Codes 119

Specific Problems with Visualization Elements 123

In situ Interviews During Eye-tracking 125

Putting the Puzzle Pieces Together 129

Summary 136

Eye-tracking Results 138

Participant Calibration Results 138

Quantitative Results 139

Qualitative Results 170

Summary 191

Discussion 193

Research Question 1 194

Research Question 2 198

Research Question 3 201

Research Question 4 204

TABLE OF CONTENTS (Continued)

Page

Limitations 206

Generalizability 212

Future Work 222

Conclusion 223

Epilogue: Reflection On A Pragmatic Framework 225

Bibliography 229

Appendices 250

LIST OF FIGURES

Figure — Page

1. Example of a Spherical Display System exhibit 9
2. *In situ* eye-tracking setup 66
3. Example of figure-ground illusion, “Rubin’s Vase” 109
4. Density curves of individual participants’ spontaneous looking fixation durations 141
5. Log-transformed histograms and density curves, individual participants’ fixation durations 142
6. Linear approximation of effect size, fixation duration versus trial, Unscaffolded and Fully-Scaffolded 148
7. All experts’ versus all novices’ eye-tracking scan paths, Unscaffolded, SL 172
8. All experts’ versus all novices’ eye-tracking scan paths, Fully-Scaffolded, SL 173
9. All novices’ eye-tracking scan paths, Unscaffolded versus Color Scaffolding, SL 174
10. All novices’ eye-tracking scan paths, Unscaffolded versus Geographic Scaffolding, SL 176
11. All novices’ eye-tracking scan paths, Unscaffolded versus Title Scaffolding, SL 177
12. All experts’ scan paths, Unscaffolded versus Fully-Scaffolded Cases 178
13. One expert’s scan path, Fully-Scaffolded, SL 179
14. All experts’ and all novices’ eye-tracking heat maps, Unscaffolded, SL 181
15. All novices’ eye-tracking heat maps, Unscaffolded versus Geographic Scaffolding, SL 182

LIST OF FIGURES (Continued)

Figure Page

16. All novices' eye-tracking heat maps, Unscaffolded versus Color Scaffolding, SL 183

17. All novices' eye-tracking heat maps, Unscaffolded versus Title Scaffolding 184

18. All novices' eye-tracking heat maps, Unscaffolded versus Fully-Scaffolded, SL 186

19. All experts' eye-tracking heat maps, Unscaffolded versus Fully-Scaffolded, SL 187

20. All novices, Area of Interest versus Time, Unscaffolded case, SL 188

21. All experts, Area of Interest versus Time, Unscaffolded case, SL 189

22. All novices, Area of Interest versus Time, Fully-scaffolded case, SL 190

23. All experts, Area of Interest versus Time, Fully-scaffolded case, SL 190

LIST OF TABLES

Table		Page
1.	Presented Gender Balance, Participants versus Population	60
2.	Stimulus Topic Presentation and Order, Clinical Interviews	69
3.	Stimulus Topic Presentation, Laboratory Eye-tracking	70
4.	Areas of Interest Associated with Scaffolding Levels	76
5.	Sub-population Comparison Tests, Eye-tracking	77
6.	Novice Clinical Interview Participant Characteristics	83
7.	Novice Eye-tracking Participant Characteristics	84
8.	Expert Participant Characteristics	84
9.	Experts' Overall Score Accuracy Index Compared to Frequency of Use and Comfort Level for Interpreting Visualizations	87
10.	Accuracy in Clinical Interviews, All Participants (*N* = 29)	88
11.	Scoring Accuracy in Clinical Interviews, Novice versus Expert	89
12.	Scoring Accuracy on Eye-tracking Interviews, All Participants (*N* = 20)	90
13.	Scoring Accuracy on Eye-tracking Interviews, Novice versus Expert	91
14.	Scoring Accuracy on Clinical Interviews by Scaffolding Level, All Participants (*N* = 29), All Novices, and All Experts	92
15.	Scoring Accuracy on Eye-tracking Interviews by Scaffolding Level, All Participants (*N* = 20), All Novices, and All Experts	93
16.	Scoring Accuracy on Clinical Interviews by Topic, All Participants (*N* = 29)	95
17.	Scoring Accuracy on Clinical Interviews by Topic, Novices versus Experts	95

LIST OF TABLES (Continued)

Table Page

18. Scoring Accuracy on Eye-tracking Interviews by Topic, All Participants (*N* = 20), All Novices, and All Experts 96

19. *In Situ* Eye-tracking Participant Description (*N* = 13) 126

20. Accuracy, *In Situ* Eye-tracking Interviews (*N* = 13) 127

21. Number of Fixations per Participant per Visualization, 10 seconds Spontaneous Looking, All Participants, All Novices, and All Experts 140

22. Durations of Fixations for All Participants, All Novices, and All Experts, SL 140

23. Truncated Linear Regression Model with Five Independent Variables, All Participants (*n* = 18), SL 144

24. Truncated Linear Regression Model with Five Independent Variables and Level and Trial Interaction Term, All Participants (*n* = 18), SL 144

25. Truncated Linear Regression Model with Gender as Main Effect on Fixation Duration, All Participants (*n* = 18), SL 145

26. Truncated Linear Regression Model with Expertise as Main Effect on Fixation Duration, All Participants (*n* = 18), SL 145

27. Truncated Linear Regression Model with Expertise, Level, and Trial as Main Effects on Fixation Duration, All Participants (*n* = 18), SL 146

28. Best Fit model: Truncated Linear Regression Model with Level and Trial Main Effects and Interaction Term on Fixation Duration, All Participants (*n* = 18), SL 147

29. Truncated Linear Regression Model with Level and Trial Main Effects and Interaction Term on Fixation Duration, Unscaffolded versus Fully-Scaffolded Case, All Participants (*n* = 18), SL 147

LIST OF TABLES (Continued)

Table Page

30. Best Fit Truncated Linear Regression Model with Level and Trial Main Effects and Interaction Term on Fixation Duration, Unscaffolded, Single Scaffolding, and Fully-Scaffolded Case, All Participants (*n* = 18), SL 150

31. Truncated Linear Regression Model with Topic as Main Effect, Pairwise Comparisons, All Participants (*n* = 18), SL 151

32. Number of Fixations per Area of Interest, All Participants (*n* = 18), SL . 153

33. Relative Risk Ratios for Areas of Interest, Unscaffolded versus Geographic Scaffolding, All Participants (*n* = 18), SL 155

34. Relative Risk Ratios for Areas of Interest, Unscaffolded versus Colors Scaffolding, All Participants (*n* = 18), SL 156

35. Relative Risk Ratios for Areas of Interest with Level as Main Effect, Unscaffolded versus Title Scaffolding, All Participants (*n* = 18), SL 157

36. Relative Risk Ratios for Areas of Interest with Level and Expertise as Main Effects, Unscaffolded versus Title Scaffolding, All Participants (*n* = 18) , SL 157

37. Relative Risk Ratios for Areas of Interest with Level and Trial as Main Effects, Unscaffolded versus Title Scaffolding, All Participants (*n* = 18) , SL 158

38. Relative Risk Ratios for Areas of Interest, Unscaffolded versus Fully Scaffolded, All Participants (*n* = 18), SL 159

39. Mean Trial Duration (ms), All Participants (*N* = 20), All Novices, and All Experts, MI 160

40. Number of Fixations per Trial, All Participants (*N* = 20), MI 161

41. Fixation Durations (ms) for All Participants, (*N* = 20), MI 162

42. Truncated Linear Regression Model with Five Independent Variables, All Participants (*N* = 20), MI 163

LIST OF TABLES (Continued)

Table | Page

43. Truncated Linear Regression Model with Topic and Gender Main Effects, All Participants (*N* = 20), MI 164

44. Number of Fixations per Area of Interest, All Participants (*N* = 20), MI 167

45. Relative Risk Ratios for Areas of Interest with Expertise as Main Effect, Unscaffolded versus Single-Scaffolding Cases, All Participants (*N* = 20), MI 168

46. Relative Risk Ratios for Areas of Interest with Expertise as Main Effect, Unscaffolded versus Fully-Scaffolded Case, All Participants (*N* = 20), MI 169

LIST OF APPENDICES

Appendix		Page
1.	Stimulus Visualizations	251
2.	Interview Protocols	257
3.	Areas of Interest	261
4.	Interview Scoring Rubrics	263
5.	Qualitative Code Book	272
6.	Lists of Jargon Words Used	274
7.	Eye-tracking on a 3-D Digital Exhibit	276

LIST OF APPENDIX FIGURES

Figure — Page

A1. Unscaffolded (US) SST (sea surface temperature) visualization 251

A2. Unscaffolded (US) SST anomaly visualization 252

A3. Unscaffolded (US) chlorophyll visualization 253

A4. Fully-Scaffolded (FS) SST visualization, including geographic (GS), color (CS), and title and key (TS) scaffolding 254

A5. Fully-Scaffolded (FS) SST anomaly visualization 255

A6. Fully-Scaffolded (FS) chlorophyll visualization 256

A7. “Larger Map” Areas of Interest 261

A8. “Smaller Map” Areas of Interest 262

A9. Comparison of typical and suspicious scan paths 283

LIST OF APPENDIX TABLES

Table — Page

A1. SST or One Month Average Sea Surface Temperature Scoring Rubric 264

A2. SST Anomaly or One Month Average Sea Surface Temperature Difference from Average Scoring Rubric 266

A3. Chlorophyll-a or One Month Average Ocean Productivity or One Month Average Microscopic Ocean Plant Concentration Scoring Rubric 269

A4. Topic-independent Questions Scoring Rubric 271

A5. Most Frequently Used Jargon by Participant Subgroup 274

A6. Additional Jargon Used by Experts 274

A7. *In Situ* Eye-tracking Stimulus Presentation Order by Level and Topic 279

For my grandmothers, Nina and Betty.

Visualizers, Visualizations, and Visualizees: Differences in Meaning-Making by Scientific Experts and Novices from Global Visualizations of Ocean Data

Introduction

Rationale

Science has been a relatively standardized part of American school curricula since the 1880's, driven largely by the establishment in the 1860's of Land Grant universities emphasizing agriculture research and mechanical education and shortly thereafter public high schools preparing students for technical jobs (Cremin, 1990; Reese, 1999). However, the rationale for promoting science understanding has changed dramatically since the 1960's. Spurred originally by the Cold War and the "Space Race," as science and technology have become an integral part of the world economy and information has become a commodity available to more and more non-professionals, a broad scientific understanding is no longer a luxury but increasingly a necessity in today's global society. Today the importance of science to society demands a person not only build basic academic science knowledge and skills as a youth, but also use and continually develop that scientific understanding throughout the lifespan in order to make decisions on a personal, professional, and societal level. However, the way in which non-professional scientists think of professional, "academic" science differs vastly from the way professional scientists and trained communicators and educators of "academic" science understand the discipline.

Professional scientists provide new data bearing on myriad problems daily. With increasing frequency, academic science data is accessible to anyone with an internet connection. Previously, the non-scientist could rely on professionals to weed out good and bad information. Today, with this glut of often unfiltered information, the non-specialist is increasingly called upon to assess the quality and usefulness of the new findings, often very rapidly, in order to refine their understanding and continue to stay truly informed. On a personal level, for example, medical information abounds, but having a conversation with one's

doctor to evaluate various possible treatment options requires understanding of who's recommending what, for what reasons, and how similar any previous research may be to your own situation. Similarly, one must be able to evaluate scientific information to make decisions about energy use, such as the relative financial or ecological consequences of choosing to insulate one's residence rather than buying more updated appliances.

Further, access to this information alone is not enough (Rennie & Stocklmayer, 2003). To create a fully-participatory scientific society, not only must individuals grapple with many academic scientific questions that inform potential policy on a city-wide, regional, or global scale, but also the public must be the ones driving the quest for further information (Briggs (2001), cited in Rennie & Stocklmayer, 2003, p. 770). First, as scientific development allows more and more access to information, weighing ethics related to health care becomes an issue for the public. For example, a vote on whether or not to allow prenatal genetic testing to be used in a certain manner calls for an understanding by the voter of what, exactly, the academic scientific knowledge resulting from such tests might clarify and how the knowledge could complicate things. Namely, additional information might enlighten parents about genetic risks that could face their children but complicate the picture by raising ethical issues around whether or not to birth children with increased risk of deformities or long-term or life-shortening illnesses. Another example is terrorism prevention. Rational debate over security versus privacy, between conducting surveillance covertly or overtly, requires understanding the technology that could be used, its capabilities, and its risks. Globally, avoiding or mitigating impacts of climate change demands understanding of not only how the Earth system works, but also how scientific modeling predicts what will happen in the future. Using evidence to choose which model is strongest in which situation, and on which to base society's actions, requires recognizing how that model was built, on what evidence, and what assumptions and unknowns still exist. People may not always choose to act

rationally when they are capable of doing so, but if they are not able even to understand and use information rationally, they definitely are not able to act rationally.

Next, increasingly, science policy advocates that academic scientists must actually be driving their research agendas based on what the public needs and wants to know (McCallie, 2010). From this perspective, academic science should be considered equally alongside other ways of knowing in a socio-scientific society, all the myriad socio-cultural norms and ways of life that society outside traditional academic science already strives to incorporate (Jiménez-Aleixandre & Pereiro-Muñoz, 2002; Kolstø, 2001a; Ratcliffe & Grace, 2003; Sadler, 2004; Sadler & Zeidler, 2003).

Thus the justification for academic science education is shifting from promoting achievement of simple academic scientific literacy, wherein a non-scientist might build a knowledge base of thoroughly-vetted academic scientific information, to working on "functional scientific literacy" (Zeidler, Sadler, Simmons, & Howes, 2005), or "civic scientific literacy" (J. D. Miller & Pardo, 2000; Shen, 1975). This is mirrored by the broader shift from a "Public Understanding of Science" movement to a "Public Engagement with Science" movement. The former developed under the supposition that "if the public understands science, they will value it," in a time of crisis for science in the public image (McCallie, 2010). The source of the problem that the public understanding of science movement sought to address was ostensibly the public, who were viewed as lacking scientific knowledge. The thought was that increased knowledge would bring understanding and with it, a restoration of science's exalted position, an assumption that has since been questioned (Allum, Sturgis, Tabourazi, & Brunton-Smith, 2008).

The more recent movement is public *engagement* with science, wherein the role of the sociocultural contexts of publics are not only taken into account in designing communications about science, but are ostensibly equally valued in a

multicultural socio-science community of both academic and everyday scientists. This shift includes a shift to a dialogue model among the members of the group and its sub-groups in which the groups “mutually [inform each other] ... in which participants listen, respond, refute, and build on one another’s contributions, and perspectives” (McCallie, 2010, p. 41). These discussions can either be specifically aimed at informing science and technology policy or simply aimed at promoting this dialogue and engagement around science (Lehr et al., 2007).

These new definitions imply an ongoing process, whereby the group uses emerging academic scientific information to inform decision-making. However, this fluency and capability for dealing with recent, almost raw research requires more than just knowledge; it also necessitates an understanding of how science works, how scientists process and build arguments from data, and how scientists represent findings in reports. Thus professional scientists, for their part, must be sensitive to the needs of the larger socio-scientific society to make clear these processes that are enculturated mainly through professional training.

Currently, much of the population of the United States performs poorly on surveys attempting to measure their formal, academic scientific knowledge (J. D. Miller & Pardo, 2000). A portion of this performance is undoubtedly due to a social and cultural gap between academic science and everyday understanding of the world. Many adults are also thus judged as not scientifically literate, that is, they lack either “a basic vocabulary of scientific concepts ... a general understanding of the nature of scientific inquiry,” or both (J. D. Miller, 2004, p. 273). In the 1990's, only 12% of adults in the United States actually qualified in surveys as scientifically literate, and only 29% of those were considered “attentive” to science, meaning they report both a high level of interest in and a sense of being informed about science (J. D. Miller & Pardo, 2000). Today, that number has risen to approximately 28% as literate, and our global competitiveness has increased, but that increase has been largely driven by college education (Raloff, 2010), to which only a minority of the population has

access. Furthermore, most adults gain more science knowledge outside of formal school than within (Falk & Dierking, 2010; Falk & Needham, 2013), such as at science centers (Falk & Needham, 2011). Finally, analyses of science literacy that seek to study science knowledge in situated contexts report much higher literacy levels (Falk & Needham, 2011).

Neither of these sources of content, however, address the larger issue of a broader definition of literacy. Self-reports of understanding on certain scientific topics and affective attitudes toward science may not be reflected in strictly content-based surveys (Falk & Needham, 2011). Adults in literacy surveys report high levels of interest in scientific developments (J. D. Miller & Pardo, 2000; J. D. Miller, 2004, 2010). Thus there is a disconnect between academic knowledge levels and interest in science, and between definitions of literacy as content and definitions that more broadly include skills and attitudes.

This difference in definitions points to the existence of another potential problem, then, a problem of accessing and using scientific knowledge for decision-making. This, arguably, is a problem of communication between scientists and the public based around different perceptions of the other population. Just as what has previously been deemed a “deficit” in certain populations of schoolchildren has been rethought in more contemporary educational parlance as cultural difference, so too are communicators and educators of academic science realizing that it is not that non-scientists lack science-related ability, but rather they use different skills or bring knowledge to bear on problems in different ways from academic scientists. These social and cultural differences in particular affect what learning academic science is, how learning academic science happens, who produces academic science, and how products of academic science learning are transferred from school-context to life-context, where personal and societal decisions are made (Bransford, Brown, & Cocking, 2000). In order to effectively reframe academic science communications to everyday science communications, the communities would be

well-served to consider “culturally-relevant pedagogy” (Ladson-Billings, 1995). I suggest modifying two of Ladson-Billings’ three broad propositions thusly to enact culturally-relevant communication between the academic and everyday science communities. Each community must conceive of the other as capable, and everyone must have a shared conception of knowledge. This will allow effective production of culturally-relevant academic science information, and dissemination and evaluation of that information through dialogue as it relates to socio-scientific issues.

The context of the conversation is also shifting from “public understanding of [academic] science” to “public engagement with science,” especially in informal settings (McCallie, 2010; Pedretti, 2004; Rennie & Stocklmayer, 2003). This move reframes the context of the reason for promoting science to increasing awareness, improving attitudes and beliefs instead of simply increasing academic knowledge. This movement also encompasses activities aimed at getting non-professionals involved in academic science through participating as “citizen scientists” or as “public participants in science” (Bonney et al., 2009), and particularly in dialogue events (McCallie, 2010).This discussion also embraces the bringing together of everyday knowledge about science with academic science and bridging the language differences and literacy differences between the two worlds from both directions, rather than unilaterally trying to bring non-professionals into the professional academic scientists’ culture.

Life-long learning can help to increase a user’s fluency with academic scientific language and skills. Jon Miller (2004) argues “a scientifically literate citizen needs to have: (1) a basic vocabulary of scientific terms and constructs; and (2) a general understanding of the nature of scientific inquiry” (J. D. Miller, 2004, p. 273). Miller and Pardo’s (2000) literacy surveys found that while college science classes had the most influence on scientific literacy, use of informal resources, including not only media but also science museums, also positively relates to literacy. The National Research Council (2009) reports that such

informal education can “bolster science education broadly on a national scale,” and do so particularly for adults who have moved beyond the reach of formal school classrooms.

Some of the most powerful tools for communicating academic science and evoking discussions about evidence, especially in informal education settings, are visualizations of scientific data. Often these take the form of graphs, tables, or charts, but increasingly scientists use images, illustrations, and animations, representations that depict vast amounts of data in concise, comprehensible pictures. In particular for Earth sciences, these often take the form of “spatially-based” visualizations, which show data as gradations of color overlaid on a map, revealing patterns to the eye that matrices of numbers obscure. The visualizations illuminate Earth's processes to scientists and are supposed to communicate research findings. It is these “spatially-based visualizations of data”, in particular overlays on global maps, which are the focus of this dissertation and are hereafter referred to as “visualizations.”

The spatial nature of the visualizations considered here indicates the need for not only scientific skill, but also graphical or spatial skill. While spatial reasoning is an important component of meaning-making from these visualizations, this dissertation focuses on the meaning-making involved in academic science communication. However, a few points regarding spatial literacy warrant mention. In particular, the visualizations under consideration here require two spatial thinking tasks: (a) describing and interpreting objects, specifically, patterns; (b) comprehending spatial properties and processes, specifically, using maps (Kastens & Ishikawa, 2006). For the first task, cognitive science work suggests experts are able to see patterns in scientific data and imagery where novices cannot because the experts understand the significance of those patterns in that context (Gilhooly, Wood, Kinnear, & Green, 1988; Lesgold et al., 1988). For the second, map use may rely heavily on spatial ability, and map users with low spatial skill may not even benefit from concurrently-

presented verbal information (Cook, Wiebe, & Carter, 2008). Fortunately, though early thinking on the participant suggested spatial skill was an inherent ability (French, Ekstrom, & Price, 1963; Linn & Peterson, 1986; National Academies Press (U.S.), 2006), more recent research reveals this ability can be learned (National Academies Press (U.S.), 2006), although there are influences of gender and chronological maturity (Piaget & Inhelder, 1956).

However, until they are taught, learners may not be able to use visualizations to make scientific meaning if those visualizations are designed for people with high spatial ability, namely expert scientists, and are not significantly modified for a non-expert audience. Other scientists and functionally-scientifically-literate people can use the visualizations to judge the scientific value of the data, grasp the current state of understanding, and apply new knowledge by supporting laws, politicians, and businesses. However, many everyday scientists cannot make sense of visualizations that are not adapted from the academic science context to be culturally relevant (Phipps & Rowe, 2010; Rowe, Stofer, Barthel, & Hunter, 2010).

In the last ten years, spherical display systems have emerged as a new avenue for disseminating such imagery to the public. In these exhibits, computers project rectangular two-dimensional images onto a three-dimensional spherical screen, purportedly removing the confusion introduced from an inevitably distorted view by displaying the imagery “intuitively” on a globe (“Products,” 2008). Viewers in science centers report seeing “more realistic” data and getting a better perspective when viewing Science on a Sphere™ (Haley Goldman, Kessler, & Danter, 2010). See Figure 1 for an example of a spherical display system (SDS) on exhibit in a science center. The flexibility inherent in the system also allows for the images to be easily updated and presented in near-real-time in many cases. The sheer size and picture quality (number of pixels, not excellence at conveying content) of imagery displayed on the globe makes them very appealing to life-long learners.

Figure 1. Example of Spherical Display System exhibit. Visitors view the Magic Planet™ 3-foot diameter spherical display system with interactive touch-screen kiosk.

However, making use of these display systems in a free-choice, hands-on environment typical of most life-long learning institutions requires careful design. In most cases, these systems are presented for the visitor to use on their own, without the benefit of knowledgeable staff or volunteers to guide their exploration or offer clarifying information. Therefore, the display and its navigational controls must essentially be self-contained. The exhibit must be easy for visitors to use, as any energy visitors must expend on understanding the navigation depletes the energy they have available for making sense of the content (Allen, 2004). The Philadelphia-Camden Informal Science Education Collaborative (PISEC) identified seven characteristics of effective family learning exhibit design; the first three relate to the physical setup: multi-sided, multi-user, and accessible (Borun & Dritsas, 1997).

The last four apply to the content of the experience. According to PISEC (Borun & Dritsas, 1997), exhibits must be multi-outcome, multi-modal, readable, and relevant. Multi-outcome and multi-modal suggest the experience be designed to foster group interaction. In particular, readability and relevance as criteria apply to visualization design. The criterion that "'text' is arranged in easily understandable segments" (Borun & Dritsas, 1997, p. 23) translates to elements of the visualization aiding meaning-making, be it accompanying text or the patterns in the data of the map itself. Any energy the visitor must spend translating the text or color scheme detracts from the attention they can pay to the meaning that the scientists or educators intend to communicate. For the relevance criterion, the content must be sufficient to build on visitor's prior knowledge and obviously connect to that knowledge; that is, it can be neither too low-level nor too high-level (Allen, 2004; Borun & Dritsas, 1997). Finally, the threshold for engagement with the exhibit and content must be low to encourage visitors to stop, interact, inquire, and work to make deeper meaning that is necessary to learn with and from visualizations. Exhibits that engage the entire visitor group are more likely to have improved inquiry and attitude outcomes than those that do not (Gutwill & Allen, 2010).

To create a solvable visualization meaning-making task for non-scientists, it remains unclear which particular adaptations are necessary to scaffold the visualizations, in what order the scaffolds should be presented, and when they should be removed. Therefore, understanding how users acquire information from the visualizations and how visualizers facilitate optimal communication is of fundamental importance to both science and science education.

Purpose of Study

This research aims to answer several questions. Can the public, who presumably lack extensive academic scientific training, make academic scientific meaning from visualizations of real-world and near-real-time phenomena to

better understand these phenomena? What features of the visualizations are salient to academic and everyday scientists? What elements are confusing to those groups, especially when the visualizations are presented without alteration from the academic context? Can I compensate for confusing elements and limited academic background through the data pattern presentation and supporting text features? Thus, this dissertation explored the following research questions:

1. What are the "bottlenecks" preventing everyday scientists from making meaning from data visualizations?
2. What strategies do academic scientists use to make meaning from the visualizations, and how do reported conscious strategies correlate with their perceptual strategies?
3. How can these academic scientists' strategies be compensated for and ultimately explicitly modeled for application to new visualizations?
4. How can methods that are based in diverse epistemologies be triangulated to provide insight into these research questions and larger problems?

I used an expert-novice framework to examine how users with specific training in making meaning from these visualizations, namely academic scientists, explore the patterns in the data and use the supporting information compared to users who have not had such training. I used both conscious self-reporting measures via interviews and measures of unconscious search patterns via eye-tracking to compare various strategies within and between these groups for making meaning. Specifically, I determined: where there is overlap in strategy use and meaning-making, where there is alternate strategy use that leads to similar meaning-making, and where there is inconsistency in making academic scientific meaning. When there were differences in strategies or meaning-

making, I examined whether that resulted from lack of information provided in the visualizations, difference in academic science training, or other factors.

I designed particular scaffolds based on earlier work (Phipps & Rowe, 2010) to further understand how meaning-making in both groups is supported or confounded by additional information and how the skills each group possesses manifest in the meaning-making process. Previous justifications of this type of work have focused on what's lacking in one population or another, a deficit model. I wished to reframe the discussion around discovering what each population can offer the other, that is, what assets each group brings to the table, their “funds of knowledge” (González, Moll, & Amanti, 2005). From there, I could work to question the assumptions academic and everyday scientists have about each other and work to build on each other’s strengths (Weiner, 2006), valuing their diversity (Fayden, 2005), rather than casting one group as “right” and the other to be brought "up to snuff.” Only when the groups’ voices are equal will true public engagement with science be possible.

The scientists that served as expert visualization users in this study do possess expertise in this area based on their extensive training particular to the task at hand. To convey the task-related expertise and allow for a more level playing field assumed in other situations, the terms “expert” and “novice” are used here as shorthand to refer to “expert visualization user” and “novice or apprentice visualization user,” respectively. On its own, “novice” could convey that a user is completely without skill in *any* area, which is of course a complete mistake; even newborn infants have some skills and tools at their disposal, though not culturally-derived skills and tools. The novice visualization readers in this experiment could be expert foreign language speakers, expert writers, or the like, without having been trained in the skill of visualization reading. On the other hand, the expert visualization readers are certainly not expert in every aspect of life, as they may completely lack skills in swimming or maintaining a budget. However, by finding what skills are more in common among novice visualization

readers that they may bring to bear on the unfamiliar task, and at the same time understanding what skills the expert visualization readers use to complete the task, I could move towards the ability to break down the task of visualization reading. This, in theory, would allow the novice visualization reader to move toward expertise by leveraging the skills they naturally use, understanding skills they misuse for the task, or targeting learning of skills expert visualization readers have learned or developed through their training.

Ultimately, it was my hope that this dissertation would lead to more insight on what disconnects exist between highly-trained scientists, who readily make meaning from spatially-based global visualizations, and the less-experienced publics with which they would also like to communicate through these visualizations. By doing so, I might be able to improve these tools as well as help experts understand what processes they have automated over time and must be sure to explicate for certain audiences. Thus, the aim was to create more opportunities for larger sectors of society to be not only consumers of but, importantly, critical consumers of (S. Miller, 2001) and eventually producers of scientific meaning.

Literature Review

One of the biggest differences between academic scientists and everyday scientists is in their methods of making decisions (Shen, 1975; Zeidler et al., 2005). If we know how experts think, we can design better interventions to allow and encourage non-scientists to work more like experts when considering scientific problems, or using scientific tools such as visualizations of data, as half of the equation to bring the two groups into a multicultural socio-scientific culture. This literature review will outline the current understanding of how experts and novices think differently in general and in academic science, specifically how visualizations of data exemplify the way experts think, and how expert thinking about visualizations can be modeled in school or out. I close with a brief review of literature pertinent to bridging multiple disciplines to make this happen.

Middendorf and Pace(2004) outlined seven steps to breaking down expert thinking and novice thinking for classroom teaching:

> Step 1. What is a bottleneck to learning in this class?
> Step 2. How does an expert do these things?
> Step 3. How can these tasks be explicitly modeled?
> Step 4. How will students practice these skills and get feedback?
> Step 5. What will motivate the students?
> Step 6. How well are students mastering these learning tasks?
> Step 7. How can the resulting knowledge about learning be shared? (Middendorf & Pace, 2004, p. 3)

The first three are aligned with the first three particular research questions in this dissertation. The first three sections of this literature review also address these questions. Research Question 1 aligns with Step 1 and is addressed in understanding how the groups think differently, Research Question 2 and Step 2 relate to how experts work, and Research Question 3 and Step 3 discuss modeling expert thinking and helping novice visualization users build skills.

Differences Between Academic and Everyday Science Thinking

Academic scientists differ from everyday scientists in meaningful ways. To be credentialed as an academic scientist in the United States, one needs to complete at least a four-year undergraduate degree in a scientific field, and more commonly a master's and doctoral degree on top of that. Beyond that additional six to ten years of post-baccalaureate schooling, many scientists who want to do research must then complete post-doctoral training before securing a university position (Ragauskayte, n.d.).

As the undergraduate degree also involves so-called "core" coursework not directly in the major subject, a bachelor's degree in science may provide only about two years of full-time coursework in science. This emphasis on breadth rather than depth offers few chances to organize what knowledge is gained (Bransford et al., 2000). Only about half of undergraduates in science, technology, engineering, or math report having any research experience as an undergraduate (Russell, Hancock, & McCullough, 2007).

Thus, it is typically not until after obtaining a bachelor's degree that future academic scientists start the practice of science required to become an expert. There are four general well-established characteristics of experts: knowledge, skills, cognition and flexibility, with components such as intuitive performance (skills), seeing the big picture (cognition), and managing adversity (flexibility) (Grenier, 2009). Acquiring expertise requires not any practice in the domain, but deliberate practice (Ericsson, Krampe, & Tesch-Römer, 1993). In science research, this deliberate practice generally means work toward publications, promotion and tenure, and presentations of one's research (K. Nandogopal, personal communication, January 24, 2011).

In particular, in one fundamental practice of science we see the gap: experts and novices observe the world differently, noticing different pieces of information and organizing that information into different structures (Bransford et al., 2000; Eberbach & Crowley, 2009; Kastens & Ishikawa, 2006). Part of that is

because the experts possess more information than novices about particular scientific subjects, but related to that, the harder part to understand and teach is how scientists interpret what they see. This interpretation is learned through participation in scientific culture and communities of practice, but is shaped by individual experiences and stored in an individual's brain and its physical collection of neurons and networks. Understanding how experts make meaning through observation requires studying both the cultural uses of tools and methods of noticing as well as examining the physical organization of knowledge and, ultimately, connections between functional regions in the brain.

The difference between "seeing" biologically, that is, perceiving physical stimuli, and "observing" or "noticing" culturally, that is, within scientific contexts, lies in recognizing what is meaningful and comparing it to other knowledge and theory (Eberbach & Crowley, 2009). Thus, as Eberbach and Crowley point out, a novice can see biologically but the culture of noticing scientifically is different from "everyday" culture of observation and must be learned so that one can use biological mechanisms to one's advantage for scientific judgments. In particular, besides the skill differences in using information once obtained that comes with expertise, experts' ways of gathering information through observing and their purposes for observing differ from novices; experts may appear to be simply describing, but "they are simultaneously inferring ... relationships and testing hypotheses about the causal order" (Eberbach & Crowley, 2009, p. 42).

Culturally, while observation is a natural practice for everyday life, practiced in many cultures around the world, the everyday practice is quite different from the expert, scientific practice (Eberbach & Crowley, 2009). Some cultures train observing as a primary source of learning, such that children are “ready to carry out an aspect of the activity being observed ... immediately” (Paradise & Rogoff, 2009, p. 110) the first time they are asked. This type of observation often involves intense concentration (Paradise & Rogoff, 2009). Novices in other cultures, such as in United States' middle-class families, who

train children to learn less by participation and observation, tend to make observations that are mainly descriptive and without thought to causes. In these cultures, parents or other adults often provide the explanations behind their observations, and consequently, children make "lots of observations but have trouble encoding evidence, making valid inferences, and connecting observation to theory" (Eberbach & Crowley, 2009, p. 47). They tend to see objects in isolation from larger systems in which they are embedded, missing complex connections. Finally, their observations are shaped by what they expect to see. They tend to notice primarily what is expected and miss what is not; they decide what observable characteristics might be important or not based on expectations rather than in spite of expectations. All of these focus children's noticing and meaning making in different directions than science experts'.

While much of the research reviewed by Eberbach and Crowley (2009) covers children's skills, and probably reflects a general dearth of literature concerning adult skills, the authors suggested that it is not simply a lack of maturity that frames children's observations. Adult novices similarly use more everyday observational practices rather than the systematic scientific observations and corresponding questions. For example, novice secondary school children and novice teachers still focus on structures or observable processes rather than causation for explanations (Hmelo-Silver, Marathe, & Liu, 2007). Eberbach and Crowley (2009) concluded that for novices, there are bigger problems than age and tendency to think concretely in general, namely lack of knowledge of the particular scientific discipline and the persistent tendency to overlook evidence that contradicts expectations.

These expert practices of seeing scientifically are learned through participation in cultures (Lave & Wenger, 1991), such as academic science, where observation is considered a primary learning or meaning-making tool. In fact, Goodwin (1994) says professional meaning making cannot be divorced from the social situation or the classification schemes and norms of discourse that

have been agreed upon by the practitioners. This meaning making must be learned by participation in the practice, especially in conversations about practice, in order to learn not only how to harness biological mechanisms such as vision to focus on relevant features, but also how to describe what is being seen in context in order to interpret the physical perceptions culturally (Eberbach & Crowley, 2009; Goodwin, 1994).

Even what constitutes relevant knowledge in science is constructed by the community of practicing scientists based on social and cultural scientific norms, the "*community's agreement* upon nature's correct answer" (Magnusson, Palincsar, & Templin, 2004, p. 133). This knowledge depends on the values, beliefs, and standards of the scientific community about "what is important to know and do" (Magnusson et al., 2004, p. 133). For example, conventions for visualizations have been developed within and agreed upon by the overall community of practice (Driver, 1995).

For making meaning in science, learning what is important comes through working as a scientist does, appropriating the tools and methods of science culture. However, one's individual experiences and knowledge bear on exactly how one appropriates that culture, so ultimately meaning must marry individual and biological background with social and cultural influences.

Though the expert has learned her skills primarily through cultural participation and thus uses her brain differently than a novice (Bransford et al., 2000; Chi, Feltovich, & Glaser, 1981; Ericsson et al., 1993; Kastens & Ishikawa, 2006), the expert and novice both encode skills in the same types of biological systems, namely neural networks. Cognitive psychologists have recognized for years the differences in mental representations in skilled performers versus novice learners (Ericsson & Lehmann, 1996; Gauthier & Tarr, 2002). Biologically, neural associations are shaped by cultural activity through the encoding of patterns of co-activation, increasing the likelihood that a particular perception in the future will also activate relevant background information for

comparison and context (Schacter, 1996). For example, chess masters extract more information from a briefly exposed chess position than novices and also encode information differently, as evidenced by their use of larger chunks of piece positions in a memory task (Chase & Simon, 1973). Skilled paleontologists find fossils where novices see plain dirt (Kastens & Ishikawa, 2006). Neural networks start out as many more connections than are ultimately needed; through a process of pruning those connections, only meaningful ones remain (Kandel, Schwartz, & Jessell, 2000). Novices must learn what is important in particular contexts in order to shape those connections into networks that have meaning; searching one's entire trove of knowledge to find pieces that are relevant quickly overwhelms working memory (G. A. Miller, 1956).

Thus, I argue biologically, meaning making is the building or refining of structures needed to connect and organize neural networks which store relationships and taxonomies (Bransford et al., 2000). This allows interpretation of new observations in light of these connections, through conjoined activation of neurons or networks associated with particular perceptions and experiences. Culturally, meaning making is understanding which connections of information are important to the discipline, on what features of the information at hand those connections are based, and how to use those connections to answer relevant problems (Chase & Simon, 1973; Cook et al., 2008; van Gog, 2006). Each influences the other; cultural understandings and participation shapes input to the brain, but prior experiences and knowledge contribute to the physical structures as well as the cultural participation, especially as one becomes a more integrated member of a community of practice and contributes more of one's own experiences to the culture.

Taken together, meaning making is a combination of the biological and the cultural: putting into appropriate context information that which is relevant to your life (based on prior experiences, as encoded in your biological systems) or purposes (based on cultural consensuses). This meaning making is the first step

in learning that new information. Scientific experts have different understandings of what is relevant to their questions about the world and what is appropriate context, and thus due to their training and expert-level attainment, they examine different information than novices do. Meaning-making in science, ultimately, is interpreting data collected with the tools and methods of science, such as observation, in terms of relevance to scientific questions but also based on one's own individual framework of science knowledge as stored in a physical network of neural connections.

Therefore, learning means forming and adopting a revised understanding of the phenomenon studied, and scientific knowledge is the scientific meaning of the observations, as analyzed in terms of prior scientific understanding, which is available only to those who have sufficient cultural training. However, the understanding and knowledge will hopefully be adopted by each individual for his or her own appropriation and use, as scientific inquiry in communities of practice provides motivation as well as cultural knowledge and experience with tools necessary to become more adept as a practitioner (Lave & Wenger, 1991). Therefore, we cannot discount the processes of individual constructivism (Driver, 1995) or cognitive or neuroscientific theories of learning and understanding of how the brain works, including memory storage and association of knowledge, emotions, and experience (Compton, Grossenbacher, Posner, & Tucker, 1991; Schacter, 1996).

How Experts Create and Use Visualizations

One of the tools scientists use most often to find patterns and convey those patterns to others is visualization. By creating graphs, charts, and the burgeoning forms of geographic information systems (GIS) visuals, scientists can harness our biological systems and incorporate cultural meanings to rapidly and efficiently process vast amounts of numerical results. However, one of the assumptions underlying this dissertation and other similar work is that imbuing

those products with content and conventions particular to academic science prevents a large portion of the everyday scientist population from being able to use them to make academic science meaning. Issues arise when academic scientists try to use those same visualizations for communicating with groups with different academic science training, such as sharing academic findings with everyday scientists in order to secure future funding, influence policy, or simply contribute to the broader knowledge base. Academic scientists often create visualizations with such complexity that everyday or novice scientists, who are unfamiliar with cultural conventions of trained professionals, do not make the same meaning or extract the same information from the visualizations as the experts do (Phipps & Rowe, 2010; Rowe et al., 2010; Rowe, Stofer, Bullick, & O'Brien, 2011). Thus, visualizations are used here as a case study of a suboptimal science-and-public communication tool in order to examine the disconnect between expert and novice understanding to allow more effective design of these communication tools.

Biology, culture, and the production of visualizations. Practicing scientists, "the scientific community," have created their own visualization culture, tacitly if not explicitly defining amongst themselves the conventions that get widespread use in the visualizations they produce, often imitating visuals they have seen published in journals by other scientists in their field (Light & Bartlein, 2004; Magnusson et al., 2004). This subculture and its tacit assumptions impact who can easily make meaning from these visualizations. Part of the problem is that one of those assumptions is the irrelevance of biological influences of vision on how untrained viewers outside of the community make meaning from a visualization. Perceptual salience of colors, perceptual illusions that distort sizes, association of colors with memories, and expectations of locations of given and new information in visualizations can all impact the way visualizations are read.

Visualizations are processed by a visual system driven both bottom-up by the eyes and arrangement of neurons, and top-down by the flexible nature of neuronal connections shaped by culture and more broadly, learning. Thus, visualizations must properly balance the use of both of these systems to optimize meaning-making. Biologically, "seeing" is the physical process of taking light in through the eyes and converting it to neural signals that convey meaning. The human visual system has been superficially well-understood for a number of years, based on clinical studies using deficits due to disease or lesions in the brain, and more recently, using functional magnetic resonance and other imaging techniques (Webvision, 2011). Research continues on the specific details of processes along the spectrum of visual activities: perception, attention, target acquisition, cognition, and the like. We know the basics of the perceptual process well: light enters the eye and hits the cells of the retina, the rods and cones, whose activation triggers release of chemicals that in turn cause electrical impulses in the optic nerve. The neurons carry the signal to the primary visual cortex of the brain, where the patterns of light, which have been translated to cellular activity, are interpreted to be what we see. This "interpreting" part of the process remains an area of active research.

Geographers and others researching and developing geo-visualizations understand that elucidating those areas of human perception, cognition and visual processing that impact meaning making from visual communication tools such as maps and data-filled graphics is a key challenge in order to move the discipline forward (MacEachren & Kraak, 2001). For example, the rainbow color scheme favored by many scientific visualizers represents middle values in the data in yellow, but yellow is perceptually the most vibrant to the human eye, so the middle range data often ends up “standing out” when it is not statistically the most important. This color scheme is also perceived quite differently by color blind individuals – both ends of the spectrum, red and blue, end up the same color to people lacking red-sensitive cones in their retinas (Light & Bartlein,

2004). Also, due to processing characteristics of the human visual system, depiction of a spread of colors might imply to the layperson that equivalent numerical changes exist where they do not, for example if shades of blue are thought to be more similar than shades of red (Slocum et al., 2001). Color can trigger memories or change perceptions by making areas seem larger or smaller than they truly are (Bevlin, 1977). In addition, Western readers look for new information at the bottom right part of an image (Graham, 2002), not necessarily where it might be located in a visualization, and information is retained far better when it is presented as complete thoughts, rather than the typical few-word titles or phrases that may accompany an visualization (Alley, Schreiber, Ramsdell, & Muffo, 2006). All of these perceptual characteristics can either afford or constrain meaning making.

However, culture plays an equally significant role in defining what one "sees." Whole societies can differ in the ways they think about the world, and that in turn influences what they attend to and how they reason (Nisbett, Peng, Choi, & Norenzayan, 2001). Nisbett, et al., argued,

> the considerable social differences that exist among different cultures affect not only their beliefs about specific aspects of the world but also (a) their naive metaphysical systems at a deep level, (b) their tacit epistemologies, and (c) even the nature of their cognitive processes—the ways by which they know the world. (2001, p. 291)

Here, the authors studied specifically ancient Greek (Western) and Chinese philosophies, as well as the differences that persist between the contemporary versions of these cultures in their interconnected ways of noticing, thought and reasoning.

Biological and cultural differences between expert and novice scientists extend specifically to using visualizations. Cook, Carter and Wiebe (2008) found that students with different levels of prior knowledge understood different parts of a biological cellular transport visualization to be important. Students that had

higher prior knowledge used the scale bar more and made better interpretations of the figure than lower prior knowledge users.

Models and meaning-making from visualizations. Visualizations are models, that is, visual representations of data. Visualization users must, therefore, understand principles of models and their construction in order to understand the visualizations and their construction, even if they may not necessarily explicitly connect the idea of model with visualizations. Therefore, I consider what models are and which of their properties are important to making meaning from them.

Instead of tackling what a model *is*, much of the recent philosophy of science discussion centers around models as they relate to theory, experiment, and the like. The majority of philosophers of science hold the semantic view, that models are the components of theories (e.g., Giere, 1999, 2010). A competing idea is that of models-as-mediators (e.g., Morrison & Morgan, 1999). The supporters of this view argue models are used to experiment on and manipulate theories, but are not themselves theories. I shall show that both of the philosophical stances relate to a characteristic of a feature necessary for understanding the representational function of models, but neither view alone addresses all the necessary features.

The semantic view holds that it is a family of models that are exactly a theory describing a target system, and that the models are (mathematical) structures that represent the theory (da Costa & French, 2000). Families of models constitute the complete understanding at the time. The proponents of the "semantic view" or "model-based view" of theories believe that models are themselves the stuff of theories, that theories are exactly made up of a group of models or varying ideas about a system or process (Giere, 1999, 2010). The semantic view tells us about what our current understanding of the particular

scientific field is, using models as *what* science knows, or thinks it knows at this point.

The main alternative philosophy sees models as “mediators” between theories and phenomena. Hacking said "the models are doubly models” (1983, p. 216), connecting the theories and the phenomena themselves, taking some aspects of both and re-elevating the role of the experiment in science. In his view, models fundamentally allow intervention in science and the development of knowledge through experimentation, “constructing models that human minds and known computational techniques can operate” (1983, p. 217). Morrison and Morgan (1999) consider models explicitly to be tools for intervention, calling them “instruments of investigation.” This alternative view tells us that modeling is *how* scientists know what they know, that is, which representations fit both the real- and hypothesized-worlds and thus “prove” theories.

Visualizations of experimental or modeled data are a special case of scientific models, and it is in terms of this relationship that the disconnect with the non-scientist public goes a layer deeper. They do not represent physical systems or theories themselves nor are they necessarily the tools for doing science, but rather they show the *outputs* of science, a “models-of-data” view (Frigg & Hartmann, 2006; Giere, 2010). The data can be the result of experiments on theories, as when one computer simulation model of a process such as cloud formation is run. Alternatively, the data can simply show observations of where clouds are on Earth at a given time, without any sort of information as to what the underlying processes of cloud formation and movement might entail. Neither the semantic or the models-as-mediators philosophical views of scientific models truly deals with this type of visualization, as philosophers are concerned with models that represent theories or experiments. To that extent, literature from philosophy of science and formal science education is of limited use for designing learning experiences whereby scientists can share current research findings with the public by means of data visualizations.

Modeling data, too, builds representations based on scientific cultural norms with which non-specialists are unfamiliar. To truly understand visualizations, the public must understand three features of models embedded within the visualizations: 1) correspondence to the real-world 2) multiplicity of layers of abstraction and 3) purpose of representation. First, one must know that the visualization is a representation of real data rather than a fictional illustration. This means that one understands that the depiction of real data in this form is possible, that it is indeed real data that is being represented. Next, one must understand that there may be multiple layers of representation of or abstraction from the real-world data, that is, the visualization, itself a model, could be representing the output of another model. Finally, viewers must understand the purpose of the representation, that is, why the visualizer is representing the data this particular way.

Understanding the purpose breaks down into three key characteristics of the visualization that one must ascertain: a) the features of the depiction and what they represent, b) the source of the data and methods for obtaining it, c) the data that is displayed and that is left out, or how the representation itself was made. These first two characteristics, a and b, are somewhat related to the views of models in the philosophy of science and science education literature, especially around reform based on emphasis on the nature of science (Develaki, 2007; Justi & Gilbert, 2000; Portides, 2007). Much of this capacity for understanding these three features of models for meaning-making from visualization relies on having significant scientific expertise due to the way these visualizations are currently produced and the conventions they contain.

These bodies of literature and their gaps reveal an incomplete picture of what the public currently understands about how models of data work. Neither the semantic view nor the models-as-mediators view of model-theory relations has been taken to be exactly representative of what scientists do. In fact, as noted earlier, it is the problems with the terms ‘model’ and ‘theory’ that contribute

directly to this insufficiency (Chakravartty, 2010). However, from theoretical and empirical works on these topics we get a sense of what skills the public has at understanding at least two of the characteristics of one of the three features of models necessary for comprehension of those models. The education and philosophy literature helps fill in the holes in research around the other features and more directly related to understanding data visualizations specifically.

Models of data, though, rather than models of systems, are more commonly presented for communication directly to the public as visualizations of results. These models may present fewer terminology complications at this time, as they tend to use pieces of all the purported functions or definitions of models rather than trying to narrow the definition of model to theory or experiment alone. However, these additional definitions have less research associated with them at this point.

Despite scant literature on current skill levels related to some particular features necessary for model comprehension (Grosslight, Unger, Jay, & Smith, 1991; Treagust, Chittleborough, & Mamiala, 2002), there are features that must be understood by anyone wishing to make meaning from a visualization, whether or not they explicitly recognize the connection to modeling. First, the viewer must understand that the visualization is depicting real data. Real data brings with it affordances and constraints that shape how the depiction is constructed. Visualizing real data should imply consistency in the conventions and symbols used to depict features and ranges of the data from one instance to the next for comparison. Visualizing real data also means that only certain values are possible. A visualization that claims to be real data yet shows the daytime temperature of Houston, Texas, to be 200 degrees Fahrenheit is not possible, so a viewer should then think that there is a mistake either in the representation or in the claim that it is real data. Some previous work in evaluation of museum exhibits displaying visualizations of suggests visitors do not have trouble assuming the data to be from the real world (Rowe et al., 2010).

Once the viewer understands that the depiction is modeling the real world in some manner, the viewer must be able to recognize that modeling doesn't occur at just one stage in the process of visualizing data. This, perhaps, is one of the biggest unknowns about public understanding of models at this point. While we know that students do understand that there can be multiple models of the same process or system (Grosslight et al., 1991), little research exists on what people might know of models within models. Our evaluation of an exhibit kiosk with visualizations from both actual satellite observation data and computer-simulated data showed that people did not easily recognize the simulation data as such, instead tending to believe it was real-world data (Rowe et al., 2010). A visual representation of a computer simulation is at least a model (picture) of a model (simulation). Depending on the complexity of the simulation, it might itself be a compound model composed of sub-models plus actual data.

Finally, recognizing the purpose of modeling in any given visualization is vital to making meaning from a visual depiction of data. Giere (2010) adds the roles of agents and their intentions into modeling, giving the 'formula': "agents (1) intend; (2) to use model, M; (3) to represent a part of the world, W; (4) for some purpose, P" (Giere, 2010, p. 269). This is, he claims, how models represent something in the world, and it must be invoked. Thus, one must understand what relationship the agent is invoking in order to understand the visualization.

To describe purposes or functions of these models, Chakravartty (2010) dispensed with the problematic idea of modeling and focused on representation in scientific practice, including in his use models, theories, diagrams and simulations among others as types of representations. He puts forth two important views of the nature of these representations: informational and functional, arguing that rather than conflicting, the two views are complementary. The informational purpose of a representation serves to convey details and features of the thing it represents, the target system. The functional purpose allows cognition about the target systems through their representations.

In the case of visualizations of satellite data, understanding these two potential functions, akin to Giere's (2010) "purposes," is crucial. A visualization conveying information might show the location of clouds on Earth as a satellite visualization of global cloud cover. For a more functional use, the public could look at that visualization to make judgments about the normalcy of the locations of the clouds, that is, whether the clouds seem to be typical of a "normal" day on Earth or a more unusual event.

But just how do people recognize the purpose? To do so, viewers of visualizations must assess three characteristics: a) the features of the depiction and what they represent, b) the source of the data and methods for obtaining it, c) what is displayed and what is left out, or how the representation itself was made. As previously discussed, characteristic 3a is akin to the "what" science knows of the semantic philosophy, and characteristic 3b is akin to the "how" science knows of the models-as-mediators philosophy.

The first step in recognizing the purpose is figuring out the "what" of science knowledge that the creator of the visualization is representing through her choices of colors and patterns. One big stumbling block the public currently faces at this step is essentially a "language barrier." Often, the relation of the image to the real world being depicted can only be understood by means of representational conventions (Giere, 2010). Giere (2010) argued that the intention of the representing agent is often obscured by the use of conventions, and the familiarity of these conventions as well as the level of background knowledge of a model user can weigh on the effectiveness of the model's representation. Scientists possess knowledge of a set of conventions standard to science. Therefore, the representations they share with the public must either contain information translating those conventions or use non-scientific conventions that are more broadly available to the public. "It is crucial to enable the [representation user] to interpret the symbols relevant to her, e.g. by using familiar conventions" (Bailer-Jones, 2003, p. 67). Empirically, this was

demonstrated in science center visitors who did not understand scientific “rainbow” color scales as easily as they did red-blue depictions of temperatures or green shades of chlorophyll (Phipps & Rowe, 2010).

In addition, in order to use representations in the functional sense, users must understand characteristic 3b, how the representations have been constructed (Chakravartty, 2010). What has been left out or simplified? What has been highlighted? This, too, is highly dependent on background knowledge of the system being represented. In viewing global visualizations of clouds on a flat map, one must have some knowledge of the distortion that goes into flattening the almost-spherical Earth in order to understand which geographic areas are misrepresented and why that is an acceptable trade-off in terms of viewing cloud patterns. As Bailer-Jones (2003) notes, makers and users of representations are trained to look at features that are being represented, (in my example, cloud locations), and not worry about those that are not (exact geographical distances). Thus, a representation of cloud patterns does not claim to represent true cloud area or height. Depicting these additional features would require adding information such as color-coding that can take the place of three-dimensional volume. Also, some satellites only collect data over a certain portion of the globe, so in the visualization there could be areas either of "no data" or interpolated data.

This last example leads us to characteristic 3c, understanding how the data was obtained. In the case of real-world (non-modeled) data, the data that is gathered is itself a model of the real situation. That means observations of cloud heights, temperatures, and analog geographies are translated into digital bits contained in the on-board computers of the satellite. This way of gathering data is a modeling process that is similar to one many users are already familiar with: the image capture of a digital camera. To an extent this is dealing with abstractions and assumptions made in data *collection.* While understanding collection is essential to understanding the scientific process, it is not directly

germane to the *representation* of the data. In that sense, unless one is interested in the operation of the satellite, that detail of collection can be irrelevant in some cases, as long as the source of the data is known to the user.

However, in different kinds of data collection, such as satellite capture of sea surface temperature, users need to know more. They need to know how scientists gathered the data and what additional layers of simplifications, highlighting, and omissions visualizers made in representing that data visually. For example, visitors should be aware that data is not captured for every single point on the globe and therefore, in the visualization the satellite passes may be "filled in" with interpolated data that is itself based on a statistical model, perhaps of how temperature changes with distance. Understanding at the very least the field-of-view of different satellites is essential.

Again, two specific examples illustrate this need for understanding data collection in terms of purpose of representation in each of the two different uses of models. In gathering information from a weather forecast model, if users do not know whether "30% chance of rain" means it will rain 3 out of 10 times, it will rain over 30% of the forecast area or it will rain for 30% of the forecast time, then they will also be likely to misunderstand the meaning the meteorologists wish to convey (Gigerenzer, Hertwig, Van Den Broek, Fasolo, & Katsikopoulos, 2005). If they know that the forecast is based on historical data where, under these same conditions, it rains three out of 10 times, the public is much more likely to understand the forecast correctly. For the functional use, knowledge of collection methods would also be necessary for a representation of data that is the output of a model, for example a prediction of air temperature in the year 2050 using a particular climate model with particular inputs to that model. Recognizing the difference in purposes between representing the "worst-case" and "best-case" scenarios demands recognizing the input to the forecast models.

These features of visualization construction that are based on the modeling of data clearly must be understood for a user to make scientific

meaning from the visualization. However, whether users must explicitly recognize the visualizations as models themselves remains an open question. Resolution of the latter is beyond the scope of this dissertation.

Modeling Expert Thinking

Learning academic science in formal schooling. Much public and private funding has been put towards the "problem" of education in the United States. Students are "failing" to obtain minimal skills (at least by formal measures). Testing has been implemented to provide benchmarks for proving that students are or are not attaining goals and whether teachers are or are not effective at helping students attain those skills and prepare for the workforce. In science education, reforms are sweeping the United States in trying to put knowledge in real-world context, add inquiry for skills and prepare students to be strong domestic science and technical workers and global citizens.

Beyond the initial development of science fluency as a youth, the pace of research development means that new information is virtually flooding the market daily throughout one's lifetime. Researchers can now go straight to things like social media to publish their work, even in the early, non-peer-reviewed stages. But how do real-world users know a) what to access, b) how to evaluate it and c) where and when to apply current research for their own ends? How can users acquire the capability to use academic scientific tools in a anticipatory, just in time, and appropriate (Roth, Lawless, & Masciotra, 2001) manner that indicates fluency and transfer, that is, enacting science fluently in other contexts (Tobin, 2005).

No scientific curriculum is complete without some level of knowledge of the current understandings of the way the world works (National Research Council (U.S.), 1996; Project 2061 (American Association for the Advancement of Science), 1993). Given that models are considered by the majority of

philosophers of science as representations of “what” science knows at the moment, it is not surprising that most of the recent theoretical science education literature focuses on the semantic view of models and how this function is or isn’t understood by students. That awareness is still a necessary part of understanding models and the components depicted in them.

Recent advocates of reform emphasizing nature of science education all focus on the semantic view of models. Particular recommendations include teaching model-building using idealization and approximation (Portides, 2007) or teaching the development of and change in models (and thus theories) over time (Develaki, 2007; Justi & Gilbert, 2000). Teaching how scientists developed the understanding of the structure of the atom through successive scientific models, rather than "hybridizing" earlier models, provides not only the facts but also an understanding of the scientific process. “Science education sees in this new view, and primarily in the cognitive approach of Giere, a satisfactory epistemological foundation for an understanding of the nature of science and its didactic transformation, for an innovative planning of analytical programmes” (Develaki, 2007, p. 728).

Empirical work in science education not only emphasizes the view of theories as models, but also underscores schoolchildren’s the lack of understanding of the nature of models. Students generally do not yet understand the relation of scientific models to theories and thus their use in the practice of science or development of scientific knowledge (Grosslight et al., 1991; Treagust et al., 2002). Students who had received no training in scientific models felt that models were necessary and useful. They had some understanding that models changed as scientific knowledge changed, but many felt that models were exact replicas and failed to recognize the use of models to make predictions and test ideas.

Some specific features of models that the public should understand were examined in studies with school children. Grosslight, et al., (1991) set forth five

particular features of models that they tested with students: kinds of models, purpose of models, designing and creating models, changing models, and multiple models. Across the board, students differed from experts in their understanding of these features. The experts viewed models more abstractly, as used for "actively formulating and testing ideas about reality" (Grosslight et al., 1991, p. 816), whereas students discussed a more concrete relationship between models and reality. A more broad curriculum on models and their various, pervasive uses in science would more realistically represent the scientific research process as a whole, not just the end-stage knowledge development.

Understanding models through thus revamped school programs at the least would work to build skills in understanding the component pieces of models, one characteristic of a necessary feature (3a) for overall understanding of data models. This focus on the theory-driven development of knowledge via models, however, leaves out the crucial data collection and representation functions of models that are fundamental to understanding visual depictions of scientific output. In this way, school programs are currently not addressing scientific understanding of modeling as it relates to working with data.

Communicating academic science through informal learning environments. Much of life is actually spent outside the classroom. In the United States, formal primary and secondary school covers less than the first quarter of an average American's 80-year lifespan, and typically no more than one-third of waking hours in that first quarter (Falk & Dierking, 2010). Many life decisions happen after formal schooling, so how do we prepare for life beyond school five days a week, with messy questions, incomplete evidence and ever-changing parameters let alone ever-changing targets?

Informal science education institutions such as museums have offered various forms of exhibits and programs to try and offer current academic science findings to the public. Multimedia exhibits with frequently changing content, staff

presentations, and “science pubs” with scientists offering lectures for general audiences, ostensibly with discussion, have been cropping up around the world (Dallas, 2006; Pedretti, 2004). A recent version combines these three as scientists go into the museum to talk with the public about their research in the Pacific Science Center’s Portal to the Public program (Selvakumar & Storksdieck, 2013).

All of these informal educational settings rely on the use of visualizations: to attract users to an exhibit, to start conversations in a science pub, or to actually be the centerpiece of the information dissemination as in the case of spherical and other displays of scientific data. Over the last 10 years, especially, there has been an explosion of museum-based and other publically available installations based around imagery. Educators, science advocates, and the public alike have encouraged this rapid deployment in order to get content into the hands of the public and harness the enormous potential of visualizations of real data (Haley Goldman et al., 2010; Nelson & Ellenbogen, 2006; Nelson, 2006). However, this deployment has often outpaced the capacity of the institutions where they are deployed and the educators in those institutions to make use of them optimally.

Typically, hands-on science centers in particular have built exhibitions around subject matter rather than science process skills (Karp & Leblang, 2004), so that modeling, for example, is rarely explicitly considered outside of the context of one or two interactive exhibit pieces, perhaps. Compounding this, many of these settings simply repurpose visualizations provided by academic scientists who created them for presentations to other academics, not to everyday scientist audiences (Rowe et al., 2010).

Scaffolding as a form of modeling expert thinking. The fundamental problem for everyday scientists in making meaning from visualizations created by academic scientists is that those unfamiliar academic visualizations are outside

the everyday scientists' "zone of proximal development." Vygotsky (1978) introduced this term to describe the level of a problem that is above what a learner can solve independently, but just above, and not far enough above the current skill level that the learner is unable to solve the problem with assistance. The academic scientist has progressed several levels above the everyday scientist in making academic meaning from visualized data, and making meaning from the visualizations academic scientists create presents the task at a level that is generally too far above the everyday scientist's current skill set to solve on his own (Phipps & Rowe, 2010). It is the role of the educator or facilitator to help the learner find sub-tasks that are accomplishable within the larger task and then move on to another sub-task stepwise until the learner solves the task, ultimately recognizes the progression of sub-tasks and can do the entire task independently (Vygotsky, 1978).

To make a too-difficult task manageable for a learner, facilitators use scaffolding, which describes any number of types of assistance to a learner to allow the learner to complete a task that would otherwise be beyond his or her current capability. Scaffolding involves more than simple guided assistance or modeling and imitation (Wood, Bruner, & Ross, 1976). Instead, scaffolding provides direct support for the learner to complete elements of the tasks that he or she could not complete alone, allowing him or her to complete the task. This allows the learner to reach and begin to recognize solutions to the problem, a type of performance before competence (Vygotsky, 1978), which he or she must do before being able to solve the original unscaffolded problem independently. Mastery of the appropriately-challenging material is scaffolded for the learner by a "more knowledgeable other" (Vygotsky, 1978), who guides the learner in making meaning from the novel information by offering supporting tools like reasoning routines or providing background information and taking over when the task becomes too difficult.

A facilitator adds scaffolding to and ultimately removes it from a task in order to advance the learner's skill step-by-step from apprentice to master understanding. Instead of the learner being immediately expected to jump from his or her current understanding to that of the master, the learner moves in increments. A facilitator creates those incremental steps by taking the master's former learning path or complete current knowledge state and breaking it down into elements to be individually appropriated. Each element or step is supported with scaffolds, which focus the task to be learned on a smaller part of the task that is within the grasp of the apprentice. Scaffolds offer the learner a place to start the task and first become proficient at a piece of the overall task. Then the learner can move on to a slightly larger or more advanced part of the task, become skilled at that part and move on, and so on until the whole task is conceptualized by the learner independently. Humans, that is, *more-knowledgeable others* (Vygotsky, 1978), may scaffold performance by offering conversation or question cues to guide the learner or simply answers to questions the learner has. Scaffolds may also be internal to the task (Wood et al., 1976). For example, a math apprentice is not expected immediately to solve an algebraic equation for a variable. Instead, they are first focused by the teacher on combining like terms to simplify the expressions on each side of the equation and, so the original problem given (simplifying the expression) is a small but integral part of the ultimate algebraic task. Then learners are progressed gradually as they master smaller steps, beginning to move added and subtracted terms from one side to another, then learning to divide and multiply with variables until finally the learner understands how to isolate a single variable through a series of steps.

Modeling expert visualization use. In many cases, academic science visualizations are unaltered when provided to or obtained by formal and informal education institutions. In addition, many of these academic visualizations lack

extensive information about how they were created, so that formal educators and informal education facilitators alike have to go to great lengths to uncover that information, if it even occurs to them to share that with visitors. Educators and facilitators also often lack the resources or permission to change the visualizations. Thus, the visualizations shared in these environments often include elements that directly compete with general human perception and non-scientific culture, slowing down or even completely blocking the communication that the scientists and educators intend (Phipps & Rowe, 2010; Rowe, Stofer, & Barthel, in preparation; Rowe et al., 2010).

Functional brain imaging using both Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI) suggests that experts automate processing of visual objects and thus use different brain regions from novices for these tasks (Bukach, Gauthier, & Tarr, 2006; Tarr & Gauthier, 2000). The less time and energy users expend comprehending the basic features of what is depicted and where it is, the more effort they can spend recognizing patterns in the data that illustrate processes and evaluating those processes in a broader context (Ware, 2004).

In particular for visualizations, conventions of color are based in cultural understandings and vary among different cultural groups, bringing potentially different intuitions about the meaning of those colors in visualizations. Lucy (2011) updated the Whorfian hypothesis that language influences cognition; in some cultures, words for particular colors do not exist; in some languages, no terms for color exist. Therefore, people from those cultures often don't automatically perceive differences between colors they don't discriminate verbally. These differences in visible spectrum parsing could conflict with the meaning intended by the developer, making the developer's meaning less accessible to those with a different cultural or linguistic understanding of the colors used. If extreme enough, this conflict could make the intended meaning unintelligible to the user if all the information is encoded only in color, defeating

the purpose of the tool. These influences of design elements may or may not play a role in small-scale elements such as visualizations, especially for English-speaking participants, but should not be overlooked without analysis.

Scaffolding experimental visualizations for broader meaning-making. To investigate the ways expert academic scientists use visualization and how that could be modeled for everyday scientists, I first removed the potential perceptual and cultural conflicts from the academic versions of the visualizations. One could argue that this is instead *shaping* or changing the task itself; however, even Vygotsky referred to his interventions by both names. Both shaping and scaffolding are used by educators regularly to support learners in mastering tasks (Greenfield, 1999). I am not interested in the behavioral origins of shaping but instead with the sociocultural origins of scaffolding, so for simplicity, I will refer to the interventions here as scaffolding. Thus, I embedded scaffolding in the experimental task itself instead of using a more-knowledgeable other to more closely resemble a naturalistic visualization interpretation task. The scaffolds took the form of replacements for higher-level cultural-specific knowledge that confounds the bigger task of academic meaning making from these visualizations for the everyday scientist.

Phipps and Rowe (2010) found that several features of visualizations used in scientific journals could be altered to better enable novices to make academic meaning from the scientific findings expressed in the visualizations. Changing conventional academic scientific color scales, typically depicting data with a rainbow spectrum, to more culturally relevant colors such as reds and blues for temperature, vastly improved comprehension of the overall meaning for both formal educators and science center visitors. Furthermore, adding geographic labels, borders, and legends based on the ways users naturally read or scan pages make the areas represented in a visualization more immediately recognizable.

Based on this earlier research, the scaffolds for this experiment were designed to break down three elements of the task, in order to empirically determine 1) which, if any, scaffold is most effective at immediately moving novice, everyday scientist visualization readers toward academic expert visualization reading, 2) which elements of the task or data and scientific meaning ostensibly being conveyed are still alternatively conceived by novices even with scaffolding in place, and 3) how participants use various scaffolds, and 4) how that use differs based on expertise. See Appendix 1 for example stimulus visualizations.

First, the colors chosen by the academic scientist visualization users often convey more specific detail about certain situations than the novice user needs to learn the basics of visualization “reading”. For example, a rainbow color scale used to demonstrate a binary result uses three hues to depict each possible result when different shades of only one hue each is necessary. That is, red, yellow, and orange depict above-average temperatures and green, blue, and purple depict below-average temperatures in an unscaffolded academic visualization of sea surface temperature anomaly. A scaffolded version would use shades of red for above average and shades of blue for below average. The six-hue rainbow color scale at the very least muddles the picture and at worst creates artificial distinctions that do not naturally exist in the data, as the human visual system responds more strongly to edges than homogeneous areas (Duchowski, 2007). The rainbow color scales also do not draw on cultural expectations of the population, and they compete with natural perceptual processes. In the United States, the middle value of the rainbow, green-yellow, does not have any cultural associations with temperature, for example, and the extremes, red and blue, do not have any cultural associations with plants (Conroy, 1998). Perceptually, the human eye is most drawn to a green-yellow hue (Faughn & Serway, 2003), again portrayed as the middle value of the rainbow scale, rather than extremes, which are generally most interesting

statistically and for meaning-making. Furthermore, rainbow color scales do not smoothly transition in absolute color values (hue, saturation, and brightness) across data values (Brewer, Hatchard, & Harrower, 2003; Light & Bartlein, 2004). For non-experts, this could lead to artificial color delineations in an visualization that have no statistical meaning, which experts have learned through experience to ignore. That is, if the transition from green to yellow happens at a temperature value that is meaningless in the data, the non-expert might not realize there is no significance to the temperature change at that point. In the experiments in this dissertation, the chosen color scales increase hue, saturation and brightness in equal increments matched to data value increments (Brewer et al., 2003), which also prevents them from competing with natural human visual perception (Faughn & Serway, 2003; Light & Bartlein, 2004). The color scales also test hues that were thought to have cultural relevance to the novice population, namely, everyday science visualization readers in the United States.

Second, the titles of the visualizations often used by experts contain scientific jargon such as *chlorophyll*, a pigment found in green plants, or abbreviations such as “SST” to refer to “Sea Surface Temperature.” These would be translated in an article or report aimed at a general audience, so I also translated them in these experiments. Measurement units such as Celsius when provided in figure legends also tend to be abbreviated and scientific as opposed to the more common use of customary units such as Fahrenheit in the United States. In these experiments, Fahrenheit values were provided as scaffolding (though the label was abbreviated “ºF” to accompany the “ºC” of the unscaffolded versions) for the temperature visualizations. In the chlorophyll visualizations, labels “very high” to “very low” were added for the scientific measurement abbreviation for concentration, mg/m^3. Finally, in the temperature anomaly visualizations, Fahrenheit values and “higher,” “average” and “lower” labels were added to Celsius values (with unit labels abbreviated) in scaffolded visualizations.

Third, I added geographic labels for continents, ocean basins, and the equator, to remove an element of the task that is less familiar and less automatic for novice visualization readers, namely geographic orientation on a global scale. Here, the geographic orientation task was further complicated for novice users by presenting an visualization centered on the Pacific rather than the Atlantic Ocean basin. Users of these global visualizations presented on spherical displays in science centers often spend time puzzling over or simply identifying geographic areas (Phipps & Rowe, 2010; Rowe et al., 2010). Provision of these labels allows the user to have confirmation of orientation that they may speculate on or otherwise dwell on to the point of distraction from or inability to complete the ultimate task of constructing scientific meaning. On regional visualizations, users were often unable to recognize the Pacific Northwest coast of North America without political boundaries displayed (Phipps & Rowe, 2010). On a global visualization, if a user gets hung up on where the exact placement of the equator is, she or he may be unable to accurately understand the precise location of upwelling in relation, or of temperature anomalies, or simply of temperature distributions that vary during the seasons of the year. If the user is unaccustomed to the presentation of the Pacific Ocean basin in the center of the visualization, she or he may be unable to even recognize the visualization as one of the Earth.

Thus, here, color scaffolding allowed users to focus on the extreme values and overall range of data in visualizations. Familiar titles and units added verbal cues to the visualization data. Geographic labels allowed users to focus on other, more relevant, areas of the visualization content in order to make meaning from the visualizations. This project seeks to explore how and why those work as well as investigating the various levels of scaffolding to shed light on which interventions or combinations prove most useful, or which elements of the expert-level task are more difficult than others. This insight could lead to better understanding of various levels in the task of constructing meaning from data visualization and could help identify additional levels of difficulty. Understanding

expert performance can point to steps and levels necessary for non-scientists to ultimately be able to make meaning from unscaffolded visualizations as well.

The ultimate aim of scaffolding is for it to be gradually removed as the learner becomes more expert at a task. In the present work, for example, as the learner grasps the jargon and learns scientific measurement units, the next step up could be an visualization reading task that does not scaffold titles and units, but instead only scaffolds colors and geographic labels. Then, in turn, as a learner becomes more facile at these other elements of the tasks, they may also be removed or progressively be made more like those comprehensible to expert visualization readers. Thus, this process of scaffolding and then gradually removing scaffolds could permit a user to tackle a task which they might have previously walked away from entirely when the cognitive load was insurmountable. Then other scaffolds that might be needed by a non-expert could be revealed when they begin to tackle a task. This could be a sign of increasing expertise, when a user is able to use what they know effectively and articulate more about what they don't know.

Bridging Multiple Disciplines

Marrying the influences of culture and biology in understanding visualization meaning-making requires drawing on multiple, to date disparate, research disciplines, each with their own norms and cultures. To bring these disciplines to bear on a third discipline, education, introduces even more variables. Many neuroscientists have claimed to produce results that can inform educational practice, but few have evinced practical ways of doing so. As Bruer (2006) points out, cellular and molecular models of learning and memory via long-term potentiation (LTP), or excitation, of neurons not only do not always show a causal link between LTP and memory, but also remain at a much different level of analysis than cognitive psychological and other social science models of learning. What work has been done at a macro or functional level more

akin to social science studies, functional magnetic resonance imaging (fMRI) of the brain also concentrated in its infancy on compartmentalizing the brain to map its fundamental functional areas (Bruer, 2006). In order to provide meaningful insight into learning, these studies need to examine associations of structures to characterize the higher-order processes that make up learning.

Increasingly, neuroscientists are examining more higher-order tasks, which involve associations between multiple functional areas as systems and a time course of activations of different areas (Abdullaev & Posner, 1997; Compton et al., 1991). This associative theory is allowing investigation of increasingly more complicated tasks. The use of functional imaging and its study of brain activation patterns could begin to refine cognitive models, though these studies that go beyond localization of activity are still “complex, subtle, and rare” (Bruer, 2006).

Part of the lack of meaningful fMRI results so far may be due to a dearth of appropriate real-world behavioral tasks that could be tested both in the sociology or psychological realm and in the neuroscience realm (Lobben, 2007). Lobben and collaborators have begun to investigate fMRI as a behavioral geography testing method, correlating findings from the brain images with more traditional behavioral and spatial testing methods by using real-world tasks tested with each discipline's techniques (Lobben, Olson, & Huang, 2005). Other collaborations between social scientists and neuroscientists and fields from gaming to economics are beginning to incorporate fMRI techniques (Lobben, Lawrence, & Olson, 2009). Gläscher and colleagues investigate voxel-based lesion system mapping compared to a psychological task (Gläscher et al., 2010). Even decision making has become a task operationalized for neuroscientific study (Grecucci, Giorgetta, van’t Wout, Bonini, & Sanfey, 2012). Investigation of deficit areas of the brain is another well-established neuroscience technique (Kandel et al., 2000). Thus, it may be that the fields are truly ready to establish crossover and help to validate and inform each other's techniques.

Much of sociocultural and indeed, education research in general, relies on self-reports of meaning, understanding, and personal characteristics such as identity. Self-reports may be, for many reasons, of questionable accuracy and reliability. In particular, self-reports may differ from mental functioning around visualization use; both naïve users and meteorologists preferred more complex, realistic graphics, despite lower performance on spatial tasks associated with the complex visuals (Hegarty, Smallman, Stull, & Canham, 2009). Thus, potential conflict between self-reports and self-understanding might be elucidated through physiological results. Coupling physiological measures with self-report then might tease apart group differences in construction of meaning.

Another area of understanding that the use of these neuroscience methods could shed light on is differences in task performance. Eye-tracking with high- and low-performers with visually-based tasks has revealed difference in patterns of eye movement with different task performance (Grant & Spivey, 2003; van Gog, 2006). Olson, Lobben, and Huang (2005) have started investigating the use of fMRI for studying task performance around map use.

Physiology-based neurobehavioral results could eventually confirm participation in a new, blended, “scientific thinker” community of practice by confirming similar patterns of brain activity and other physiological measurements among those trained extensively in science and those not when using supposedly socioculturally-scaffolded tools. Neuroscience techniques could confirm brain activity in areas particular to recognition as a correlation to understanding and eventually lead to correlations for characteristics such as identity with activity in areas involving ideas of self. Newberg and collaborators have conducted the beginnings of this sort of exploration of self, using a related brain imaging technique (Newberg et al., 2001). They found that while meditating, Buddhist monks tend to have decreased activity in areas of the brain thought to be associated with a sense of self, at the same time that the monks report entering deeper meditation states where they disconnect from a sense of

self (Newberg et al., 2001). Other researchers have noted differences between expert and novice meditators using fMRI specifically (Davidson, 2003), suggesting that perhaps these techniques can be used for observing such changes in identity as becoming a scientific thinker as membership in a community of practice changes.

By ascertaining what, if any, differences in attention, place recognition, spatial visual, or other areas of the brain exist between experts and novices through the use of eye-tracking or other neuroscience tasks such as fMRI, we can potentially add to our understanding of what parts of visualizations are salient, confusing, or otherwise confounding expert use of visualizations. This could lead to design of more scaffolds for use with visualizations as well as add a potential tool for concurrent if not construct validation of self-report methods concerning recognition and use of mental tools or strategies.

Methods

This chapter will detail the methods and procedures used in this dissertation. I start with the conceptual framework for the study, followed by the overarching research plan. The procedures which follow detail the sampling, stimuli preparation, physical setting setup, pilot testing and data collection. I outline the quantitative and qualitative analysis I performed, and finish with a discussion of the triangulation. The final section considers limitations of the methods and procedures.

Conceptual Framework

This study draws on more than one tradition in the social sciences for understanding meaning making, learning, and science literacy. As a result, instead of relying on one tradition of theories or methodologies to elucidate meaning making from scientific visualizations, I will be engaging in a type of bricolage. Bricolage, a term primarily associated with Levi-Strauss (1966), describes the process of bringing together what otherwise might be considered disparate or even non-compatible materials in one artistic or cultural production. Originally a term used primarily by anthropologists and sociologists, the notion of bricolage has, in recent years, been adopted by social scientists working in a variety of fields to support complex, interdisciplinary work. In this case, my bricolage will bring together components of several epistemologies into a new combination to explain our relation to the world, a strategy suggested by a growing number of methodologists (Kincheloe & Berry, 2004; Patton, 2002). Specifically, in this project, I will draw on specific theories and methods usually associated with (social) constructivism, socioculturalism, and neuroscience. Using the same types of visualizations across methods within these three paradigms allows triangulation of the results and elucidation of ways to allow

findings from each approach to shed light on findings from the other approaches attempting to answer the same questions.

The methods have been chosen pragmatically, that is, for their usefulness in addressing the questions at hand. As William James put it, “[a pragmatist] is willing to take anything, to follow either logic or the senses if they have practical consequences” (James, quoted in Thayer, 1982, p. 225). So, here the conceptual frameworks, methods, and stances were chosen principally with regard to how best they contributed to answering the research questions. In particular, pragmatism suggests studying actions, as they are a person’s “practical reality,” regardless of what separate internal reality a person may profess to have. For, from this perspective, it is only the practical actions, born out of the participant’s experience, knowledge, and understanding, that matter to understand the concrete consequences of the participant’s reality at, in this case, solving a visualization meaning-making task. Since the practical consequences for and actions of the study populations while using the visualizations was of interest, then clinical interviewing was chosen from the constructivist tradition (Posner & Gertzog, 1982) to uncover what experiences participants bring to bear on their meaning-making. Eye-tracking from the neuroscience and cognitive psychology tradition (Duchowski, 2007; Holmqvist et al., 2011) was incorporated to bring to light unconscious neural actions that are driven in part by a participant’s reality. That is, the physical organization of the human sensory system drives information acquisition in a bottom-up manner, but the organization itself is shaped by how an individual’s understanding of the world is built through relating that information to existing information, creating a participant’s reality and thus affecting perception in a top-down fashion (Kandel et al., 2000; Schacter, 1996). Overarching is the contribution of the socioculturalists, specifically in the way the visualizations themselves afford or constrain meaning-making, based on the cultural-historical perspectives embedded in the visualizations and the social and cultural groups (including professional groups) to which participants belong

(Wertsch, 1994). Further, scaffolds for the visualizations were specifically designed to probe these social, cultural, and historical perspectives.

Social constructivism. Constructivism suggests that the individual constructs her own knowledge and reality based on personal experiences, as opposed to discovering an objective world (Piaget, 1967). Prior knowledge and experience shape what the individual considers relevant from the ongoing stream of sensory input. Then individuals either assimilate new knowledge into their existing frameworks, or alter those existing frameworks to accommodate both the prior and current knowledge and experiences.

The tenets of *social* constructivism postulate that the individual builds a mental representation of the environment based not only on internal knowledge, but also on external affordances and constraints. For example, “whether an individual's ideas are affirmed and shared by others … has a part to play in shaping the knowledge construction process” (Driver, 1995, p. 393). In terms of learning science, for example, it is not enough to put novices into situations where they will “discover” science. They also need to be given access to experiences, concepts, and models of science and assisted in the construction of their personal models through social mediation. Further, they need to be enculturated into academic science through help using all of their scientific understandings, especially where their constructed academic scientific understandings apply, namely what those understandings afford them to do, what problems they can tackle, and what constraints they impose and what questions those understandings cannot answer (Driver, 1995). An example employing scientific visualization is the CoVis project, which aimed to network teachers, students, and scientists online to collaboratively study atmospheric and environmental science through inquiry-based activities (“CoVis Information,” n.d.; Edelson & O’Neill, 1994). CoVis used just such a social constructivist lens on

classroom-based lessons involving creating and making meaning from scientific visualizations. Here,

> students play an active role in formulating research questions and pursue them in collaboration with their peers, teachers and scientist mentors. In the process of these investigations, students acquire a deeper understanding of the subject matter addressed, and learn other valuable lessons about how scientific inquiry is accomplished (Edelson & O'Neill, 1994).

In social constructivist environments such as these, the teacher in the classroom plays a central role not only in guiding students, but also more importantly in helping them "establish and maintain transformative learning conversations about these activities" (Edelson, Pea, & Louis D. Gomez, 1996, p. 153). Students in these studies have found flaws in models used to create the data they were using in their visualization environments, indicating they were able to construct meaning from the data and also make great strides toward participating in the practice of scientists by using the same tools.

The constructivist perspective offers two more advantages to science education as well as to my personal conceptual framework. First, it makes explicit the nature of science as provisional rather than absolute (Driver, 1995), opening the possibility of participation in construction of meaning. This underscores the culture-dependent nature of science and the need to be a part of the culture to be able to make shared meaning. This makes the divide between trained, enculturated experts and untrained, external novices more than just a matter of a lack of knowledge. It is instead a lack of enculturation that would enable one to be able to contextualize and make scientific (cultural) meaning of the knowledge that science develops. However, it also raises the possibility for newcomers, once they are enculturated, to fully participate in the negotiation of meaning, even to the point of challenging previously agreed-upon meaning.

Second, the social constructivists offer the chance to look at learning in a group as well as from a more traditional individual perspective. Without the ability

to consider how the group changes during meaning negotiation, or how group negotiation affects an individual, we cannot consider how meaning is ultimately arrived at in science. Group learning is also more reflective of the collaborative way scientists conduct their work (Edelson, 1997) and people learn outside of school, particularly in museums (Falk & Dierking, 2000). While learning in a group is not the focus of my work here, as we understand different starting points for novices and experts, we could work toward studying changes in the larger group of "scientific thinkers" as we scaffold novices toward full participation in the community of practice of science experts. For the visualizations particularly, I aim to work toward a set of practices that make meaning for a wide variety of audiences, based on groups of novices and experts working together to create shared meaning. After all, these visualizations, as communication objects, are designed for two purposes: "to convey meaning adequately and generate new meaning" (Lotman, quoted in Rowe, 2002, p. 27). The visualizations need not only scaffold individual learners in meaning-making, but also scaffold the group in making decisions based on their shared meaning and in communicating their meaning to a wider audience.

Finally, drawing on the social constructivist position will hopefully aid the connection between a more purely sociocultural perspective and an individual, cognitive, neuroscience view of learning. Cognitive scientists have recently been working to find practical implications of their laboratory-based work for naturalistic learning and education situations, with as yet limited success (Bruer, 2006; Cocking, Mestre, & Brown, 2000).

Socioculturalism. For this dissertation, I designed the scaffolds in the task to compensate for presumed differences in cultural backgrounds between the academic, expert scientists and the everyday, novice scientists in accordance with a sociocultural perspective. From a sociocultural perspective, teaching works best when the material to be learned is in the learner’s zone of proximal

development (Vygotsky, 1978). If the task in outside the zone of proximal development, scaffolding in the social or cultural surroundings, namely by another person or supporting contextual information, can help guide the learner to be able to solve the task. In the case of visualizations, educators provide scaffolding either through their own presence as a guide who draws attention to important elements of visualizations or who offers missing context or background knowledge. When those educators cannot be present, the visualizations themselves can contain material in such a way as to scaffold the visualization by backgrounding expert-level or tangential data, providing more novice-level supporting information, and foregrounding easier-to-master constituent tasks overall. In this dissertation, the scaffolds were designed remove previously revealed academic cultural conventions (Light & Bartlein, 2004; Phipps & Rowe, 2010) and replace them with more broadly culturally familiar colors, words, and arrangements of supporting elements.

Similarly to social constructivism, the sociocultural approach I take here is founded on a tenet of relativistic, rather than absolute, knowledge. To socioculturalists, knowledge and all higher mental functions are rooted in, emerge from, and are negotiated in social interaction first but eventually are internalized by an individual as that knowledge or those functions become relevant to him or her and is integrated into their identity and social interactions (Vygotsky, 1978). For professionals, that means meaning making cannot be divorced from the social situation or the classification schemes and norms of discourse that have been agreed upon by the practitioners (Goodwin, 1994). This meaning-making must be learned by participation in the practice, especially in conversations about practice, in order to learn not only how to harness biological mechanisms such as vision to focus on relevant features, but also how to describe what is being seen in context in order to interpret the physical perceptions culturally (Eberbach & Crowley, 2009; Goodwin, 1994).

A more purely sociocultural approach contrasts with social constructivism by focusing on the interactions of a group, such that any considerations of individual learners cannot be done apart from understanding of their social environment and contexts. Beyond individual knowledge, even what constitutes relevant knowledge in science is constructed by the community of practicing scientists based on social and cultural scientific norms, the "*community's agreement* upon nature's correct answer (emphasis in original)" (Magnusson et al., 2004, p. 133). This knowledge depends on the values, beliefs, and standards of the scientific community about "what is important to know and do (Magnusson et al., 2004, p. 133).

Conventions for academic visualizations were developed within and agreed upon by the overall community of practice (Lave & Wenger, 1991; Light & Bartlein, 2004). Practicing academic scientists, often referred to as members of "the scientific community," have created their own visualization culture, tacitly if not explicitly defining amongst themselves the conventions that get widespread use in the visualizations they produce. They often imitate visualizations they have seen published in journals by other academic scientists in their field (Light & Bartlein, 2004; Magnusson et al., 2004). Further, academic scientists have appropriated certain everyday words and imbued them with specific academic meanings as a Discourse (Gee, 1999). This distinction causes academics to judge some everyday scientific explanations with these words as incorrect when in fact, everyday explanations often accurately describe the physical world but are using the alternate, everyday meaning of words (Gee, 1999).

This academic science culture effectively shuts out everyday scientists who are unfamiliar with the agreed-upon knowledge, values, and goals of the expert scientist culture, unless the novices are willing to undergo a long training process. Currently, there is little reward for completing this process, except for an academic scientific career. The price is high for everyday scientists simply wishing to use the knowledge of the field for personal or societal decisions.

In articulating these ideas related to expertise and expert culture development, the sociocultural approach begins to tie into the more purely cognitive approach of neuroscientists. Eventually, by internalizing a field's relevant knowledge and practices and participating in the negotiation of knowledge, a novice becomes an expert as the new practices become automatic. For Vygotsky, expertise is "flexible and fluent" use of cultural tools (Wertsch, 2007, p. 190). To Leontyev (2009), goals and actions become integrated and form skills and habits rather than independent thoughts. For Lurija, expertise in speech is essentially development of a new "sense" or context for meaning (1981, p. 45). Adults, for example, know of and can use several different connotations of words depending on the situation, whereas young children have only concrete meanings for words (Halliday, 1975). For all of these types of expertise, changes in neurological connections accompany the changes toward automation, as sensory cortex use diminishes and higher cortical system connections come into prominence (Lurija, 1979).

Cognitive psychology and neuroscience. Geographers and others researching and developing geo-visualizations understand that elucidating those areas of human perception, cognition and visual processing that impact use of visual communication tools such as maps and data-filled graphics is a key challenge in order to move the discipline of geography forward (MacEachren & Kraak, 2001). Color is again very relevant when viewed through this lens. For example, the rainbow color scheme favored by many scientific visualizers represents middle values as yellows and highs as reds, but yellow-green is perceptually brightest to the human eye (Faughn & Serway, 2003), so the middle range data often ends up “standing out” visually when it does not statistically. This color scheme is also perceived quite differently by color blind individuals – both ends of the spectrum end up the same color to people lacking red-sensitive cones in their retinas (Light & Bartlein, 2004). Also, depiction of a spread of

colors might imply to the layperson that equivalent numerical changes exist where they do not, i.e. if shades of blue are thought to be more similar than shades of red, due to the lower relative numbers and sensitivity of blue- versus red-sensitive cone cells in the eye that structure the processing of the human visual system (Slocum et al., 2001; Ware, 2004). Baker and Dwyer (2000) and Slocum et al. (2001) suggest that the brain is limited in its capability for visual interpretation, so simplified representations are easier to interpret than realistic images, including photographs. Neuroscience techniques applied to geographic tool use investigations could examine the cognitive load imposed by different versions of visualizations, or unconscious patterns of attention to visualization elements that users cannot verbalize (Duchowski, 2007; Holmqvist et al., 2011).

In the cognitive science technique of eye-tracking, infrared light records the angle from the recorder to a participant's pupil. From this, the participant's gaze at a stimulus visualization, object, or video is calculated to within generally one degree (approximately 20 horizontal pixels in a 800x600 pixel resolution visualization, nominally 3% of the visualization) ("ExperimentCenter™ Manual version 3.1," 2012). Also calculated are the duration of gazes on any given location, the order of gaze positioning, and the direction and timing of micro-level subconscious movements called saccades that eyes make to gather visual information even while seemingly remaining fixed (Holmqvist et al., 2011). "Tracking" the user's gaze is taken to be a measure of their overt visual attention to stimuli based on both the bottom-up physical processing capabilities as well as the user's top-down, consciously-directed concentration (Duchowski, 2007). Fixations are "directed toward task-relevant information as it is needed for ongoing visual and cognitive computations" (Henderson, Brockmole, Castelhano, & Mack, 2007, p. 539).

Many academic studies have focused on eye-tracking with traditional fundamental neuroscience and psychology tasks aimed at uncovering the mechanics of human vision (cf. Duchowski, 2007; Holmqvist et al., 2011).

Recently, more naturalistic, real-world tasks have been the subject of study, especially reading, and less so visual search (Rayner, 1998).

Investigations of natural scenery perception (e.g. Hughes, Kitterle, & Nozawa, 1996) and most recently, natural task performance, made possible by the advent of head-mounted tracking systems (Duchowski, 2007), are the ones of most potential relevance to understanding expertise and meaning-making from visuals, especially in an informal education setting. van Gog (2006) found differences in fixation duration in an electrical circuit problem solving-task between higher- and lower-expertise participants, and Reingold et al. (2001) used eye-tracking to suggest chess expertise is based on training, not on general memory superiority. In particular, Grant and Spivey (2003) showed improved performance when increasing visual salience of a task-relevant feature of the visualization, a type of scaffolding akin to that used here. In museums, eye-tracking yields useful information about parts of art pieces visitors focus on while working with museum educators, and fixation durations of visitors at a science museum exhibit about a controversial topic increase as visitors view authentic objects versus photographic versions as primary sources (Filippini-Fantoni, Jaebker, Bauer, & Stofer, 2013). Finally, in particular, eye-tracking with studies of charts, graphs, and tables find differing tracking patterns on differing versions of visualizations of the same information (Burch, Konevtsova, Heinrich, Hoeferlin, & Weiskopf, 2011; Cheng & Peebles, 2003; Kim, Dong, Xian, Upatising, & Yi, 2012; Steinberger, Waldner, Streit, Lex, & Schmalstieg, 2011).

In terms of the bricolage in this study, a cognitive approach allows me to study the unconscious physical effects on the participant's visual perception system while they are participating in a visual meaning-making task. Accompanied by the interviews, I can investigate whether the eye-tracking pattern discrepancies are due to effects of visual salience or cognitive control. Specifically, fixations that focus on perceptually salient features of the

visualization, rather than features important for meaning-making, could point to areas of communication tools that need improvement.

A cognitive approach also allows me to consider the impact of cognitive load on a user viewing visualizations. The amount of mental work required to make sense of and ultimately make meaning from any given type of visualization differs based on what they make explicit and what is left to be deduced by the user (cf. Ainsworth, 2006; Larkin & Simon, 1987; Scaife & Rogers, 1996; Zhang & Norman, 1994). Thus different visualizations are likely to have different affordances and constraints for learning based in part on their differential computational loads imposed by the physical processing demands resulting both from cultural and biological salience match or mismatch.

Research Plan

The overall research plan involves comparing adult oceanography experts with adult novices while they observe and make meaning from global ocean data visualizations. The populations were chosen to create as much potential disparity in formal science background knowledge and professional experience as possible to maximize discriminatory power. Performance was compared while participants viewed visualizations in three experiments: semi-structured clinical interviews to ascertain the nature and extent of their knowledge about the domain in question (Posner & Gertzog, 1982, p. 195; Piaget, 1929), and two eye-tracking experiments with concurrent interviews to look at unconscious eye movements and search strategies. Both experts and non-experts were recruited for the clinical interviews and the laboratory eye-tracking; a separate general public population was recruited for an *in situ* eye-tracking experiment on a spherical display system (SDS) in a science center. The *in situ* was conducted to investigate both a more generalizable population sample and a more naturalistic setting for the imagery presentation for outreach. The interviews and eye-tracking experiments were all conducted by the author.

Sampling

Due to the use of color schemes in these studies, color-blind participants were excluded from all phases. Some participants who wear corrective lenses, either glasses or contacts, were not able to be part of the eye-tracking experiment due to limitations of the equipment and were excluded at that time, though that exclusion did not preclude the use of their clinical interview data, if applicable. As appropriate, recruitment materials and screening and consent processes included these possible exclusion criteria.

Clinical interview participants. Four pilot participants were recruited from a convenience sample of other graduate students and oceanography researchers: one doctoral student in science education, two master's students in marine science, and one university oceanography faculty member with less than five years' experience beyond completing the doctoral degree. Pilot participants for the clinical interview questions received a five dollar gift card to a local coffee shop for participation.

For the clinical interviews, adults of at least 18 years old who had not yet completed two years of college and who were not pursuing a science or engineering major were recruited. Recruitment methods included fliers posted in buildings and other public posting sites around the campuses of a large four-year science and technology university in the Pacific Northwest of the United States, a two-year college nearby, and the surrounding community. Participants were offered $10 in gift cards to a local coffee shop as incentive for their participation. Prospective interviewees completed an online survey providing contact information and confirming their age of at least 18, level of college completion, major, normal color vision, and availability to complete the one-hour interview.

For the expert population, faculty of the university who had held a doctoral degree in Oceanography for at least five years were listed alphabetically, using

listings provided on the university department web site. The qualified faculty were numbered in order, then contacted in random order based on a random number generator (“Random.org - True Random Number Service,” n.d.). Faculty were offered the same $10 gift card incentive and invited via their university email or phone to complete the online screening process, which was similar to the novice screening but confirmed oceanography doctorate completion plus five years’ experience and normal color vision.

Groups of approximately seven faculty were contacted at a time via email until the desired number of 12 total expert participants agreed to participate, with smaller groups contacted as the number of remaining spots were diminished. Faculty who did not complete the online screening or respond via email that they could not participate were contacted with a follow-up phone call to their on-campus offices approximately one week after the invitation email. If they were then not reached by phone and did not return the phone call or complete the online screening at that point, they were not contacted further. Reasons cited for not participating were travel or simply being too busy. In addition, two more faculty were recruited in a similar manner to complete the eye-tracking experiment only. Overall, 26 faculty were contacted between June 2012 and February 2013.

An informant in the department was asked to suggest which experts were regular users or creators of visualizations of data and which were not, to sub-divide the population into those oceanographers who work closely with visualizations of ocean data and those who do not. Since the informant was familiar with the work of only approximately half of the population with requisite oceanography experience, I also asked faculty themselves at the start of their interviews whether they worked with visualizations of data frequently, occasionally, or not at all, in order to sub-divide them. All 14 expert participants self-identified on a 3-point Likert-type scale of frequency of visualization use as 2 (*frequently*) or 3 (*very frequently*). This may indicate either that all oceanography

faculty at the university use visualizations regularly, or that only those faculty that do use visualizations frequently responded to the recruitment. Due to the informant's lack of familiarity with the entire department and the wide-ranging experience with visualizations as evidenced by the self-report, this criterion proved less discriminatory in this manner than hoped. A better question needs to be developed if a degree in oceanography is not considered the keenest measure of expertise with these particular spatially-based visualizations of ocean data.

Novices who completed the online screening, and faculty who completed the online screening or were reached by phone, were contacted via their method of choice and invited to schedule an interview time slot of 90 minutes. Participants were scheduled and recruited on a rolling basis until 17 novice and 12 expert interviews were completed.

Gender distribution of participants relative to their overall populations are shown in Table 1. To maintain gender balance roughly equivalent to that of

Table 1

Presented Gender Balance, Participants versus Population

Presented Gender	Recruited Experts[a]	Expert Population[b]	Recruited Novices[a]	Novice Population[c]
Male	75%	85%	35%	48.6%
Female	25%	15%	65%	51.4%

[a]As judged based on presentation and first names.
[b]As judged based on first names and web site photos on the university web site.
[c]U.S. Census estimated 2010 population 18 and over (U. S. Census Bureau, 2008).

the overall oceanography faculty, the final few expert participants recruited for the clinical interview were females only, though they were chosen at random from the qualified female faculty population otherwise. On the other hand, to maintain gender balance roughly equivalent to that of the general population, novice males were preferentially recruited toward the end of the recruitment

phase through emails to fraternities and wording asking for male participants in emails to university major departments.

Laboratory eye-tracking participants. Two pilot participants were recruited for the study: one science education professor as a stand-in for the "expert" population, and one student from the master's in science teaching licensure program to serve as "novice".

Contact information was retained for participants who were willing to participate in further experiments as indicated on the consent form, and the subset of willing participants from the interviews were used for initial recruitment for the laboratory eye-tracking experiment. Participants in this phase of the study were offered $20 each for their participation.

Two additional experts were recruited from the expert population as described in the clinical interview recruitment to provide a full complement of 10 participants for this experiment, due to dropout by some of the original interview participants.

***In situ* eye-tracking participants.** For the *in situ* eye-tracking, 13 participants at least 18 years old were recruited from visitors to a university marine science center visitor center on two weekend days in February 2013 as they approached the spherical display system exhibit. The first adult visitor to approach the exhibit room from every other group was recruited. If they declined to participate, a volunteer from their group could participate, or the group count restarted and the first visitor from the next group was approached. Visitors were not offered an incentive to participate.

Visualization Preparation

Three baseline "scientific" visualizations were used in the various stages of this study. Two were global versions of the same topics used in Phipps and

Rowe's (2010) study of museum visitors and teachers: sea surface temperature (SST) and chlorophyll in the ocean. The third was a related global visualization: sea surface temperature anomaly. Each scientific visualization was derived from satellite data averaged from all observations over a month. For the anomaly, the climatological average data from the Advanced Very High Resolution Radiometer satellite instrument from 1985-1997 was subtracted from the January 2010 sea surface temperature average data (Casey & Cornillon, 1985). See Appendix 1 for example visualizations.

Five visualizations were originally produced for each topic. One version had no scaffolding (“unscaffolded”, US). Three each had one element of scaffolding: geographic labels (GS); culturally-relevant colors (CS); or title, measurement units, and key placement (TS). The fifth had with all three elements of scaffolding (“fully-scaffolded”, FS).

For SST, I used the following scaffolding: color ramp of black or dark purple through pale yellow and white, increasing equally in saturation, hue, and brightness at each scale step, as described by Harrower and Brewer (2011); geographic labels for five continents (Asia, North America, South America, Australia, Africa), the Pacific, Atlantic, and Indian ocean basins, and the equator (a dashed line, no word label); and title “One Month Average Sea Surface Temperature,” with measurement units in both degrees Celsius and Fahrenheit, though abbreviated “ºC” and “ºF”, respectively. For the SST anomaly visualizations I used scaffolding as follows: color ramp of dark blue through pale blue for lower-than-average values, white for average or middle values, and pale red through dark red for higher-than-average values; geographic labels identical to the SST visualizations; title “One Month Average Sea Surface Temperature Difference from Average,” and measurement units +/- ºC and ºF. For chlorophyll the scaffolding is as follows: color ramp of dark blue through pale yellow (again increasing equally over hue, saturation, and brightness), geographic labels as before, and title: “One Month Average Ocean Productivity” or “One Month

Average Microscopic Ocean Plant Concentration.” The first title variant was used in the initial interviews until it was pointed out by an expert participant that productivity is a rate, with an element of time, rather than strictly a mass per unit volume. It was then substituted with the latter for all subsequent interviews. In each set, for the scaffolded title level (TS), the key, composed of color bar and measurement units, was moved from the bottom of the visualization to the left-hand side. For all three topics, the fully-scaffolded visualizations with all three elements (FS) will combine the elements of the single-element-scaffolded visualizations (CS, TS, GS) without alteration.

During the clinical interviews, 11 of 12 expert participants recognized that the data from later visualizations was the same as in the visualizations already presented, though colored and labeled differently. They also struggled to determine season of the year for the visualizations presented (Northern Hemisphere winter). Therefore, for the eye-tracking experiment, an additional two visualizations from April 2010 and July 2010 for each topic were created to try to eliminate the recognition effect and allow further investigation of the confusion about seasons. The geographic (GS) and color (CS) scaffolding visualization versions used April 2010 data, and the title (TS) and full (FS) scaffolding visualization versions used July 2010 data; January 2010 remained the source of data for the unscaffolded (US) versions.

To ensure consistency in visualization presentation across experiments, all visualizations were created at 640 x 480 pixels resolution due to the limitations of the functional magnetic resonance imaging (fMRI) scanner, planned for subsequent studies with these visualizations. Colors for areas with no data are dark gray for land without data, lighter gray for ocean without data. Fonts were 12-point, colored black for maximum contrast with the visualizations’ white background. Map projection was Robinson, an almost-ellipse compromise projection created to be visually appealing by not totally eliminating any form of distortion (Dean, n.d.). This is similar to the Hammer-Aitoff projection used for

NASA's Earth Observatory web site global visualizations (“Global Maps,” n.d.). The stimulus visualizations were centered on the Pacific Ocean basin due to the focus on ocean data in these experiments.

Location Setup

Interviews and laboratory eye-tracking were conducted in an office on the university campus. Windows were closed, blinds drawn, and fluorescent overhead lights were used for all interviews and laboratory eye-tracking to maintain consistent ambient lighting and minimize background noise. The telephone was disconnected.

For clinical interviews, participants sat in front of a 22” external computer monitor at a comfortable distance away. The researcher sat to the left side of the participant, controlling visualization presentation manually via a laptop computer using ExperimentCenter™ presentation software from SensoMotoric Instruments, Inc. (SMI) to maintain consistency of presentation for the eye-tracking experiment. For the clinical interviews, the “dry run” function was used to randomize stimulus presentation to prevent fatigue and learning effects but not collect eye-tracking data. Visualizations were manually advanced by the researcher for the clinical interviews. In-between stimulus visualizations, a "noise" image produced by White Noise Generator 1.0 matched to the contrast and brightness of the test visualizations was presented briefly to the participant, while they may have been asked if they needed a break, or reminded to look at the next stimulus visualization for 10 seconds, before the researcher manually advanced to the next stimulus. This noise image served as a break during the clinical interviews, and during eye-tracking, prevented pupil relaxation or dilation that might affect precision of the eye-tracker while allowing the participant to return to a “start” position of their eyes instead of starting where their eyes had finished tracking the last stimulus. An audio recorder captured the interview, while a video camera over the shoulder of the participant captured the audio as

backup and also recorded any gestures the participant made toward the screen or during explanations.

For the laboratory eye-tracking, a similar viewing setup and distance were used, with the SMI remote eye-tracking device (RED-m™) eye-tracker attached to the bottom of the monitor. The RED-m™ video-based remote eye-tracking device uses a binocular infrared camera with an angular resolution of 0.5 degrees. Due to constraints of the system, the participants were asked to stay still, to maintain their eyes within 60-80 cm in front of the monitor and within a 32 cm range left to right and 21 cm range up and down as required by the eye-tracking device. Thus, the monitor height and participant chair were adjusted to approximately center the participant's eyes with the eye-tracking system. Again, the researcher controlled a laptop to the left side of the participant to present the standard five-point calibration procedures with a four-point validation using ExperimentCenter™ presentation software ("ExperimentCenter™ Manual version 3.1," 2012). Calibration was repeated until all participants were calibrated within ideally 0.5 degrees, but definitely one degree, in both the horizontal and vertical directions. The RED-m™ captured gaze data at 120 Hz via iViewX™ software.

Visualizations were presented using ExperimentCenter™ software to control randomization and auto-advance; the researcher controlled manual advance in the software after the subsequent concurrent interview period for each of the five versions of the visualizations shown. Visualization contrast, brightness, and resolution matched that of the clinical interviews presentation. Audio responses were recorded with an external webcam placed on top of the stimulus monitor, connected to the laptop, and controlled by the stimulus presentation software to allow synchronization with eye-tracking data.

For the *in situ* eye-tracking, participants stood 60 -- 80cm in front of the same SMI-RED-m™ eye-tracker used in the expert-novice experiments. In this setup, the eye-tracker was on its stand-alone stand on top of a tripod that was height-adjustable. After inputting the screen size and selecting standing rather

than seated participants, iViewX™ software gave a recommended position for the eye-tracker height and distance from the screen so that the tracker was the required 60 - 80 cm from the visitor's eyes. The spherical display system with the stimuli was approximately one meter behind the eye-tracker, for a total distance of approximately 170 cm from the participant. The eye-tracker was placed at a height of 151 cm from the floor to the bottom of the tripod and an initial angle of eight degrees, as recommended by the software. The researcher and laptop stood to the left and in front of the participants, next to the sphere. See Figure 2.

Figure 2. *In situ* Eye-tracking setup. The participant, right, stands in front of the eye-tracker on a tripod in front of the exhibit. The researcher, left, controls the eye-tracker via computer and also presents the stimulus image on the globe using the touch screen monitor at the lower-left of the globe. Here, the stimulus image is the unscaffolded version of the Sea Surface Temperature Anomaly visualization, with color bar visible on the globe at the bottom of the visualization.

The researcher obtained informed consent, verified the visitors were not color blind, and then angled the eye-tracker as necessary to locate visitors' eyes on the laptop screen. Participants were instructed to view five calibration points shown on the spherical display in order based on the researcher's verbal instructions as she verified the fixation on screen in iViewX™. Stimuli were presented using the spherical display system's proprietary software, StoryTeller™. For both eye-tracking experiments, the participant position was monitored by the author and adjusted with verbal instructions if necessary. The SMI-RED-m™ eye-tracker under the control of iViewX™ software recorded gaze data at 120 Hz.

Participant audio was recorded using the eye-tracking Experiment Center software via an attached web camera that recorded an image of the spherical display system to synchronize eye-tracking data with the participant's audio and the image presented.

Pilot Testing

Pilot testing for clinical interviews primarily assessed the time to complete the entire interview with 10 visualizations, as well as clarity of questions and instructions. The pilot participants underwent color vision screening and then were asked the clinical interview questions. After pilot testing, the semi-structured interview protocol was finalized (see Appendix 2).

Pilot testing for laboratory eye-tracking tested one graduate student in the university's Master's in teaching licensure program and one faculty member from Education. Testing assessed performance of software and refined the experiment setup in the software. Pilot testing also assessed participant ability to answer questions while having pupils tracked, plus participant comfort and time to fatigue without breaks. Pilot work established clarity of instructions, where to position the tracking system, and how to run the presentation software. The pilot testing for eye tracking determined a think-aloud, abbreviated clinical interview

concurrent to the eye-tracking to be the best method (Holmqvist et al., 2011) for assessing comprehension of these visualizations to dovetail with clinical interviews. Finally, pilot testing suggested that calibration of participants at the beginning of the experiment was sufficient for the duration and did not need to be redone during the experiment.

Experimental Procedure

Clinical interviews. Clinical interviews presented participants with two of the three topics of visualizations. Topics were randomly selected for participants, so that six novices (except for the chlorophyll first, anomaly second case) and four experts looked at each possible pair of topics. Order of presentation of the topics was also randomized, but as the presentation order SST then SST anomaly was equivalent to SST anomaly then SST, there were total topic pairs (SST and SST anomaly, SST and chlorophyll, or SST anomaly and chlorophyll). Within each set, order of presentation was randomized to prevent fatigue and learning effects. Each participant was shown up to 10 visualizations, one for each of the five levels of scaffolding for each of the two topics. See Table 2.

The total number of visualizations viewed by each participant varied based on the thoroughness of their answers, their indication of similarity of previous visualizations, or on the time constraint of no more than one hour for the clinical interview (for further explanation see chapter 4). For the clinical interviews, all data was from January 2010.

For each visualization, participants were asked about five themes: geography (based on location of highest data values), colors, main idea, measurement unit and scale, and time of year, with up to several content questions and tasks of varying levels for each theme. After each content question, probes were used as follow-up, generally one or more of the following: a) why do you think that, b) what in the visualization suggests that to you,

Table 2

Stimulus Topic Presentation and Order, Clinical Interviews

Participant[a]	Participant Population	Topic 1	Topic 2
Keith	Expert – Non-Visualizer	SST	SST anomaly
Ray	Expert – Non-Visualizer	SST	chlorophyll
Matt	Expert – Non-Visualizer	SST anomaly	SST
Alice	Expert – Non-Visualizer	SST anomaly	chlorophyll
Eric	Expert – Non-Visualizer	chlorophyll	SST
Rick	Expert – Non-Visualizer	chlorophyll	SST anomaly
Mark	Expert – Visualizer	SST	SST anomaly
Charlie	Expert – Visualizer	SST	chlorophyll
Lindsey	Expert – Visualizer	SST anomaly	SST
Brent	Expert – Visualizer	SST anomaly	chlorophyll
Janet	Expert – Visualizer	chlorophyll	SST
Jay	Expert – Visualizer	chlorophyll	SST anomaly
Samantha	Novice	SST	SST anomaly
Ferdinand	Novice	SST	chlorophyll
Virginia	Novice	SST anomaly	SST
Gina	Novice	SST anomaly	chlorophyll
Ivan	Novice	chlorophyll	SST
Linda	Novice	chlorophyll	SST anomaly
Veronica	Novice	SST	SST anomaly
Allison	Novice	SST	chlorophyll
Shelby	Novice	SST anomaly	SST
Emma	Novice	SST anomaly	chlorophyll
Glenn	Novice	chlorophyll	SST
Jeff	Novice	chlorophyll	SST anomaly
Vanessa	Novice	SST	SST anomaly
Brad	Novice	SST	chlorophyll
Mikayla	Novice	SST anomaly	SST
Eden	Novice	SST anomaly	chlorophyll
Kyle	Novice	chlorophyll	SST

[a]All participants' names are pseudonyms.

c) how do you know, or d) tell me more about that. See Appendix 2 for full interview protocol.

Laboratory eye-tracking. The laboratory eye-tracking experiment presented participants visualizations from the topic they did not see in the clinical interviews. See Table 3 for full layout of data collected per participant and per

Table 3

Stimulus Topic Presentation, Laboratory Eye-tracking

Participant	Participant Population	Topic
Keith	Expert – Non-Visualizer	chlorophyll
Ray	Expert – Non-Visualizer	SST anomaly
Matt	Expert – Non-Visualizer	chlorophyll
Rick	Expert – Non-Visualizer	SST
Mark	Expert – Visualizer	chlorophyll
Brent	Expert – Visualizer	SST
Janet	Expert – Visualizer	SST anomaly
Jay	Expert – Visualizer	SST
Jack	Expert – Visualizer	SST anomaly
Maureen	Expert[a]	SST anomaly
Ferdinand	Novice	SST anomaly
Ivan	Novice	SST anomaly
Veronica	Novice	chlorophyll
Allison	Novice	SST anomaly
Glenn	Novice	SST anomaly
Jeff	Novice	SST
Brad	Novice	SST anomaly
Mikayla	Novice	chlorophyll
Eden	Novice	SST
Kyle	Novice	SST anomaly

[a]Maureen was not asked her level of frequency of using visualizations.

topic. Each participant in the eye-tracking experiment saw all five levels of scaffolding for the topic. A standard "Blue Marble" visualization (a satellite montage that mimics a view of the Earth from space as if it were in perpetual daylight and visible all at once), a "South-up" version of the Blue Marble (with the South Pole, rather than the North Pole, at the top of the visualization), and a “no data” visualization with background colors only (dark gray for land, light gray for

ocean as in the experimental visualizations), were incorporated into the presentation for control and randomized with the experimental visualizations, for a total of eight stimuli.

Each visualization was presented to the participant for a 10-second spontaneous looking period (Libarkin, Clark, & Simmon, n.d.). Afterwards, while the stimulus was still shown and the eye-tracker still collected data, the researcher asked five content questions and generally a single follow-up probe in a shorter, more structured version of the semi-structured clinical interview. See Appendix 2 for full protocol.

***In situ* eye-tracking.** Due to the constraints of the eye-tracker and the size of the SDS, participants were only able to view about 15% of the globe, instead of the whole globe as in the expert-novice experiments. Participants were shown one fully-scaffolded (FS) visualization and one unscaffolded (US) visualization, centered on the Pacific ocean basin, from either the chlorophyll or sea surface temperature anomaly topic, as sea surface temperature was either most readily recognized or the standard guess of non-expert participants in the clinical interviews and laboratory eye-tracking experiments. The visualizations were taken from the previous experiments, with the key and title moved to be presented on the sphere, meaning they covered part of the data itself (see photo). Order of presentation (scaffolded or unscaffolded first) and topic were alternated between participants, so that every fourth participant saw the same topics and versions in the same order.

Participants were instructed to stand as still as possible and view the visualization shown on the globe while the interviewer asked them questions. Participants were first given 10 seconds to look at the visualization before the first question was asked, the spontaneous looking condition akin to that in the expert-novice eye-tracking experiment. Then participants were asked similar

questions to the expert-novice eye-tracking while viewing the visualizations, in a semi-structured, semi-clinical manner:

What is the main idea of the image shown?
What do the colors represent?
Where are the highest values in the image?
What time span is represented?
What season is represented?

I then repeated the procedure of 10 seconds of spontaneous looking followed by the interview for the second visualization.

After the participant had looked at both visualizations and been interviewed, the eye-tracker was turned off and the participant was asked about their comfort with interpreting such visualizations (not at all comfortable, comfortable, or very comfortable), what high school, college, and professional science background they had, and what sorts of science-related hobbies, if any, they pursued, including watching science documentaries. Gender was recorded based on presentation. These answers were recorded in the interviewer's field notes.

Analysis

Each method provided data to be analyzed both qualitatively and quantitatively. Clinical interviews and the interviews that accompanied eye-tracking were qualitatively coded for strategies for meaning-making and "confusion," defined as any verbal indication by word or tone of uncertainty in an answer. The interviews were also quantitatively scored for correctness of answers based on a rubric described below. Eye-tracking provides qualitative data in the form of scan paths and quantitative data in terms of focus areas (fixations) and dwell times (durations) on those areas.

Independent variables considered throughout the analysis were: gender, expertise, level of scaffolding, topic, and experiment trial number. Level of scaffolding and topic were randomized for presentation. Dependent variables

were: correctness of answer, number of fixations, duration of fixation, and area of interest (AOI) “Areas of interest” are geometric subdivisions of the stimulus visualizations to look at the variables as they impact particular elements of the visualization in relation to one another.

Clinical interviews. The author coded interview transcripts of individual participants, then bundled related ideas were bundled into themes and to identify common themes across participants, recreating subthemes or sub-ideas as necessary as coding progressed, based on standard qualitative interpretive coding methods (e.g., Patton, 2002; Bernard, 2005).

To establish accuracy and performance ranges, participant answers were also coded according to a rubric. I developed the rubric in conjunction with a collaborator based on anticipated subject matter of answers, which sometimes allowed leeway from an absolutely objectively correct answer based on the data underlying the visualizations. Possible answers were rated “incorrect” or “don’t know” (scored 0), “partially correct” (0.5), “correct” (1), and “sophisticated” (1.25), in which the participant volunteered more detail than the minimum sought by the question. If a question was asked and an off-topic or irrelevant answer was given, the question was not counted as asked. See Appendix 3 for full scoring rubric.

Two colleagues also scored the first two visualizations viewed by one expert and one novice participant for inter-rater reliability. Agreement with the first colleague was 100% within one score category, with 79% of scores identical. Agreement with the second colleague was 95% within one score category, with 36% of scores identical. The one question that was not scored within one score category was “Are the conditions in the Equatorial Pacific normal?” on which it was not possible to score partially correct.

Since the interview was semi-structured, not all participants were asked every question for each visualization. Thus, scores for the main idea questions

for each of the visualizations were summed and divided by the total questions asked of each participant for that visualization for comparison as an visualization sub score. All sub scores were combined into an overall participant score.

Laboratory eye-tracking. SMI BeGaze™ software converted participants' three-dimensional gaze coordinates, pupil direction, and timing data into three types of data considered here: fixations, that is, points at which users gaze for a length of time above a certain minimum threshold, durations, the amount of time a participant looks at a point, and scan paths, the order in which participants view features or make fixations. Fixations were calculated as groups of points with dispersion of no more than 100 pixels and with minimum dwell 80 ms.

These numbers were overlaid onto the stimulus visualizations to determine first which, if any, fixations existed, then which fixations corresponded to which features of the stimulus visualizations, and which fixations on interesting features were viewed in what order (scan path). Areas of interest (AOIs) were drawn around the title, scale bar, and map portions of the visualizations in BeGaze™, and fixations were grouped into these regions for analysis of probability of looking at each. See Appendix 4 for examples of the AOIs on the stimulus visualizations. These patterns of probability were generated for each individual participant and then across expert participants and across non-expert participants, across topics, and across scaffolding levels.

I drew four areas of interest on the US, CS, and GS visualizations, referred to as the "larger map" scaffolding levels. The first three were: the map portion with data overlay, the title, and the key, consisting of the color bar and measurement unit labels. The map was surrounded using the polygon tool in BeGaze™ to approximate the flattened oval of the Robinson projection of the map. I surrounded the title and key each with a rectangle. In the FS and TS cases, due to the repositioning of the key from the bottom to the left-hand side of

the visualization, the map became smaller, so these cases are referred to as the "smaller map" case. Therefore, I drew a fourth AOI, called "overlap" in the larger map case based on the position of the key in the smaller map version. This "overlap" AOI was drawn to ensure that participants were not spending a significant amount of time on that portion of the larger map (US, CS, GS cases) so that I could be assured that the participants looking in that area in the smaller map cases (TS, FS) were indeed interested in the key and not just the left-hand side of the map, mostly depicting Africa and Western Europe.

The "larger" or "smaller" distinction here in the AOI names is by size of the map, which means it is not associated with geographical labels or color scaffolding, but rather with title scaffolding. In the TS and FS visualizations, the position of the key changed along with the measurement unit and title scaffolding. That is, the US, CS and GS cases have the larger map AOI of the unscaffolded case, and the TS case has the smaller map size of the fully-scaffolded case, even though all three intermediate cases are one level of scaffolding, with no presumed effect-size difference among them. See Appendix 4 for example AOIs.

In the TS and FS smaller map case, four AOIs were also drawn for most of the images. The first three were analogous to the larger map AOIs. I surrounded the map itself with a polygon approximation and the title and key each with a rectangle. The fourth AOI in the smaller map case for SST anomaly and chlorophyll visualizations is a polygon drawn around the key to minimize its spillover onto the map itself ("color cutout"), as the scaffolded text of the key was longer, and the width of the area smaller, after moving the key to the left than in the smaller image case. The SST key did not run onto the map itself, so there was no "color cutout" AOI for that smaller map and the TS and FS levels. See Table 4 for the AOIs associated with each scaffolding level. In all cases, I drew the AOIs as conservatively as possible around the elements. Holmqvist, et al., (2011) note that the eye's foveal size of 1 -- 1.5 degrees should be the minimum

Table 4

Areas of Interest Associated with Scaffolding Levels

Topic	US, CS, GS Levels				TS, FS Levels			
	larger map	key	title	overlap	smaller map	key	color cutout	title
SST	x	x	x	x	x	x	N/A	x
SST anomaly	x	x	x	x	x	x	x	x
Chlorophyll	x	x	x	x	x	x	x	x

border size around an AOI; here that translated to approximately 20 horizontal pixels, given a 22” monitor with 800-pixel width with an average viewing distance of 70 cm, and fewer vertical pixels for the screen height of 600 pixels. Given the visualization limitations and BeGaze™ software limitations, the 20 pixels, or about 3% of the visualization width, were approximated as the minimum border for drawing AOIs.

Due to the overlap of the “overlap” AOI with part of the “map” AOI in the larger map case, and the overlap of the “color cutout” with “key” in the smaller map case, some fixations were reported by BeGaze™ as in both of these AOIs. When that was the case, the smaller of the AOIs was presumed to be the target (AOI overlap or color cutout), and that AOI label was carried forward for AOI analysis. Given this criterion, the overlap regions accounted for only 10 of 2564 fixations (six in the anomaly visualizations, four in the chlorophyll visualizations) in the spontaneous looking condition, and thus the overlap was presumed not to be an area of the map on which participants generally relied. In addition, when key and map overlapped, it generally did so within an area of previous or subsequent key and color cutout overlap, indicating a likely continuation of looking at the key, rather than the map. For these two reasons, the key AOI in smaller map stimuli, which encompassed the smaller color cutout, were treated as key rather than map hits in cases when they overlapped. Altogether, that provided 46 anomaly key AOI hits in smaller map cases, and 19 for the chlorophyll, still relatively small compared to the overall number of smaller map

hits (575). In the main idea condition, 161 of 4201 total fixations (3.8%) were duplicates and were considered as described above. Of these, 49 (1.2% of the total) were duplicates of the overlap and map AOIs in the SST case, and were thus considered as SST map AOI. This was the largest percentage in any of the cases of the overlap; thus, in the main idea condition overlap of the key with the map was also not considered to be a problem.

Using R open-source statistical software, I examined the fixation and duration data to determine which, if any, variables influence numbers and durations of participant fixations. I considered expertise, topic, scaffolding level, trial, and gender as potential variables of interest and used truncated linear regression to model the data. In Table 5, I show several sub-population

Table 5

Sub-population Comparison Tests, Eye-tracking

Test #	within/ between	Sub-group	Visualization	Group	Visualization	Hypothesis
1	between	Expert	first viewed	Novice	first viewed	different
2	between	Expert	first viewed	Expert	last viewed	same
3	between	Novice	first viewed	Novice	last viewed	same
4	between	Expert	last viewed	Novice	last viewed	different
5	between	Expert	US	Novice	US	different
6	between	Expert	GS	Novice	GS	different
7	between	Expert	CS	Novice	CS	different
8	between	Expert	TS	Novice	TS	different
9	between	Expert	FS	Novice	FS	different
10	between	Expert	US	Novice	FS	same
11	between	Novice	US	Novice	GS	different
12	between	Novice	US	Novice	CS	different
13	between	Novice	US	Novice	TS	different
14	between	Novice	US	Novice	FS	different
15	between	Novice	GS	Novice	FS	different
16	between	Novice	CS	Novice	FS	different
17	between	Novice	TS	Novice	FS	different
18	Within	Expert	US	gender		same
19	Within	Novice	US	gender		same

comparisons that I investigated with statistical analysis based on hypotheses about the influences of those variables. Finally, multinomial logistic regression was used to evaluate relative probabilities of fixating on a given AOI using the same five dependent variables as potential influences.

The concurrent interviews were transcribed, coded, and analyzed in the same procedure as the clinical interviews, though gestures were not used for eye-tracking because the participant was asked to remain still for proper calibration and due to camera position. Scan paths were compared qualitatively and the order of feature viewing described in BeGaze™ to determine whether they differ. I also produced visualizations of AOI timing and order of entry.

In situ* eye-tracking.* Data was imported from iViewX™ into BeGaze™ software for analysis in a manner similar to that described in the laboratory eye-tracking condition. Eye-tracking data was overlaid onto the images of the participant's purported view of the Magic Planet as captured by the external camera. However, due to data quality issues, the quantitative data was highly suspicious and was not analyzed extensively for this dissertation. See Appendix 7 for a full discussion of *in situ* eye-tracking data issues. Interview data were transcribed and coded similarly to the previous two experiments.

Triangulation of Methods

The final research question of this project deals with the use of diverse methodologies to answer the same research question at different levels of detail. To that end, each experiment will attempt to determine accuracy of participant performance at judging a) the main idea of the data visualization, b) the meaning of the colors used in the visualization, c) the area of the highest data value in the visualization, and d) the time of year depicted. To coordinate the data, I looked for patterns in the data from each experiment, based on accuracy of performance, that indicated a) whether high-performing participants perform

equally well across scaffolding levels, b) whether both high-performing and low-performing participants perform equally well across topics, c) whether a certain type of scaffolding improves low-performing participants' accuracy better than others, and d) whether low-performing participants perform equally well as high-performing with all three levels of scaffolding included.

The qualitative set of data from these experiments will deal with strategies used by the various individuals and groups of participants, and will also be compared across methodologies to shed light on the ability of these methods to create a more complete picture of differences that exist among groups and individuals. By comparing verbalized strategies from the clinical interviews with perceptual strategies from the eye-tracking experiments, I determined what high-performing visualization users have internalized or automated, whether those strategies are conscious, semi-conscious, or unconscious, and, to an extent, how they arrived at those strategies, for example through training or because of fundamental individual differences in capability.

Methodological Limitations

There are several limitations to the methods used here. First, because the experiment was designed to counterbalance topics and scaffolding across participants with the aim of equal and repeated data collection for all, there is a possibility that data saturation was not reached on particular questions. This was compounded by the randomization of stimulus presentation order; some participants may have seen the scaffolded titles before the unscaffolded. This is especially the case for the question about the main idea in the unscaffolded versions of the visualizations; participants who saw the scaffolded versions first may have applied those titles to subsequent visualizations, precluding a true answer about what the main idea of the unscaffolded visualization seemed to be in the presumed absence of context. This could be served better for visualization design purposes by prototyping just the proposed color schemes with a larger

sample of potential users to probe the relationship between culture and color expectations for main idea more fully.

The presentation of so many versions of the same topic with the same data certainly might have allowed the experts to recognize the same patterns more frequently than not. This could have been the case with more novices who either did not voice such speculation or unconsciously applied previous visualization information to subsequent stimuli, causing results to be skewed more correct than might be true of a more naive or naturalistic population where visualizations are viewed singly.

The semi-structured interview format meant that the data was not collected identically for each participant, topic, or scaffolding level, limiting the quantitative statistical power of the correct answer tests. However, the in-depth nature of the explanations was balanced with the correctness per version of scaffolding and the combination outweighed the statistical power consideration here. The interviews also did not use exactly the same wording for each question, which could have biased some participants' answers differently than others. Finally, the clinical interview format might not have been used fully to probe explanations to their absolute limit in the interest of time and presenting multiple visualizations rather than single versions to each participant. Further research might include an open-ended interview on fewer versions of the visualizations.

For the eye-tracking in the laboratory, all but two participants had seen the same sorts of visualizations and were asked similar questions, in the interview previously, though with a different topic. All attempts were made to allow sufficient time (at least three weeks in between experiments) for learning effects to be minimized from one experiment to another. However, spontaneous looking in the eye-tracking condition was perhaps not entirely spontaneous, especially on the part of the experts. Also, given the semi-structured nature of the accompanying interviews during eye-tracking and the open-ended questions, trial

length differed within and between participants. Finally, despite the presence of control visualizations interleaved randomly with the experimental visualizations in this experiment, and experimental visualizations that depicted different seasons of the year, visualizations were still relatively similar in this experiment, and thus, learning effects may have occurred in the eye-tracking. Data compared in the first trial across participants, and from first to last trial, should give insight into these effects.

Participant populations were not identically balanced by gender, but rather were chosen to be roughly similar to the demographics of the population. All the expert participants also appeared to be older than all the novice participants except one, which means that I cannot rule out the effects of age on performance, rather than expertise from disciplinary training. On the other hand, the novices, mostly undergraduates, may have made up for this lack of age experience by being closer to formal training in pattern recognition and graphical data inference tasks such as the ones posed in this experiment. The general public who does not work frequently with such visualizations may perform worse than undergraduates for that reason. The data presented here cannot answer that question, although the general appearance of the *in situ* population showed a similar level of performance to the novices.

Finally, *in situ* eye-tracking did not allow for participants to be presented with multiple versions of the same task. Participants also did not see the full globe visualization that laboratory participants saw. The stand-alone eye-tracker also limited the examination of the full exhibit, including kiosk and 360-degree affordance of the spherical screen that visitors could normally walk entirely around. For full discussion of the *in situ* eye-tracking constraints, see Appendix 7. Also, interviews with participants in the *in situ* condition did not allow for extensive probing of “How did you know” and instead simply collected answers to main idea questions.

Interview Results

This chapter will discuss the qualitative and quantitative results of the interviews, both the semi-structured interview experiment and the interviews that accompanied the eye-tracking experiment. I begin with a discussion of the participants' backgrounds and experience with academic and some relevant informal science, as well as their self-reports of comfort with the visualization meaning-making task, as asked after the clinical interview.

First, I report the accuracy of participants' answers, segregated into discussion by clinical interview and eye-tracking interview. Scores are broken down into novice and expert groups and then by topic and scaffolding level as well. Next, I discuss the findings from qualitatively coding the interviews to discuss the how's and whys behind particular answers and shed light on differences in score accuracy among participant groups, topics, and scaffolding levels. I cover the codes I expected going into the experiment, with a specific section on the use of the elements of the visualizations, then cover the codes that emerged from the participants' answers. Due to the similarity of the questions asked in both interviews, the results of coding are discussed together here. Then I report the scoring from the *in situ* eye-tracking interviews, and conclude with sections discussing how novices fail to, and experts succeed at, putting together all of the pieces of information they gather from the visualizations.

Participant Characteristics

General background. Novice participants ranged in college experience from none, either entering in the fall or having completed only vocational training after high school, to starting their fourth term as an undergraduate at the time of the clinical interviews in Summer or Fall 2012. They reported a range of majors,

and were approximately 65% female by presentation and reported first name. See Table 6. For the eye-tracking portion, the novice participant subset had

Table 6

Novice Clinical Interview Participant Characteristics

Years in College		Majors	
No college	1	Psychology	3
Before freshman year	1	Exercise and Sport Science	2
In first term	6	Business/International Business	2
In third term	1	Undeclared/Exploratory Studies	2
Completed three terms	3	Fitness and Nutrition	2
In fourth term	5	Economics	1
		Math	1
		Pre-therapy and Allied Health	1
Presented Gender		Housing Studies - Architecture	1
Female	11	Anthropology	1
Male	6	Vocational Training	1
Total 17			

similar characteristics, with the exception of the gender balance, which was reversed at 60% male for the eye-tracking experiment. At the time of the eye-tracking experiment, all participants had matriculated, but none had completed more than their fourth undergraduate term at the time of the experiments in Fall 2012. All novice participants had at least three weeks between completing the interview and the eye-tracking portion. See Table 7. For the interviews, 25% of experts presented as female based on appearance and first name. For eye-tracking, two additional experts were recruited, one female and one male, for a total of two females and eight male experts participating in the eye-tracking portion. Experts ranged in years of experience beyond receiving the Ph.D. from six years to over 30 years. Of those participating in the clinical interview, 33% explicitly mentioned they work with satellite data. Specific research that the group reports working on varies. See Table 8.

Table 7

Novice Eye-tracking Participant Characteristics

Year in School		Major	
In first term	4	Exercise and Sport Science	2
In fourth term	6	Math	1
		International Business	1
		Fitness and Nutrition	1
		Psychology	1
		Exploratory Studies	1
Presented Gender		Housing Studies – Architecture	1
Female	4	Pre-therapy and Allied Health	1
Male	6	Economics	1
Total 10			

Table 8

Expert Participant Characteristics

Presented Gender		Research Sub-discipline		Research Geographic Scale	
Female	3	Physical oceanography	3	Estuarine/River	1
Male	9	Geophysics	2	Coastal/Continental Shelf	7
		Chemical Oceanography	3	Global	2
Total	12	Biogeochemical	2	Inside Earth/crust	2

Note. Numbers may not sum to 12 because not all participants reported answers in each sub-category. Experts recruited for the eye-tracking only were not asked these same questions.

Scientific background. Many of the novices reported at least three years of science classes in high school, typical of both high school graduation for the state and the requirement for the university that most attend, typically biology and chemistry. Seven reported high school physics; four had anatomy, three Advanced Placement biology. One each took marine systems, ecology, and Advanced Placement chemistry in high school. However, none of the novice participants had ever taken an oceanography-specific class beyond middle school, except the marine systems class, though that participant reported seeing

more of the visualizations shown here in a general science course in high school than in the marine systems class. Several were taking one or more science classes at the university: four were in or had taken college chemistry, two were in organic chemistry, one each were taking Introduction to Environmental science and microbiology at the time of the interview. None had participated in oceanography-related scholastic competitions such as the National Ocean Science Bowl. One had done a science fair project on giant kelp, and one did a high school report on Kiribati that touched on effects of climate change on the island nation. One had participated in an environmental youth corps in high school.

Few reported extensive experience with geography beyond a class or two in middle or high school, and as parts of some high school or college classes. One, who also reported a love for travel, “went far” in the geography bee in elementary and junior high school, participating in the state-level competitions. Interestingly, that participant (Brad) struggled mightily to identify the continents in the first visualization (extensive transcript below, p. 106).

Comfort with interpreting visualizations. Experts, despite their reported experience, say they’re only comfortable with interpreting data-based visualizations, averaging 2.05 on a Likert-type scale of 1-3, with 1 being “not comfortable” and 3 being “very comfortable”. They are especially uncomfortable with imagery outside their area of expertise or when given limited information. The novice average was 1.625, slightly less than “comfortable,” when asked specifically about their comfort interpreting spatially-based visualizations such as in this experiment, with similar comments about limited information and background knowledge. Novices were slightly more comfortable with interpreting charts and graphs of data more broadly (1.875 out of 3), but this was still lower than the middle score of 2 on the scale.

Expert frequency of visualization use. Seven of 12 experts (58%) reported working with spatial visualizations *very frequently* (3 on a scale of 1 - 3). The other five reported an average of 2 (*frequently*), with a range of 1 - 2.5. Four work extensively with satellite data, either global-, regional-, or coastal-scale. I computed an overall accuracy index by scoring the participants' answers based on the rubric described in Chapter 3 and attached in Appendix 3. The overall index for each participant represents the average score for all questions of the average scores on each individual question. This average of averages compensated for the semi-structured protocol, which resulted in different numbers of visualizations shown to and questions asked of participants. I compared expert accuracy indices with their self-reports of frequency of use of visualizations and comfort level interpreting visualizations, both on Likert-type scales of 1 (*not very frequently* or *not very comfortable*) and 3 (*very frequently* or *very comfortable*). There was a trend for experts who reported more frequent use with visualizations to have higher scores, but it was not exclusively the case. For example, Eric reported only occasional use of visualizations and a middle comfort level, but he had a better accuracy than six participants who reported using visualizations more frequently, four of whom had equal or higher self-reported comfort levels with the task. Scoring was in the range of 0 (incorrect) to 1.25 (correct and sophisticated) as described in Chapter 3. A score of 1.0 was fully correct. See Table 9.

Scoring

Overall novice versus expert.

Clinical interviews. Interview results showed a range of accuracies for individual questions in the entire participant population. Each average score represents the average for all participants of their individual average scores for

that question. The overall average accuracy, the average scores for each question averaged together, was 0.71, approximately halfway between a score of

Table 9

Experts' Overall Score Accuracy Index Compared to Frequency of Use and Comfort Level for Interpreting Visualizations

	Score Accuracy Index[a]	Frequency of Visualization Use[b]	Comfort Level Interpreting Visualizations[c]
Alice	0.57	2	2
Matt	0.75	2.5	2
Janet	0.84	2	1.5
Keith	0.85	3	2
Mark	0.85	3	[d]
Rick	0.86	3	3
Eric	0.88	1	2
Charlie	0.88	3	[d]
Lindsey	0.90	2.5	2
Jay	0.92	3	2
Brent	0.96	3	3
Ray	1.04	3	2

Note. Arranged in increasing order of score index.

[a] Scores range from 0 (*incorrect*) to 1.25 (*sophisticated)*, with 1.0 equaling *fully correct*.
[b] Likert-type scale of frequency of visualization use from 1 (*not very frequently)* to 3 (*very frequently*).
[c] Likert-type scale of comfort with visualization interpretation from 1 (*not very comfortable*) to 3 (*very comfortable*).
[d]These two participants were not asked about their comfort level due to oversight.

0.5, partially correct, and 1.0, fully correct. See Table 10. Overall, performance was best on the question of measurement unit, with an average score of 1.01 (1.0 was correct, 1.25 was the maximum indicating a "sophisticated answer"), and on the question of color meaning, at 0.90. It was worst on the question of

season, at only 0.21, and not much better for time span of the data represented, at 0.47.

Table 10

Accuracy in Clinical Interviews, All Participants (*N* = 29)

Question	Maximum	Minimum	Mean
Main Idea	1.25	0.30	0.79
Evidence	1.25	0.50	0.86
Color Meaning	1.25	0.46	0.90
Timespan	0.93	0.00	0.48
Season	0.75	0.00	0.21
Measurement Unit	1.25	0.17	1.01
Equatorial Pacific Condition	1.25	0.08	0.65
Is Equatorial Pacific Normal?	1.25	0.00	0.64
Grey Meaning	1.25	0.00	0.68
Equator Location	1.25	0.00	0.99
Overall	0.99	0.41	0.71

Note. Scores range from 0 - 1.25.

Novices and experts differed on accuracy on almost all individual questions, with experts surpassing novices except in the case of timespan, where novices performed slightly better (0.54 versus 0.39), though both accuracies fell in the range of “partially correct” to “incorrect”. See Table 11. Experts scored *correct* or *sophisticated* (≥ 1.0) on questions of main idea, evidence of the main idea in the map, color, measurement unit, and conditions at the Equatorial Pacific. They were nearly correct (≥ 0.8) on average for the normalcy of the Equatorial Pacific conditions and meaning of the grey coloring. Novices were, on average, correct only for the measurement unit and otherwise indexed 0.8 or lower, with worst performance in the same categories as the

experts, timespan and season. Overall, novices averaged solidly in the partially correct category (0.55), while experts on average were nearly correct at 0.83.

Table 11

Scoring Accuracy in Clinical Interviews, Novice versus Expert

Question	Novice (*n* = 17)			Expert (*n* = 12)		
	Maximum	Minimum	Average	Maximum	Minimum	Average
Main Idea	1.00	0.30	0.62	1.25	0.65	1.04
Evidence	1.00	0.50	0.76	1.25	1.00	1.13
Color Meaning	1.11	0.46	0.81	1.25	0.88	1.03
Timespan	0.93	0	0.54	0.80	0	0.39
Season	0.40	0	0.11	0.75	0	0.36
Measurement Unit	1.25	0.17	0.97	1.25	0.50	1.08
Equatorial Pacific Condition	0.79	0.08	0.39	1.25	0.88	1.04
Is Equatorial Pacific Normal?	1	0	0.47	1.25	0	0.91
Grey Meaning	1	0	0.55	1.25	0.25	0.86
Overall	0.67	0.36	0.55	1.00	0.56	0.83

Note. Scores range from 0 – 1.25.

Individuals varied, however, among each other and among questions and topics. Looking at the maximum scores, while no novices scored in the "sophisticated" range except on measurement unit, at least one novice scored correctly on at least one question in every category except season and Equatorial Pacific conditions. However, novices also generally had lower minimum scores than the experts as well.

Eye-tracking. Patterns of performance on eye-tracking interviews were similar to clinical interviews in the overall, novice, and expert populations.

Average scores were worst on the season question, followed by timespan, and best on the measurement unit question. See Tables 12 and 13.

Table 12

Scoring Accuracy on Eye-tracking Interviews, All Participants (N = 20)

Question	All Participants		
	Maximum	Minimum	Average
Main Idea	1.25	0.00	0.81
Evidence	1.13	0.50	0.68
Color Meaning	1.00	0.00	0.72
Timespan	1.00	0.00	0.54
Season	1.00	0.00	0.38
Measurement Unit	1.25	0.60	1.12
Equatorial Pacific Condition	1.25	0.50	0.90
Grey Meaning	1.00	0.50	0.68

Note. Scores range from 0 – 1.25. Equatorial normalcy question was not asked.

Scaffolding level.

Clinical interviews. Scoring accuracy in the clinical interviews improved on the main idea with increased scaffolding in the overall population (0.63 to 0.83) and almost by double (0.43 to 0.80) in the novices. This improvement not significant from the unscaffolded to fully scaffolded case in the overall population W_s (n_1 = 22, n_2 = 27) = 269, p = .56, but it was significant for the novices, W_s (n_1 = 16, n_2 = 16) = 70.5, p = .02. The experts' performance was not significantly different with scaffolding *Ws* (n_1 = 6, n_2 = 11) = 42, p = .34, remaining essentially completely correct (score of 1.0) on the main idea in both cases. See Table 14. Expert performance on the timespan question, where the time span was particularly put in the title, also was not significantly different between the

Table 13

Scoring Accuracy on Eye-tracking Interviews, Novice versus Expert

Question	Novice (n = 10)			Expert (n = 10)		
	Maximum	Minimum	Average	Maximum	Minimum	Average
Main Idea	1.25	0	0.58	1.25	0.80	1.04
Evidence	0.63	0.50	0.52	1.13	0.50	0.87
Color Meaning	1.00	0	0.64	1.00	0.10	0.81
Timespan	1.00	0	0.49	0.90	0.33	0.60
Season	0.50	0	0.23	1.00	0.00	0.54
Measurement Unit	1.25	0.60	1.12	1.25	1.00	1.12
Equatorial Pacific Condition	1.00	0.50	0.67	1.25	1.25	1.25
Grey Meaning	1.00	0.50	0.55	1.00	0.50	0.89

Note. Scores range from 0 – 1.25.

unscaffolded and fully-scaffolded versions, W_s (n_1 = 6, n_2 = 7) = 13, p = .18. On this question, novices again significantly improved W_s (n_1 = 13, n_2 = 16) = 47, p = .003, and the overall population also significantly improved with scaffolding W_s (n_1 = 20, n_2 = 22) = 111, p = .001. Otherwise, the overall population did not improve on other questions. Novices had a small improvement (0.68 to 0.89) on the color meaning question and a possible decline on the evidence in the map (0.81 to 0.69) question. This last result could be reflective of the general lack of cultural familiarity that seemed to accompany the scaffolded color scales and the confusion that remained about the main idea.

Eye-tracking. When scoring was broken down by scaffolding level, overall accuracy improved between the unscaffolded (US) and fully-scaffolded (FS) versions for the main idea (0.7 to 1.0), evidence of the main idea in the map (0.56 to 0.94) and timespan (0.26 to 0.88) questions. See Table 15. In the case of the main idea question, this effect was accounted for almost entirely in

Table 14

Scoring Accuracy on Clinical Interviews by Scaffolding Level, All Participants (N = 29), All Novices, and All Experts

Question	All Participants				
	No Scaffolding	One Level Scaffolding			Full Scaffolding
	US	CS	GS	TS	FS
Main Idea	0.63	0.62	0.71	0.90	0.83
Evidence	0.88	0.84	0.81	0.81	0.72
Color Meaning	0.78	0.96	0.84	0.86	0.92
Timespan	0.32	0.28	0.33	0.83	0.77
Season	0.15	0.13	0.21	0.18	0.15
Measurement Unit	1.03	1.01	1.04	0.85	0.95
Grey Meaning	0.58	0.78	0.73	0.50	0.45
Novice Participants (*N* = 17)					
	No Scaffolding	One Level Scaffolding			Full Scaffolding
	US	CS	GS	TS	FS
Main Idea	0.43	0.48	0.57	0.81	0.80
Evidence	0.81	0.75	0.79	0.67	0.69
Color Meaning	0.68	0.91	0.73	0.82	0.89
Timespan	0.34	0.32	0.34	0.83	0.81
Season	0.04	0.08	0.06	0.17	0.13
Measurement Unit	0.97	1.03	0.98	0.83	0.90
Grey Meaning	0.43	0.70	0.58	0.63	0.42
Expert Participants (*N* = 12)					
	No Scaffolding	One Level Scaffolding			Full Scaffolding
	US	CS	GS	TS	FS
Main Idea	1.02	0.96	1.07	1.15	0.93
Evidence	1.13	1.13	1.00	1.25	1.00
Color Meaning	1.00	1.04	1.03	1.06	1.00
Timespan	0.25	0.20	0.29	0.82	0.67
Season	0.39	0.23	0.56	0.22	0.29
Measurement Unit	1.16	0.96	1.20	1.25	1.13
Grey Meaning	0.75	0.85	0.94	0.25	0.50

Note. Scaffolding increases left to right, with CS, GS, and TS considered equal as one level of scaffolding each. Scores range from 0 – 1.25.

Table 15

Scoring Accuracy on Eye-tracking Interviews by Scaffolding Level, All Participants (N = 20), All Novices, and All Experts

Question	All Participants				
	No Scaffolding	One Level Scaffolding			Full Scaffolding
	US	CS	GS	TS	FS
Main Idea	0.70	0.73	0.75	0.92	1.00
Evidence	0.56	0.70	0.9	0.75	0.94
Color Meaning	0.77	0.74	0.6	0.74	0.75
Timespan	0.26	0.26	0.43	0.84	0.88
Season	0.36	0.29	0.32	0.45	0.40
Measurement Unit	1.15	1.08	1.15	1.02	1.15
Grey Meaning	0.67	0.70	0.88	0.58	0.67
Novice Participants, (*N* = 10)					
	No Scaffolding	One Level Scaffolding			Full Scaffolding
	US	CS	GS	TS	FS
Main Idea	0.43	0.43	0.43	0.78	0.93
Evidence	0.50	0.50	0.75	0.50	
Color Meaning	0.67	0.67	0.50	0.70	0.69
Timespan	0.10	0.28	0.50	0.70	0.78
Season	0.11	0.25	0.30	0.28	0.15
Measurement Unit	1.13	1.04	1.16	1.00	1.16
Grey Meaning	0.50	0.63	0.50	0.50	0.50
Expert Participants, (*N* = 10)					
	No Scaffolding	One Level Scaffolding			Full Scaffolding
Main Idea	US	CS	GS	TS	FS
Evidence	0.98	1.03	1.08	1.09	1.08
Color Meaning	0.75	1.00	1.00	1.13	0.94
Timespan	0.83	0.81	0.69	0.79	0.81
Season	0.44	0.25	0.35	1.00	1.00
Measurement Unit	0.61	0.33	0.33	0.64	0.72
Grey Meaning	1.19	1.15	1.14	1.06	1.13
Main Idea	0.88	1.00	1.00	0.75	1.00

Note. Scaffolding increases left to right, with CS, GS, and TS equal. Blank cells indicate no data collected. Scores range from 0 – 1.25.

improved novice performance, from 0.43, partially correct, to 0.93, fully correct. The experts remained essentially unchanged at 0.98 to 1.08. However, in the case of the time span, both groups improved. Novices improved from a 0.10 (completely incorrect) to 0.78 (edging toward fully correct), while experts improved from 0.44 to 1.0, partially to fully correct. No real pattern can be seen in either sub-group for the evidence of the main idea in the map question, however. This may be due partially to the absence of any answers to this question for the novices in the fully-scaffolded case, but with the exception of the geographic scaffolds case, neither of the other single scaffolds actually improved novice performance. Again, this is most likely due to a small number of observations; overall in both the clinical interviews and eye-tracking interviews, this question was infrequently asked. Scaffolding did not seem to affect performance clearly on any of the other questions asked.

Scores by topic.

Clinical interviews. In the clinical interviews, performance did not vary much across topics in the overall participant group, but did within sub-groups. While experts showed similar performance across topics, novices generally performed best on the SST visualizations, and worse on chlorophyll and anomaly visualizations, with some variation across questions. See Tables 16 and 17.

Eye-tracking. In the eye-tracking interviews, accuracy was similar, generally greatest on the SST visualizations, or SST and chlorophyll were equal, and worst on the anomaly visualizations, with some exceptions. See Table 18. Performance on timespan for all participants was worst on the chlorophyll visualizations, and performance was generally equal on the measurement unit and meaning of grey in the visualizations for all topics.

Again, different patterns for experts and novices emerged. Expert performance on the main idea question did not differ greatly among topics,

Table 16

Scoring Accuracy on Clinical Interviews by Topic, All Participants (N = 29)

Question	All Participants		
	SST	SST anomaly	Chlorophyll
Main Idea	0.89	0.73	0.79
Evidence	0.91	0.86	0.83
Color Meaning	0.99	0.75	0.99
Timespan	0.54	0.47	0.43
Season	0.18	0.14	0.25
Measurement Unit	1.07	1.12	0.86
Equatorial Pacific Condition	0.72	0.57	0.56
Grey Meaning	0.63	0.55	0.72

Note. Scores range from 0 – 1.25.

Table 17

Scoring Accuracy on Clinical Interviews by Topic, Novices versus Experts

	Novice Participants, (*N* = 17)			Expert Participants, (*N* = 12)		
	SST	SST anomaly	Chlorophyll	SST	SST anomaly	Chlorophyll
Main Idea	0.77	0.49	0.58	1.05	1.01	1.08
Evidence	0.82	0.50	0.80	1.13	1.25	1.00
Color Meaning	0.95	0.54	0.95	1.05	1.02	1.05
Timespan	0.63	0.52	0.39	0.38	0.45	0.50
Season	0.16	0.09	0.07	0.20	0.50	0.54
Measurement Unit	1.04	1.15	0.69	1.11	1.14	1.08
Equatorial Pacific Condition	0.54	0.32	0.31	1.04	0.92	1.13
Grey Meaning	0.56	0.38	0.63	0.81	0.63	0.88

Note. Scores range from 0 – 1.25.

though it was marginally lower (0.2 difference) for SST anomaly. Novice accuracy was nearly one full category lower (0.47 difference) on the anomaly main idea question than the other two topics. Experts could provide evidence in

the map for the main idea in two of three topics by referring to specific patterns or features, while novices provided only partially correct information for all three, generally noting that the data was only located on the ocean parts of the map.

Table 18

Scoring Accuracy on Eye-tracking Interviews by Topic, All Participants (N = 20), All Novices, and All Experts

Question	All participants		
	SST	SST anomaly	Chlorophyll
Main Idea	1.01	0.60	1.03
Evidence	0.77	0.50	0.72
Color Meaning	1.00	0.45	1.00
Timespan	0.66	0.56	0.38
Season	0.42	0.31	0.54
Measurement Unit	1.23	1.15	0.95
Is Equatorial Pacific Normal?		0.90	
Grey Meaning	0.75	0.65	0.76
Question	Novice Participants (*N* = 10)		
	SST	SST anomaly	Chlorophyll
Main Idea	0.85	0.38	0.98
Evidence	0.56	0.50	0.50
Color Meaning	1.00	0.40	1.00
Timespan	0.75	0.53	0.10
Season	0.25	0.21	0.25
Measurement Unit	1.25	1.18	0.80
Is Equatorial Pacific Normal?		0.67	
Grey Meaning	0.50	0.58	0.50
Question	Expert Participants (*N* = 10)		
	SST	SST anomaly	Chlorophyll
Main Idea	1.12	0.92	1.13
Evidence	0.90	0.50	1.00
Color Meaning	1.00	0.53	1.00
Timespan	0.60	0.59	0.60
Season	0.53	0.46	0.63
Measurement Unit	1.21		1.07
Is Equatorial Pacific Normal?		1.25	
Grey Meaning	1.00	0.75	1.00

Note. Counts are not equal for each cell. Blank cells indicate no data collected. Scores range from 0 – 1.25.

For both groups, color meaning was similarly partially accurate in the anomaly case, and similarly correct for the other two topics. Experts were essentially partially correct for time span for all three topics, as were novices for SST and anomaly. However, novices were completely incorrect for the timespan of the chlorophyll visualizations. Novices also struggled slightly with the measurement unit in the chlorophyll visualizations, but on other topics they were correct, as were the experts in the two of three cases for which they were measured. For the question of equatorial pacific conditions, novices were worse than experts, generally describing the entire area as similar, failing to note the east-west variation at the equator. More evidence of this is found in the qualitative data described below. Novices accounted for much of the effect of the less-than-correct scoring in the case of the meaning of the grey; without fail they recognized the land in dark grey but many did not distinguish between the dark grey and the light grey areas of ocean without data unless prompted further.

These findings about topic are somewhat surprising, as had hoped to pick visualizations and color scales, at least in the scaffolded case, that provided fairly easy meaning-making even for non-oceanographers. However, this evidence supports the need for testing every topic individually with users and at the very least, not assuming that carrying over what works in one visualization will be immediately usable by every public audience. It also justifies the use of randomization of assignment of participants to topic and presentation order.

Overall, due to the semi-structured nature of the interviews, the participants were not all asked the same questions on all versions of the visualizations. In the clinical interviews portion, due to the depth with which some experts, especially, answered the questions, they were not always shown more than two versions of each visualization. Thus, the numbers of scores in each cell in the preceding discussion were not consistent. Though interviews accompanying the eye-tracking were more structured, it still varied somewhat

due to the detail participants answered within answers to particular questions. However, averages were computed by participant for each question, and then those averages were averaged together to make the “expert”, “novice” or “all participants” averages reported above to alleviate this issue to the extent possible.

Despite these limitations, in general, all of the findings on scoring accuracy match well with the qualitative data described below; scoring accuracy varied between experts and novices, among scaffolding levels, and among topics, and participants’ open-ended answers reflect these patterns. The ranges of scores also provided opportunities for probing questions to understand better where the discrepancies in meaning-making occurred and why.

Analytic Coding – Expected Codes

As is typical of open coding, I approached the analysis with some initial codes in mind. Particularly, from a constructivist perspective, I was interested in prior knowledge and experience as elements of meaning-making and evidence of confusion or insight as evidence of disequilibrium. From a sociocultural perspective, I was interested in references to specific elements of scientific thinking and culture such as jargon, use of scientific language, or reference to scientific or visualization practices. I looked for references to salient parts of the visualizations or things that “stuck out” as evidence of perceptual influences. Based on previous work (Phipps & Rowe, 2010; Rowe et al., 2010), I also sought evidence of use of particular scaffolded visualizations elements.

Open coding assessed commonalities across participants and within and across expertise groups. Generally, open coding looks for themes reported by multiple respondents about similar reasoning, use of evidence, or sources of external information, aiming to find commonality. However, as these visualizations are designed to transmit information accurately (Rowe, 2002), if even one user fails to obtain the intended meaning, the visualization is deficient

in its ability to serve that purpose. Thus, I am also interested in instances of single or only a few participants reporting particular areas of difficulty. Examining these cases reveals ways to encourage mutual understanding among audiences and designers by assisting accurate transmission through visualizations.

Constructivist-derived codes of prior knowledge and experience formed the first set of codes examined for meaning-making. Particular specific codes about visualization or data-specific background also emerged from these.

Prior knowledge. All but one scientist reported prior knowledge that they gained in graduate studies or beyond as important elements of their making meaning from the visualizations. The one who did not was an eye-tracking-only participant who was not asked as many probing questions as clinical interview participants. Graduate school as the source of background brought to bear for meaning-making precludes the novices from the opportunity to obtain much of the same prior knowledge or at least to learn to use it in the specific context; they had not yet completed undergraduate studies, let alone graduate work. One expert noted that some of the information about seasons, though it generally came from earlier school work, was only put into oceanography context through graduate school:

> Lindsey: Well, I mean, thinking of it in an oceanography context, probably, and temperature cycles, probably from grad school, but, you know, it's all making sense from basic information on seasons learned from a very early age. (clinical interview).

Novices did report learning about units of measure, seasons, and temperature in elementary or middle school. However, consistently, novices judged visualizations with roughly equal north-south distributions of temperatures (symmetric about the equator) as spring or fall, if they attempted to determine the season at all. Visualizations were in fact from January in the clinical interviews and while symmetric in temperature distribution, represented symmetry typical of an *ocean* winter due to heat capacity of water that makes it warm up and cool

down more slowly than air. Surprisingly, experts and novices used that same incorrect reasoning of symmetry about the equator for judging the visualizations to represent one year (or more), if one month wasn't in the title or wasn't in the titles they'd seen previously. This faulty reasoning could explain much of the poor performance on the scoring accuracy on the time span and season questions. It may be an indication of the lack of familiarity with the particular topic (temperature, as many of the satellite oceanographers indicated they worked with chlorophyll), and certainly points to a need to include information explicitly even for other scientific audiences.

When asked about the other "stories one could tell about the visualizations," novices struggled, especially when they didn't grasp the full understanding of even the scaffolded titles. Allison says she would describe the visualization to someone else as "Just showing productivity in the ocean of something, I know that it's darker in the middle parts as compared to closer to the land. So that must show some biological effect of that, other than that I don't know." More typical of experts was Brent, who said, "There's a thousand different answers to that question," then referenced several different geographical phenomena within the chlorophyll visualization, such as the "the productivity signature of the Patagonian shelf, that shows up really strong ... the persistence of high productivity around the continents, that shows up really strong." Overall, experts suggested ocean circulation for the SST visualizations, El Niño almost universally for the SST anomaly visualizations, and eddies, ocean carbon dioxide uptake, and fisheries for chlorophyll. Only two novices suggested global warming for the SST visualizations, one mentioned El Niño in connection with SST, and two hesitatingly suggested the idea of fisheries for chlorophyll, though one expressed that the visualization was depicting ocean mercury levels and thus thought it should show where not to fish.

Experience with visualizations. Novices occasionally mentioned having seen visualizations similar to the stimuli here in the news, in a certain class, or with a parent. One participant reported having seen them in science classes, on TV news and on the internet, in the context of El Niño and hurricane reports. Other places these visualizations had been shown were weather forecasts, geography class, global warming or other science studies, or from a particular teacher. None of them, however, reported extensive experience with these, and five said they had never seen an visualization like this.

No experts reported they had never seen an visualization like this. In fact, scientists say “do you know how many times I’ve seen this image?” (Janet, clinical interview) or “part of [my job] is to develop algorithms that will help us to get better estimates of [this data] from satellites,” (Ray, clinical interview) or “I’ve made those measurements ... I’ve been working with this data for 30 years” (Charlie, clinical interview). Many also produce imagery to visualize their data, spatially-based or not, so are familiar with choices made in representation. Nine of 12 in the interviews reported teaching about oceanography and teaching about this sort of visualization as part of that at some point in their careers. At the least they were familiar with them from graduate school or seminars and conferences, to which novices did not report any access.

Twelve of the 14 scientists across both experiments commented spontaneously on the appropriateness or quality of the color scales, units, or titles, based on their own use or production of visualizations, such as when Janet noted in the clinical interview: “I'd be careful with those colors maybe just a bit because the darkest red and the purple are starting to look a little bit similar.” Jay (clinical interview) said, “This is actually a color scale I use a lot and I kind of like because it emphasizes positive and negative anomalies with a sort of drastic color change from red to blue.” Many of their critiques tended to the negative, such as Matt’s gentle rebuke:

> “Alright, I can guess as to what that means. It’s not worded very well, but let’s call the long term averages there, and then what the ... I don’t even

> know what this means. (sigh/laugh) You did that intentionally right? This isn't a real slide that somebody actually produced." (clinical interview)

One expert commented on aspects of the visualization quality 27 times in the clinical interview alone.

While 12 of the novices commented on the quality of the visualizations, no one commented more than seven times in the interview, and six of the 12 only commented once. Their critiques also tended to be a bit vaguer, "These are pretty colors," (Shelby, clinical interview), blame themselves instead of the visualization designer, "even though I'm sure SST does stand for something very specific I don't know the meaning of it," (Samantha, clinical interview), or simply forgive more. Gina commented on the same title as Matt, but put it this way, "Yeah, this one, this one's kinda hard 'cause it's using average twice, in this, in the label. And that's confusing to me." (clinical interview).

Extensiveness and detail of answers. Given only an visualization title (or even one that confuses them) and a color key, oceanographers can talk at much greater depth and length about a global visualization than novices, whether or not the experts directly or only tangentially work with satellite data or even the ocean topic presented. When presented with the first visualization, one of the researchers who studies the interior of the Earth rather than the fluid ocean itself, answered this to the question of the main idea:

> It's a sea surface temperature anomaly on a global map. Probably derived from satellite data. And it's uh, got a scale in um, degrees centigrade, going from plus or minus five degrees, and the scale looks pretty appropriate to the range of values in the data set with a few points that are really at the far ends of the scale, and enough resolution to um, show the major variations on a global basis. (Alice, clinical interview).

On the other hand, when there are gaps in their knowledge, experts can often consider several possible explanations and are aware of what extra information they would need. When asked about time span, Rick said:

> Uh again, it's difficult to say. And I think again this is probably ... you know. Yeah, it's difficult to say. I think potentially we're missing a key piece of information in this image as well in the other image and that is the Northeast pacific, uh, Northeast Atlantic, the region between Newfoundland, Great Britain, and Iceland. Because that region in terms of time, we know that region has a well-defined spring bloom in April, May of each year, of each of each year. The fact that we can't, for example, if that was observable and showed high concentrations of chlorophyll we would be able to imply that it was a certain time of the year. Since we don't have information there it's difficult for me again, I don't know enough about the southern ocean in other regions to say unequivocally when this is. But, I'm thinking still it's probably a long term average, the annual average, perhaps. (clinical interview).

At the very least, experts can say what information they might need, such as Alice's answer to a question about season, "I would need more information about what's being represented: where the data come from." Novices typically did not hazard educated guesses, and simply answered, "I don't know." In fact, 14 of 17 (82%) novices answered "I don't know" with no more explanation at least once in the clinical interviews, while only four of 12 (25%) experts did, though similar numbers of each group, six experts and seven novices, answered "I don't know" at least once for the eye-tracking. This may be explained by the lack of opportunity in the eye-tracking interviews to expand upon answers when such an answer would have been probed in the clinical interviews to determine if there was any sense of an answer. In the clinical interview, I only coded an answer as "don't know" if the participant could not offer any suggestions at all as to the answer; otherwise it was coded as either an area of confusion or simply a correct answer where the participant was not fully confident.

However, the rates for the frequencies of "don't know" answers were very

different: experts in the clinical interview were coded as “I don’t know” 23 times, or 5.75 per expert who were coded thus, and 1.9 times per expert overall. Novices, meanwhile, had 105 instances in the clinical interview, or 7.5 per novice coded this way and 6.2 per novice overall. In the eye-tracking case, there were 12 instances for experts, a rate of two per expert coded with at least one “I don’t know” or 0.2 per expert overall, versus novices at 45 total, 6.4 per coded novice and 4.5 per all novices.

Satellites. Scientists may work with satellites, produce or teach about visualizations. Nine scientists spontaneously suggested these are satellite visualizations, and the other three correctly identified them as such when asked. Many even went so far as to say there was no other way to get the data, giving rationale from the visualization why it is so based on prior knowledge and experience of the extent of coverage and the data gaps and reasoning for those gaps such as ice cover, darkness, or cloud cover. Some novices refer to manual data collection, as when Shelby suggested “take a thermometer and stick it in the ocean” (clinical interview). Those that don’t only had partially-formed conceptions of how “technology” could handle the problem, as Samantha describes:

> I’m assuming they just didn't go out there and do it manually, so I'm thinking they obviously have some sort of technology to do that for them, the researchers, but I don't quite know, I'm trying to think of how they would get some thermometer of sorts to take a temperature probably around the exact same time every day and then eventually compile that all into one ... I have no idea how that would actually... (clinical interview)

Some reply vaguely about taking a “bunch” or “ton” of readings over time and space, from as many as 50 locations in the ocean and several times a day. Those who guess or respond “satellites” don’t know thoroughly how they work. Virginia: “Um. Hm. probably from Satellite. Using like infrared or something, really tech-y (laughs) ...I'm not sure if the satellite can really cover [the whole Earth in a day]. Hm” (clinical interview). This gap exists despite the fact that many younger

participants have “grown up” with satellites as part of the popular media surrounding them, while several of the oceanographers noted that satellite data of this type did not exist while they were in graduate school.

Yet again, scientists were not perfect in their understanding of the visualization. Despite recognizing the data was derived from satellites, they were unable to accurately determine the time span represented in most cases without scaffolding. They tended to assume the data was a longer average, either a year or multiple years, based on the symmetry as described previously. A few suggested something on the order of a day or a few days, and one even thought it was a snapshot, indicating they either didn’t know or didn’t talk about the need in those short-term cases for extensive interpolation due to satellite paths, despite their reports that the reason for their assumption for the satellite basis of the data was the extensive coverage obtainable only by satellites. Novices, however, had more tendency to assume shorter time periods of a snapshot or a day, with little reasoning to back up their assumptions, although three made reference to the idea that the visualizations were unchanging, such as Virginia, when asked in the clinical interview why she thought the visualization depicted just an instant, “Um, ah, well, the picture's not really changing very much, so it can't really be over time I guess?”

Geography. Novice participants were less familiar with global geography than were the experts. Both groups were able to accurately point out the equator when asked to do so, with an overall average score of 0.93, and they all described knowing the equator was in the “middle” or “center” from a very young age. However, novices used less specific names for places, generally naming things on a continental scale, though a few used countries and some United States state names. Experts, on the other hand, named many ocean features, including Baffin Bay, Labrador Sea, the Gulf Stream, South Pacific gyre, Pacific warm pool, the Kuroshio and Humboldt currents. One novice admitted making

use of the geographic labels when they were offered to name the specific ocean basins: “Notice how I am identifying them now by their names, because I didn't know that this was the Indian Ocean and Atlantic Ocean before. Because now, like, 'cause now that they're labeled I can say their names with it instead of just pointing.” (Samantha, clinical interview). Three others explicitly noticed the labels, though particular geographic place names were never asked for directly. Three experts remarked on the presence or lack of labels.

Some novices reported this was a different depiction of Earth than they were used to. In the eye-tracking interview, novice Eden talked about how unfamiliar it was to have the Western Hemisphere on the right side of the globe:

> Eden: It's the reversed image from the first one, so not backwards, it's turned. So now Asia and Africa is on the left and North and South America is on the right.
> Researcher: And which first image do you mean it's reversed from? Or switched from?
> Eden: I think it's just switched from normal maps that I see.

Most could orient to the visualization given that it was a globe, but some did not do so immediately. When asked about the main idea of the visualization, in the clinical interview, Emma replied, “How much heat, like, is being, is given out in different areas of the world. Oh, those are the oceans, like how the temperature of the water is in different areas.”

The novice, Brad, who indicated a lot of experience with geography in elementary and middle school through participation in the state geography bee, was extremely confused for much of the first fifteen minutes of his interview.

> Researcher: Okay. Great. So which coast of which continent do you think has the highest value?
> Brad: (pauses [1:42 - 1:51]) Continent? But, like this isn’t up to date contin-..I mean continent panoram.
> Researcher: How do you mean?
> Brad: Like it’s not, this is like an old continent view.
> (About eight minutes later, during which time additional questions have been asked and answered)

Researcher: How about what season of the year do you think this is representing?
Brad: {~20 sec pause} Winter?
Researcher: OK. Winter in which hemisphere?
Brad: The continents look normal. Did you do that on purpose?
Researcher: What's that?
Brad: To make the continents not the normal iconic looking continents?
Researcher: Instead of?
Brad: Mmm, I feel like it's... oh man. OK. Ask the question again? {Laughing}
Researcher: Do you want to change your answer? Did something change? Did you realize something?
Brad: No, I should stay with my first guess, because that's always my best guess.
Researcher: OK. So you said winter, but which hemisphere is it winter?
Brad: So that's why I want to change my answer, because that means, because certain continents experience difference seasons at different times. It's not a whole panoramic, like multiple continents then.
Researcher: OK. So how would you change your answer?
Brad: Well that's not the equator then, but, now I'm confused. Oh what the heck?
Researcher: That's OK. You can be confused, but tell me what you're thinking and I can try to ask the question differently.
Brad: Well because when you say what season is it in this hemisphere, well like, the northern hemisphere is experiencing a different season than the southern hemisphere, and that's like why birds migrate. Cause they go to where ever it's warmer. So that means this isn't multiple continents, it's like just one.
Researcher: OK. So, what continent did you think, what did you think this was depicting then? Maybe what continent?
Brad: I don't know what...
Researcher: OK. How about this, let's step back. Do you recognize anything on this as familiar continents or something?
Brad: To me it looks like the formation of Australia, and then the like all those tiny islands.
Researcher: OK.
Brad: Like Indonesia and all that. But I honestly don't really know.
Researcher: OK. So nothing else familiar in terms of...
Brad: Yeah but like it looks like an Australia to me, but that's just off my best educated guess.
Researcher: OK. Then let's just stick with winter. So tell me why you think it's winter then.
Brad: Well I was going to say it's winter because of the blue markings cause it's like colder up in the north then it's warmer down in the south

during the winter. And then it's pretty much warm everywhere in the northern hemisphere when it's during the summer.
Researcher: OK. So it looks like winter to you?
Brad: Yeah.
Researcher: So, what do you think the grey is depicting in the image? What do you think the grey is showing here?
Brad: Oh, fail. Oh my god, not... Whoa, now my brain totally registers. There is South America, there is northern America, there is Australia, and there is...Oh, and that's the ocean's climate.
Researcher: Ok, so, it didn't seem like that before? It was confusing?
Brad: I was looking at just the color and not the grey. (laughs)
Researcher: That's ok, this is exactly what I'm trying to find out; what people think of these, so...
Brad: That's so weird. (laughs) Maybe if it wasn't grey it would be a little more pin point. And the, there is Antarctica, but (pause) Yeah this is definitely winter... (pause) and this, this is the Northern hemisphere right here, Well, there is multiple hemispheres.
Researcher: Sure. So Northern versus southern. Which Hemisphere is winter then?
Brad: (pause) um (pause) I think the Northern is experiencing the colder climate. Because the dark blue is more down and if it was, like, summer or springtime the green would be more apparent and even the yellow and orange, cause, California is experiencing cold...But yeah, this is the temperature of water, like, the ocean. And that's Australia and that's Indonesian islands, that's New Zealand, and there is Africa and Asia. So yeah, cause there's definitely multiple continents and it's not an old depiction of the world. (laughs)

The second exchange ends after six and a half minutes. Here, the participant was likely experiencing a physical perceptual confusion of the foreground and background of the images (Koffka, K., 1935), as with this image of Rubin's vase. Is it showing a two faces or a vase? See Figure 3.

Visualization Element Use

In the examples above, participants explicitly used prior knowledge and experience to make sense of the experimental stimuli, an important part of learning and meaning-making from the constructivist's perspective (Roschelle, 1995). On the other hand, use of specific elements of visualizations is an

example of an enculturated skill, generally learned either through schooling or professional work. Previous research (Phipps & Rowe, 2010) indicated that some viewers of visualizations do not make use of the supporting elements of the visualization, namely the key or title, when attempting to make meaning of

Figure 3. Example of figure-ground illusion, “Rubin’s Vase.” Looking at the yellow in the left-hand image, one sees a vase. Looking at the light blue, however, one sees two faces. John Smithson 2007 / Wikimedia Commons / Public Domain (2006)

visualizations of data. Scientific experts have extensive experience in reading these visualizations, however, reflected in their use of these elements for meaning-making and to some extent, their use of them simply for confirmation of what they assumed based on the patterns in the visualizations. Further, even when they reported making use of the elements, the novices did not always arrive at correct answers.

Participants were not specifically asked how they knew to use the various visualization elements. However, as detailed previously, most novices reported vague familiarity with these sorts of visualizations from various school or life

experiences, while most experts expressed extreme familiarity, to the point of incredulity that the interviewer would even ask. Given also the scientists' particular spread of training-specific sources of familiarity, such as graduate school, professional seminars and conferences, research work, and even collecting the data and producing these types of visualizations, it is apparent that reading spatial visualizations of global data represents an the end result of a process of enculturation, socialization, or internalization that scientists experience primarily through specific oceanographic training in graduate school and beyond.

Title and key use. Only six of 17 novices (35%) referenced using the title in the first visualization they viewed in the clinical interviews, though all of them used it by the end of the experiment. On the other hand, nine of 12 experts (75%) reported used the title on the first visualization, and all by the end of the experiment. More novices (14 of 17, 82%) and experts (11 of 12, 92%) used the color bar on the first visualization, though they were all asked what the colors meant.

Indeed, some participants in my experiment failed not only to use but even to notice the key or title, especially upon first viewing the visualizations. This was evidenced by either their lack of specifically referring to those visualization elements when asked how they arrived at their answers, or later evidence of their noticing those elements. When one novice, Veronica, was asked in the clinical interview what the measurement unit for the SST visualization was, after being asked both the main idea and the meaning of the colors, she replied "K" (Kelvin). She revealed she was from China and knew the United States used a different unit of measure than "C," which they use in China. A few moments later, however, she said, "Oh, this one is C. Oh, it shows the unit here! I found it. It should be this," laughed, and pointed to the key, with the units of C marked. Another, Gina, had this exchange with the researcher:

> Gina: am I allowed to like use this label up here. Am I, I'm allowed to see that?
> Researcher: Sure.
> Gina: OK. So I can use that information, um, what's that, um, that's a good question. OK.
> Researcher: Is that the first you saw the label?
> Gina: Yeah, I didn't notice before.
> Researcher: OK.
> Gina: Because the first time, I - the first screen it didn't really help me, I'm like, oh chlorophyll, duh (laughs).
> Researcher: OK.
> Gina: So I didn't even notice the label because I have been concentrating so much on pictures. (clinical interview)

There were also indications that participants ignored the information in the title or key if it had no meaning to them, especially with the SST Anomaly case. Here, in the eye-tracking interview, a novice reports reading the title but not incorporating the difference from normal piece:

> Researcher: So, what do you think the topic of this one is?
> Glenn: Standard surface temperature of the ocean?
> Researcher: OK. And how do you know that?
> Glenn: It says "SST" at the top of this picture. I remember in the past, that is the standard surface temperature. And also, the scale at the bottom reads in degrees Celsius, which is a way of measuring temperature.

Only later, when probed about the visualization, does Glenn admit that the temperatures depicted represent *differences* from average temperatures rather than “standard surface temperature” as he says above.

Finally, even when they used the title to understand the main idea, and had been through the set of visualizations in the interview portion, some still didn’t use the title easily, especially for the time span. Here, in the eye-tracking interview, Veronica had read the title aloud when asked about the main idea, but a few minutes later, couldn’t immediately find the time period.

> Researcher: All right. How about, what time period is represented here?
> Veronica: I don't know.
> Researcher: No idea?
> Veronica: Oh, one month. [Laughter]

> Researcher: And, how do you know it's one month?
> Veronica: It's in the title.

Or similarly for Jeff, “oh never mind, one month average, just kidding. It’s a month. (chuckles)... I finally found the title again, and saw that it said a month average” (clinical interview).

All participants also used the key both to judge color meanings or values specifically and also to make meaning for other questions, such as main idea, by confirming the units displayed on the key were reasonable for the main idea they suggested. An example is provided by expert Keith, who in the clinical interviews draws on the title and units together, “it says ‘one month sea surface temperature difference from average.’ That's what I define as an anomaly. The range is -5 degrees C to +5 degrees C,” but later uses the colors, “There's a color bar at the bottom that associates red with 35 degrees C and purple with -2 degrees C.” Five out of 12 experts (42%) and five out of 17 novices (29%) used both the units and the title to answer the question of visualization main idea in the clinical interview portion. Six out of 10 experts (60%) and seven of 10 novices (70%) used both together in the eye-tracking interviews. This increase in proportion between experiments among the novices may be due to a learning effect, as all participants had seen similar visualizations in the interview, even though at least three weeks had passed between the two experiments, supposedly sufficient to prevent memory consolidation (Roediger, Dudai, & Fitzpatrick, 2007).

Participants, especially novices, may also rely on the units part of the key in the place of seeing or making sense of the title. In the clinical interview, Gina (novice) starts to piece together the main idea of the chlorophyll visualization thusly, saying,

> They are measuring a substance in the water ... Um, they are measuring something milligrams per meters cubed, um, and so the cubed is a volume ... so I'm assuming that they are measuring something in the water because the water is highlighted, so yeah. (laughs).

Here, in fact, she also draws on the key to clue her in to the location of the important parts of the visualization in judging where the data is. Another novice, Virginia, also uses both at the same time when she is asked how she knows the locations she pointed out have the most extreme values in the visualization. She replies, “They’re the darkest colors, and also um, on the little bar, the darkest colors have the highest and lowest numbers” (clinical interview). By attempting to make meaning using elements that they do recognize, the novices demonstrate a level of sophistication in their reasoning that puts them on the path toward reasoning like experts.

Visualization pattern use. Novices used the patterns in the data as part of their reasoning less than the experts did, instead relying more heavily on the title bar and key as evidence when they noticed those elements. When pressed about what in the map part specifically lent them evidence for their assertions, in the clinical interviews, half of the novices referred only to the fact that the oceans were where the colors (meaning data) were located as evidence. In contrast, only two of 10 experts (20%) mentioned that. Similarly, in the eye-tracking interviews, 50% of novices and 30% of experts used that as evidence. As novice Mikayla says: “I guess that it's the oceans because that's what it's colored in, but I wouldn't know it’s chlorophyll,” (clinical interview) but expert Mark explains,

> Well so the title is one. And also, again, mostly from having seen figures like this before. From an oceanographic point of view it makes sense. ... It shows that the plant life is not distributed evenly through the Ocean; there's higher concentrations in the margins and some of the upwelling areas, I guess. Um, and then there's also differences between the open gyres of the central Ocean, and some of the high latitude.
> ... Researcher: How do you know that the colors are correctly distributed, as you were saying?
> Mark: Oh, um, mainly just from the background of having seen, you know, things like Chlorophyll concentrations, or pictures and satellite pictures. And also from just knowing that productivity tends to be enhanced along the margins and upwelling areas. (clinical interview)

Units of measurement decoding. As evidenced in the scoring, novices generally scored accurately on the question of the measurement unit; with three exceptions, all individual participants scored 0.8 or higher (1.0 was fully correct). One of the three failed to mention that the temperatures were in degrees, but otherwise vacillated between scoring partially correct (0.5) for noting simply “C” or “F” and scoring fully correct (1.0) for “Celsius” or “Fahrenheit.” This was the same participant who initially did not see the bar and assumed the units for temperature were “K” (presumably Kelvin). The two others struggled particularly with the chlorophyll units, each just focusing on half of the mg/m3 ratio, one on the milligrams part and one on the meters cubed. Otherwise, the novices used the units frequently to confirm their decoding of the main idea particularly, and reported that they were familiar with most of the units from elementary or middle school. Some in particular reported they knew mg/m^3 represented concentration or density from chemistry class. Otherwise, they mentioned a relative lack of comfort with Celsius versus Fahrenheit but recognized the abbreviation C as representing Celsius.

However, without further understanding of the title, recognizing the units’ abbreviations was as far as the novices could go, especially in the unscaffolded cases and when trying to judge the time span or season, or the temperature in Fahrenheit. The two examples below are both examples from versions of the SST anomaly visualizations.

> Samantha: ... maybe- maybe winter again? I mean, just because everything seems to be colder. There's not very many red or warm color spots here, so- but then again, I don't really know my Celsius in degrees so I'm not quite sure how the compares to the last one which I was kind of reading mostly in Fahrenheit. (clinical interview).

> Jeff: well since I know that zero [Celsius] is 32 degrees Fahrenheit, I would say that it's about ... goes up to nine ... hm, aw, that's probably .. so it'd be 32 plus like two or three, 35 maybe, degrees Fahrenheit.
> Researcher: so is that pretty warm or pretty cold there, at the equator?

> Jeff: ah, it's right about freezing or a little above freezing, so it's not warm by any means, but as far as the temperature of the ocean goes, apparently it's on the warmer half of their scale, here. (clinical interview)

Logarithm scale. About half of the novices who saw the chlorophyll stimuli in the clinical interviews recognized the visualizations for that topic have the numbers closer together on the scale than their distribution of linear values would suggest, though they do not speak of it in logarithmic terms. This is indeed the spirit of the use of the logarithmic scale – to be able to show larger differences across a large numerical span,

> Ivan: (pauses) Well, I mean, four and 0.04 is, is, uh, a big difference when the scale is from 0.01 to 10. But, uh, in terms of, uh, like how close they are, it seems like they're relatively close to, to what the, the bar is that's the standard. So, it doesn't, it doesn't seem like the, if it is water, it doesn't seem like the water is very much, that much different in Japan (clears throat) in comparison to the difference in, uh, by South America.
> Researcher: OK. How about the, that you said the numbers themselves, the four and the 0.04, would you say they are pretty different?
> Ivan: Yeah. I mean, four is a lot greater than the 0.04.
> Researcher: OK.
> Ivan: So, that is much different.
> Researcher: OK. But on the image, they look kind of similar. OK. (clinical interview).

However later, when considering what would be average on this scale after he described something as average spontaneously, he started to become confused:

> Ivan: Average? Well, for average for the bar I'm kinda confused on that - looks like it's in the green, but in terms of the numbers that could be around the orange, so I don't know.
> Researcher: Mm, OK. But, so when you say less than average, what do you mean by that then? I guess.
> Ivan: Well, less than average is, kind of looks like what-, whatever I think average is uh, if it's like the green color or if it's like the actual numerical value. (clinical interview).

However, they may not be using the scale correctly:

> Ivan: I just, uh, look at the bar, um, like the ratings they have for the different colors.
> Researcher: OK. And so, um, you said orange was about three, and there's not a three on there. How did you decide that it was three?
> Ivan: Oh, I just kind of estimated it about one-third of the way up from, of just like this section of one to 10. (clinical interview)

Seven of the eight experts (88%) who saw chlorophyll in the interviews mentioned logarithm spontaneously, while none of the novices did. Eight of the 11 novices (73%) seemed to be on the cusp of understanding logarithm in the manner of Ivan, with none of them actually mentioning the word. One mentioned the scale seemed “exponential,” and two more explicitly mentioned factors of ten or a decimal place difference. Ivan divided the color scale in a linear fashion despite arriving at the idea of the large difference in the values he reported relative to the range represented on the scale. Again, while this is likely a concept some if not most of the novices could have encountered in high school starting at the Algebra II level, it is doubtful that they encountered it in a scientific, let alone specifically oceanographic context.

Vocabulary. Scientists demonstrated a greater facility with academic scientific vocabulary than the novices, as is typical of expertise in a domain (Halliday & Hasan, 1976), both fluently decoding it by making academic scientific meaning from jargon used in the visualizations and encoding it by introducing it in their explanations. No experts struggled with the unscaffolded versions of the titles, specifically the terms anomaly and chlorophyll, or the SST abbreviation, beyond confusion about the time period of the averages represented especially in the anomaly case, as the definition of average used in the definition of anomaly isn’t actually encoded even in the academic scientific meaning of the word ‘anomaly’.

Novices don't decode this academic scientific jargon generally. Some even struggled to understand the specific context of a form of the word "equator", "equatorial", confusing it with north-south delineation:

> Glenn: Um, the Equatorial Pacific has low productivity, like, on the Equator, except *for the farther north and south you go*, the productivity increases--except for a strip that has a little bit higher productivity than usual off the west coast of South America. (clinical interview, emphasis added).

This contrasts with near-perfect performance when asked to point out the location of the equator, in both their gestures and references to how they know where the equator is, as discussed previously.

I also examined the use of jargon, defined as words with either exclusively scientific meanings, based on their inclusion in an oceanographic dictionary ("Glossary," n.d.) or used with their scientific denotation, based on their inclusion in Oxford's *A Dictionary of Environment and Conservation* (Park, 2007). See Appendix 6 for the full list of jargon used by both groups. Among the 27 total novice interviews (17 clinical interviews and 10 eye-tracking interviews), novices used 11 different words judged to be academic scientific language (excluding measurement units): concentration (three interviews), photosynthesis (two interviews), density (two interviews), currents, exponential, intensity, magnitude, solstice, equinox, model, and 45th parallel (all one interview each). Scientists used 96 in 22 interviews. Ten words were used in five or more interviews: upwelling (15), distribution (seven interviews), gyres (seven interviews), eddies (six interviews), tongue (six interviews), composite, warm or cold pool, austral, margin, and phytoplankton (five interviews each). Many words were used multiple times in an interview, especially by experts. Seven interviews also had the use of the word "anomaly" to reference specific features in visualizations or otherwise talk about the visualization outside of the context of the main idea or title. Novices might say "anomaly" in reading the unscaffolded title "SST Anomaly", but then they ignored the word, either reporting that the visualization

was showing only average temperatures, or expressing that they were at the best unsure about what the word meant, let alone what it meant in context. Further evidence of the confusion about titles and abbreviations is reflected in the alternative main ideas suggested for the visualizations by the novices, discussed below.

Main ideas. Twelve novice participants saw each topic in the clinical interviews. Six of them (50%) offered alternative main ideas for the SST visualization, four (33%) offered alternatives for anomaly, and ten (88%) offered alternative main ideas for the chlorophyll visualizations over the course of the interview. Some participants offered more than one alternative topic when they were unsure. For SST, two instances were *climate*, and one instance each were: *temperature inside Earth*, *air temperature*, surface of the temperature," "spread of heat across the globe", *heat transfer or distribution*, *heat across the water*, *geography*, and temperature alone without referencing the ocean. For anomaly, suggestions were: *depth*, *currents*, *heat*, *heat in the ocean*, *fishing*, *temperatures* (accompanied by confusion because they had seen temperatures in a previous stimulus and the patterns looked different here), *ocean temperature*, and *temperature anomaly* (with explanation that they did not know what SST was, but that "anomaly means something different") were each mentioned once. For chlorophyll, *depth* and *harvest*, *fishing*, or *life* were each suggested three times, "measuring a substance in the water", *dissolved oxygen* or *oxygen production*, and *mercury* were mentioned twice, and *climate*, *rain*, "currents or wind direction for water," water, "something to do with where sunlight hits," and "salt or minerals" were each mentioned once. The total participants offering alternatives is not equal to the sum of the alternatives offered as some participants offered multiple alternatives for the same version or across multiple versions of the visualizations for that main idea. For some instances in SST, the participants did not explicitly state that the visualization was depicting temperature data in the

ocean specifically. Others confused heat and temperature. The anomaly and chlorophyll visualizations prompted references to the broadest range of alternative main ideas: fishing, depth, currents, sunlight, substances in the water. Participants who recognized that the chlorophyll visualization was depicting measurement of a substance in the ocean (often by referencing the units of mg/m3)were uncertain exactly what the substance was; mercury, oxygen, salt, and fish were all suggested. Rain was suggested due to the greenish coloration of the scaffolded chlorophyll color scale as related to depictions in weather forecasts, probably from radar. In general, these results suggest that no color scheme, scaffolded or unscaffolded, offered here was sufficient on its own to provide accurate culturally-relevant associations, possibly because the main ideas or relation of the main ideas to the ocean (in the case of SST) are so unfamiliar to the novices in general.

New Codes

There were also several unexpected response patterns that emerged from the open coding.

Tendency to compare. Both groups actually naturally tend to compare visualizations to ones they've seen previously; all experts and all but one novice compared the later to the earlier visualizations within the clinical interview experiment. Eleven of 17 novices (65%) explicitly mentioned they recognized one or more as the same data, but 11 of 12 experts (92%) did. However, only experts have a vast trove of visualizations they have seen over their graduate training and professional careers to draw on in the absence of enough information presented in a single visualization, as evidenced by the previous familiarity discussion.

Comparison can provide a useful way to make meaning, except when academic scientific visualizations conflate elements such as the color scale and it

actually can interfere with academic scientific meaning-making. Previous work (Rowe et al., 2011) and some evidence here indicate that temperature is a widespread default association with the rainbow color scale. Several experts and novices alike commented along the lines of a general (probably United States or Western world) association of red with hot and blue with cold. At least one expert, however, also stumbled with one of the rainbow-colored chlorophyll visualizations: Janet: “Green, high productivity. Oh, hold on, no it's not. Um, the red is highest.” (eye-tracking interview). While the red-blue diverging color scale provided cultural association with temperature in the SST anomaly visualizations, many participants failed to specify that those temperatures were actually anomalies, or differences from average, when asked what the colors meant. In the expert case, given other evidence in the interviews, such as definitions of anomaly, references to El Niño, and specifically, accuracy on the question asking whether the equatorial Pacific conditions were representing “normal” in the visualization, this was likely another evidence of shorthand and assumed shared knowledge, However, in the novice case, this truly represented failure to make academic meaning because unless and even sometimes when pressed, participants did not indicate understanding of the temperatures being different from average, either indicating that two degrees Celsius meant the equator was indeed close to freezing (see Jeff’s comments, above in the discussion of measurement units), or otherwise having their reasoning fall apart.

Word choice: Conflation amongst “normal,” “typical,” “average” and “usual.” The tendency of users to compare visualizations was also supported by answers to questions about whether the visualization was indicating normal, typical, average, or usual conditions. The terms were used interchangeably throughout the interviews. Novices indicated overwhelmingly that they did not know what “normal” or “typical” was in most cases, despite the presence of the scaffolded titles that indicated “average” or “different from average.” Linda offered

another form of confusion, reflecting her relative lack of academic experience: “I think of ‘difference from average’ as normal since temperatures change.” (clinical interview).

On the other hand, scientists could judge the value ranges of the temperatures and in some cases, chlorophyll, as appropriate based on their prior experience. Their struggles with the scaffolded anomaly time span in particular revolved more around precisely how the averages were constructed, both the “one month” and the “average” from which the month’s average was subtracted, indicating the vagueness inherent in those words (as well as “anomaly” itself):

> Brent: ... You know we always argue about what (pauses) it means to be a one month average. Whether it's the average of every data point in that specific month or if it’s the average of all months, (pauses) and all instances of that month, over the lifetime of the satellite. So if that's, you know, a decade long satellite record they could average all of the Aprils, or something like that, as opposed to just averaging this one April. I can't tell [which] from the plot. (clinical interview)

When asked whether the equatorial pacific conditions in the anomaly visualization represented something ‘normal’ or ‘different from normal,’ Brent said, “Who knows what’s ‘normal’? – that’s a judgment call” (clinical interview). El Niño has only been understood as a global phenomenon in the satellite age, as Charlie reported in the clinical interview.

Time span. Novices don’t always even understand the question when they are asked about what time is represented in the visualization, whether the question is posed as time period (as it was originally), time interval, or time span. Some confuse it with a geological epoch when the question is posed as time period, but some don’t even seem to understand the idea that more than a snapshot could be depicted, as evidenced by their clarifying questions to the interviewer: “Researcher: what time interval, maybe, do you think this image represents? Vanessa: Uh, like time of day? Time of year?” (clinical interview).

This could be a problem with the question (especially when it was posed as “time period”), but few experts stumbled with what the question was designed to elicit, even if they were incorrect. This suggests further that novices don’t think of the visualizations as representing any sort of time interval. Again, experts acknowledged the ambiguity of words: “Charlie: When I say an annual average I you know it’s hard to say what year or how many [years are averaged together]” (clinical interview). However, both groups have trouble pinning down the time period if they are not familiar with the particular content, especially, as evidenced by the scoring described earlier.

Season. Experts, despite familiarity with satellites, still don’t all use what that implies as an indication of season, from the eye-tracking interview, an expert:

> Keith: Well I was trying to see if one hemisphere had more chlorophyll than the other. And it sorta looks like the southern hemisphere does, but it's difficult to say because a lot of the northern hemisphere is covered by the light grey. (eye-tracking interview)

This pattern of data loss should indicate winter in the southern hemisphere as the darkness at that time of year precludes this satellite from getting data. In this visualization it is very clearly delineated as an almost-latitudinal horizontal cutoff, which some experts comment on yet don’t make the connection to time of year spontaneously.

Both groups actually struggle when pressed about why the seasons occur in the way they do; the oceanographers fail to recognize the influence of the specific heat of the water on the temporal pattern of the ocean. The greater heat capacity of the ocean means it is warmest in the Northern Hemisphere towards fall, when hurricane season peaks in the United States (September), and is actually more roughly symmetric in December and January (the month represented in the clinical interview visualizations) because of the longer cool-down period. The novices sometimes struggle with the reason for the seasons at

all, confusing the influence of the tilt of the Earth with the distance of the Earth from the sun. However, they recognized the connection of the sun with the seasons, so ultimately that particular factual error might be less of a problem than the unfamiliarity of specific heat of the ocean.

Specific Problems with Visualization Elements

Finally, the clinical interviews and eye-tracking interviews revealed some elements of the visualization that caused or could potentially cause confusion. Problems with the key labeling in SST anomaly visualizations

Perhaps as part of the confusion surrounding the use of the terms "typical," "average," "normal," and "usual" or "unusual," the attempt to scaffold the SST anomaly key with "average" for zero change was confusing for some:

> Veronica: ... I, I actually I don't understand why, maybe it's not average, but it's that the, um, why it's a, why it's zero is average. No, no. It's not, *zero is not average*. Zero is a difference between average and average. *If, if a place is at average temperature, the difference between this place and the average is zero.* (clinical interview, emphasis added)

Or this one:

> Linda: The red is (pauses), see and I don't, I don't quite, I can't quite puzzle this one out. My brain's not feeling like I'm understanding how you can have negative degrees difference. (laughs).
> Researcher: Okay. Tell me; tell me a little bit more about that.
> Linda: You know like if I'm, if I'm trying to see the difference in temperatures, if I'm trying to get temperature difference from average, it would always seem to be a measurement that would be in degrees, that you, your difference couldn't be less than zero. (laughs).
> Researcher: Mmm. Okay. Okay. Makes sense. I mean I understand what you're ...
> Linda: Right, yeah. Yeah that if I'm really thinking about this as temperature difference, that and it's a difference from average, you're either right at average which would be zero degrees difference, or if it's different from average, then it's a positive measurement. ... Um, the red is five degrees temperature difference. And the violet would be negative five

> degrees temperature difference. But then I, I haven't been able to imagine how you have a negative measurement of difference. (clinical interview).

And a third novice's explanation of the confusion:

> Shelby: um, they have degrees Celsius and degrees Fahrenheit (pause) on the yeah.
> Researcher: okay, great.
> Shelby: but the degrees Fahrenheit is kind of weird. 'Cause it's saying average is zero degrees Celsius which is like thirty-two and it says nine degrees higher instead of writing like, forty-one degrees Fahrenheit, (pauses) which is kind of interesting. Probably intentional. (clinical interview).

Finally, an expert, who recognized the visualization as displaying anomaly, read across the left- and right-hand labels on the scale when it was placed vertically, and was confused as to whether the top value was 59 degrees, which he thought was not feasible for anomaly, or 5.9, instead of 5 degrees Celsius *or* 9 degrees Fahrenheit higher than average. Perhaps a better labeling would be "no change" or removing the numbers of degrees altogether, and instead leaving higher and lower, and indicating a cutoff of significant deviation from the long-term average. Again, this would benefit from prototyping with real-world users.

Color scale and judgments of particular values. Participants were asked to judge values on the visualization partly to assess their use of the color scale and partly to assess their retrieval of information in the visualization accurately. However, this turns out to be a more difficult and somewhat disingenuous task for true meaning-making from these visualizations. For one, experts commented that they don't typically judge precise values from these visualizations but rather rely on them at most to discover patterns and at the least to communicate those patterns to others.

One recurring complaint about the chosen scaffolded scales when judging values was that the extremes were harder to distinguish, probably due to lack of perceptibility of differences in luminance, which more-or-less replaced variation in

hue from the rainbow color scale. Another criticism along those lines was that the continuous color scales were essentially too continuous and differences among all values were difficult to judge. One comment raised was that the dark pink or purple at the bottom of the rainbow color scale could be confused with the dark red at the top. This was a particular issue in the anomaly visualization when some eddies contained both extreme high and extreme low deviations in close proximity.

At least one expert participant explicitly raised another problem of color perception based on surrounding colors, noting

> Well, the juxtaposition of the colors next to colors, next to different colors will make colors look different if that makes sense. And those are surrounded by greys, and so they might, more grey than normal. So they might somehow be tricking the eye to be looking darker. (Rick, clinical interview).

This is also an example of gestalt tendencies of the brain (Koffka, K., 1935).

In situ Interviews During Eye-tracking

Six males and seven females participated in the eye-tracking with concurrent interviewing at the science center. They were roughly equally distributed amongst comfort level in interpreting these visualizations, levels of formal science education (though skewed toward completing less than college-level science), and level of participation in science hobbies. See Table 19.

Overall, these demographics are similar to the general science center visitor population. The science center typically sees a higher level of visitors with a bachelor's or higher degree than national averages.

Participants overall had a "partially correct" score of 0.4, lower than both experts (0.83) and novices (0.61) in the semi-clinical interviews and interviews accompanying laboratory eye-tracking (0.93 experts and 0.61 novices), with lower scores on all individual questions as well.

Of the five questions asked, participants scored most poorly on the questions of time span (0.06 out of 1.25) and season (0.25) represented, similar

Table 19

In Situ Eye-tracking Participant Description (N = 13)

Background	Number of Participants
No college science	6
Degree in science/engineering	4
Some college science	2
No answer	1
Number of Science Hobbies	**Number of Participants**
Few	6
Many	7
Comfort Level with Interpreting Visualizations	**Number of Participants**
Not Very Comfortable	5
Comfortable	5
Very Comfortable	3

to the expert and novice participants in the first two experiments. Thus, I computed a sub-score of the first three questions asked: visualization main idea, color meaning, and location of highest values, which dealt more with explicit interpretations of the scaffolded information given. Participants averaged "partially correct" (0.56) on those three questions. See Table 20.

Overall, performance on the SST anomaly questions was higher than for chlorophyll, with the exception of main idea, where performance was similar (0.48 versus 0.46 correct), and season, where chlorophyll was higher (0.31 versus 0.19). This may be due to the relative familiarity of the temperature condition and the red-blue scaffolded versus blue-white scaffolded color schemes. However, on the main idea question, the response patterns differed: roughly equal numbers of participants scored completely correct (1.0) for the main idea in the chlorophyll question (six, including one who was scored as sophisticated (1.25) for associating chlorophyll with algae, 46%) as got zero

Table 20

Accuracy, In Situ Eye-tracking Interviews (N = 13)

	SST anomaly	Chlorophyll	Viewed first	Viewed second	Unscaffolded	Fully Scaffolded	Overall
Main Idea	0.46	0.48	0.37	0.58	0.44	0.5	0.47
Color Meaning	0.73	0.27	0.38	0.62	0.65	0.35	0.5
Highest Value Location	0.79	0.42	0.46	0.75	0.63	0.58	0.61
Questions 1 - 3	0.7	0.42	0.46	0.65	0.62	0.49	0.56
Time Span	0.08	0.04	0	0.12	0	0.12	0.06
Season	0.19	0.31	0.15	0.35	0.27	0.23	0.25
Overall	0.46	0.32	0.3	0.48	0.42	0.36	0.39

credit (seven of 13, 54%). No partially correct score was awarded. In the SST anomaly case, only two participants (15%) scored completely correct for noting it was both temperature and an anomaly, three participants (23%) scored completely incorrect, and the remaining eight (62%) scored partially correct for deciphering that the visualization concerned temperatures, but missing the anomaly part. If no partial credit were awarded, the SST anomaly main idea performance would have dropped to 0.15.

Viewing order showed that the participants improved on the second visualization they viewed in all cases, indicating perhaps a learning effect wherein they better understood the questions or were more familiar with answering questions using the visualization. They also may have noticed the title or key after examining the visualizations for a longer period of time while answering questions.

Scaffolding slightly improved understanding of the main idea, indicating perhaps better clarity of titles, but worsened performance on the color meaning and highest value location questions. The general unfamiliarity of the scaffolded chlorophyll color scale to the laboratory participants indicates a possible reason for poorer performance. However, given the problems with the visibility of the key and the colors themselves when displayed on the exhibit (for further discussion see Appendix 7), particularly in the SST anomaly case, I am hesitant to conclude that the scaffolding of colors actually interfered with meaning-making in this case. Instead, I rather suspect that the confound of lighting, which especially washed out the colors in the anomaly visualization, made the “brighter” more saturated colors of the rainbow color scale easier to pick out on the globe for the location question. In addition, the move of the color scale from the bottom of the visualization to the left may have put it out of view of some participants. For the chlorophyll case as well, the relative size of the coastal features that were brightest compared to the equatorial feature which stood out centered in front of participants and was indeed brighter than the surrounding ocean, may have

been a confound. This indicates a need to prototype color schemes *in situ* as well as in the lab for a balance between visibility and scaffolding for meaning-making based on cultural familiarity. Confound with depth may also indicate that flipping the color scheme in this instance, so that “more” means “darker” may be both easier to see *in situ* as well as more culturally relevant while providing less confound.

Putting the Puzzle Pieces Together

Novices. Overall, novices rely on relatively sound, if incomplete, ways of thinking about the visualizations, looking for information in the title and key that provides information, and trying to make sense of the patterns by comparing them to those they have seen before. Only twice in the entire set of clinical and eye-tracking interviews did novices mention any sort of “pseudoscience” understandings; two participants referred to learning that the North Pole was cold at an early age as the place where Santa lives. However, while the sources of their knowledge might be scoffed at by academic scientists, it is precisely the sort of folk knowledge that could be built upon and incorporated into a more academic understanding of why that area is cold.

Some information in particular about satellites and oceanography conflicts with other knowledge on which novices rely in their active meaning making. They have little unprompted spontaneous concept of these visualizations as representations of anything more than a snapshot, even though they cannot explain how vast amount of data is collected.

> Allison: Umm just testing it randomly for a month and then adding them together, and then dividing it, like finding the mean temperature.. in certain areas.
> Researcher: Okay. What do you mean by randomly?
> Allison: Like... I don't know (whispered) ... I feel like they would either umm... have to test a certain area, see how hot it is throughout the day periodically er Once a day at the same time of the day to figure out what it would be, on average, in certain areas

> Researcher: Okay and what sorts of areas do you think or..?
> Allison: Umm I would guess they would try to do at least one like section in there. I don't know how they would do that unless from space they took a picture and could take the heat off of it but that wouldn't be one month average so...
> Researcher: Why do you think that couldn't be a one month average or why wouldn't that be?
> Allison: Umm.. I mean they could if they took a picture everyday and then did it together but I don't know, I feel, we're probably technologically advanced enough to do that but I don't know for sure so I feel silly saying that. ... 'Cause I've never took- thought into how they create the maps. (clinical interview).

As Allison also indicated, novices seem to be thinking about this application of scientific ideas for the first time. Once they are guided to consider specific things that could factor into the creation of these visualizations, they start to reason, though not always well. Without familiarity about the visualizations, particularly perhaps about doing more than simply reading a predicted temperature (often also scaffolded with number) off a television or newspaper map, they may not have an adequate schema or background experience to think about how to make meaning from these things. If they do have a schema, it is only to start with the key and the title, but often they have to spend time decoding these elements, where experts spend virtually no time doing so as they are familiar with the jargon, the data and the task:

> Researcher: Ok, how bout um tell me about what time span you think might be represented here?
> Jeff: Um if the SST does actually mean Standard Seasonal Temperature or something about seasons then it could be four months or it could be just like the first one would be one month.
> Researcher: Ok
> Jeff: So it would kind of have two options I guess?
> Researcher: Ok, but it doesn't seem like a year? Doesn't seem like a day?
> Jeff: I wouldn't, I dunno I don't think so.
> Researcher: Ok, do you have any idea what it would look like if it were a year or a day or?
> Jeff: Umm I bet if it were a year it would be similar maybe a little bit more of the green color just a little more towards the middle.
> Researcher: Ok

> Jeff: Versus if it were a day it would be very - none of the colors would like blend in with each other it would be all distinct like red yellow green blue for the one time only sort of thing.

Similarly, in the *in situ* case, participants improved from one visualization to the next, perhaps demonstrating that they looked specifically for information in the second visualization which they had been asked about in the first visualizations. Eye-tracking patterns might help confirm this; see Appendix 7 for issues surrounding the collection of the eye-tracking data *in situ*.

Novices employ prior by-and-large academic scientific knowledge, seeking the information given in the visualization pattern, title, and key. However, they draw the wrong conclusions because still they don't have enough information, either because they cannot decode jargon or because they have limited familiarity, especially with what is normal in an oceanographic context. Experts and novices make very different assumptions from the given information based on their experience. Here, Ferdinand thinks of the average as a spatial, not temporal, one, which leads him to confusion about the high-latitude warm anomalies:

> Ferdinand: Oh, it's saying sea surface temperature difference from the average. From what I'm looking at maybe like some of the places are much colder and warmer have a place or that's *has more heat than other locations.*
> Researcher: Okay and how do you know that's what those places are or what makes you think those places have more heat or less heat?
> Ferdinand: More because of the color, the color that's been represented in that location.
> Researcher: Okay and what color means what?
> Ferdinand: Well, the blue would be something, a place where it's normally cold or like, yeah like the temperature is colder and other locations like I guess reddish color would be little warmer or has, yeah little warmer. And I'm not really sure about there like red, the blood red I don't understand that so, that's where like backfires at me.
> Researcher: Okay, tell me what you don't understand? What would you expect or something?

> Ferdinand: I would expect that the top part like at the very top, I would expect that place to be like freezing and cold.
> Researcher: And instead the image is showing that it's what?
> Ferdinand: It's higher so it's like saying that place has warm sea surface temperature. (eye-tracking, emphasis added)

Even once the novices understood the data that was being represented, or at least could accurately report the title and color meaning, they still failed to see the larger significance. When asked "what would you tell someone this image is about?", Eden replied haltingly, "I think this is the...I'd say it's the same one as the surface temperature in the oceans. So where the average temperature is and where the highest difference...the where it's warmest and where it's the coolest...from average" (clinical interview).

In fact, novices often eventually manage to gather all the pieces, but the final picture is askew from the academic picture, sometimes missing pieces when they rely overmuch on single features and ignore the overall picture, partly because they don't know what the final picture is supposed to look like. For example, Emma states that she does not know what normal is for the SST anomaly visualization:

> Researcher: How about, does that area indicate that the equatorial Pacific... Does that look like it's, what's represented is considered normal or is it different from normal in this image?
> Emma: I don't know what normal would look like so... I don't know. (clinical interview).

They often cannot reconcile conflicting information. Here, viewing the fully-scaffolded SST anomaly visualization, Brad wonders about the patterns he sees in the data versus what he "understands" from the title and key about the temperature difference from average, and prior knowledge of ocean temperature patterns:

> Brad: Darker red represents um, maybe a greater difference in change, and then lighter blue represents like a consistent temperature, or like a relatively consistent temperature. ... And it's also colder, but doesn't make

> sense to me that the oceans up in the north are really hot, 'cause they're like blood red. (eye-tracking interview)

It seems that given their limited lack of prior knowledge, limited experience with the visualizations and position on the fringes of the academic scientific culture, novices are still trying to correctly assemble all the pieces of the puzzle.

Experts. Almost without fail, experts' data revealed their comfort and familiarity with the visualization interpretation task compared to novices. Their score accuracy was overall higher than novices in almost every category. In the interviews, I found they used all elements of the visualizations and especially the patterns in the data to make meaning. When there were struggles, conflicts among elements or with their prior knowledge, by and large they did not blame themselves, they either blamed the visualization creator for leaving out or poorly portraying the information,

> Brent: Well I'd need the caption. I mean this could be a study of, of you know seasonality if this were a climatological month relative to the climatological mean um and you know it doesn't say. This just says one month. It doesn't say the specific month or anything like that. (clinical interview).

or recognized that the information they needed was specific to the sub-discipline with which they were not immediately familiar:

> Jay: I don't know. I'm not a Global physical oceanographer - If someone pointed things out to me, I could say "oh yeah, that's what it is," but no. I don't know. I'm not even convinced that it's winter looking at what's going on here. It's all awfully warm there in the - around Greenland and the Denmark Strait that area. I know that in the winter time there's a lot of cold water mass being produced there. So, I may be entirely wrong on the season. Being I might be wrong in the season I can't really say what stands out as unusual. (clinical interview)

These ways of resolving or managing conflicts indicate familiarity with the task, its requirements, and the field at large, products of the academic oceanographic enculturation that novices had not experienced.

When the experts struggled with the source of the data as satellite, and what that information could tell them, they were unfamiliar with technical details rather than the overall concept. This discrepancy with the level of understanding reflected a difference from the novices again due to varying levels of enculturation. The experts again attributed problems here to a lack of the data source provided in the accompanying information in the visualization, as when Janet notes in the clinical interview, “They’ve also left off their source. So, I don’t know what satellite it is.” However, for these particular visualizations, even the technical details were important to accurate information transmission; when the expert participants and missed or alternately understood the patterns of lack of satellite data, they had to rely on the patterns of data as symmetric about the equator. This led several to erroneous conclusions that a spring or summer month was depicted, as they failed to draw on specific knowledge of the ocean’s seasonality which actually means that winter and summer are more symmetric, due to heat capacity of the ocean versus the atmosphere. It could be that time span and date of data collection are so often provided in these visualizations that judging the time span and season are actually novel tasks for the experts.

Experts did not struggle to decode the abbreviations; in fact, some of them criticized its use, “[SST] should be spelled out. Not good labeling,” (Janet, clinical interview). They also had no problems with the jargon, and even criticized its use, especially in the case of chlorophyll,

> Charlie: ... it's concentration of microscopic ocean plants which is kind of a misnomer actually and the reason that's a misnomer is that that assumes it's the number of cells or plants or volume say per cubic meter but different phytoplankton vary greatly in size. So the smallest, the largest ones are several thousand times the size of the smallest ones, so if you say concentration you know you can have thousands or even you can have millions of the small ones it would be equal to a few thousand of the large ones ah in terms of the sac- in terms of chlorophyll milligrams

> per meter cubed, but it would be a different concentration of microscopic ocean plants. (clinical interview).

Another expert, Jay, pointed out that productivity should have a different unit than what was portrayed on the scale bar:

> Production is biomass produced per unit time, so if I would, I would guess that production should have a different unit symbol what were presented in the figure two slides ago whereas Chlorophyll A, you know, I think that's referring specifically to how much phytoplankton is there, so the units, to me, looking at the title Chlorophyll A, the units of milligrams per meter cubed makes more sense to me. (clinical interview).

Expert confusion around the words “average,” “typical,” and “normal” also revealed a level of experience with the words in the context of oceanography and visualizations that differed from novices. Experts quibbled that “there are no ‘typical’ periods, but really on an average, yes.” (Ray, clinical interview), or, as described earlier, Brent pointed out that what is ‘normal’ is still argued about in the field of oceanography. However, at the least, the experts tended to think of the time span or meaning of average as a long-term average, as that was more meaningful than a short-term one:

> Anomaly is general term used to describe, the departure from an average. So....That implies that we know average sea surface temperature ... for this to be meaningful that average temperature needs to be based on a fairly long measurement interval, so like "global" or it can be "decadal". (Rick, clinical interview).

For all of these reasons, experts were able to accurately describe the main idea of the visualization regardless of scaffolding level. Where they struggled, they did so because of a lack of information provided that they knew was important. Though they did not always arrive at the correct answer for time span, the experts considered several alternatives when met with conflicting or missing information and weighed them against one another:

> It’s Summer in the Southern Hemisphere. That's why we see this high Chlorophyll and really there is no ice, as I said, with this region. But, also, this tells me that there is a bloom taking place in the Northern Hemisphere

> so, this suggests that this is a composite image ... And it cannot be an annual average because you have, unless, unless, what they did was to take the annual average for the times that they have pixels open, but then that would not explain why you don't have data in this area here, because during summer we have plenty of data in this region, so it doesn't explain why we don't have uh the uh why we have that grey uh bar across. That suggests to me that is that is taken during winter time. Winter time in the north hemisphere and summertime in the southern hemisphere. So it may be it may be an issue of scale. (Ray, clinical interview).

Finally, as explained previously, besides recognizing the intended main idea of the visualization, the experts were able to offer other ideas for what the data could be used to represent.

Summary

The interviews revealed many differences among the three populations studied. Visitors to the science center were the least accurate, though they were also offered the least opportunity to use the guiding questions of the interview to improve with practice. Despite the short experiment time, there were indications in this group, however, that practice with the task and the questions did improve their accuracy, and that when they could see the more culturally-familiar visualization elements of title and key, they were also more accurate. Their relative distance from formal schooling likely influenced their ability to draw on specific academic science knowledge that the laboratory novice population used, and there were also problems with the visualization quality on the spherical display. However, these early results show promise for designing interventions with visualizations in museums and other education settings.

The laboratory-based interviews offered deeper insights into the struggles that academic science novices have when trying to make meaning from spatially-based visualizations of data. Novices revealed a general lack of familiarity with the task as well as with almost all of the elements provided to assist with meaning making, such as the title, the key, and the patterns in the data

themselves. They also had a limited background of academic knowledge and experience on which to draw compared to the academic experts. However, the novices, too, showed improvement with practice and with more culturally-familiar visualization elements. Ultimately, though, the novices were generally unable to come to the same conclusions about the meaning of the visualizations that the experts could and that were intended by the visualizers to be communicated.

Eye-tracking Results

This chapter presents the quantitative and qualitative results related to eye-tracking. I begin with a short introduction to the quantitative measures and the participant demographics for the eye-tracking experiment. The quantitative results include numbers of fixations on the visualizations, duration of those fixations, and relative likelihood of fixating on different parts of the visualizations. Qualitative results are visualizations themselves of which parts of the visualizations and in what order individual participants look at, and to the extent possible, what patterns are common within and between participant groups. In both cases, I consider the differences among scaffolding levels, and in the quantitative case, among visualization topics as well.

Eye-tracking captures participant gaze position and timing data. These are translated into fixations, records of positions on which participants "stop" and dwell for a minimum time. The human eye is rarely "fixed," even when attention is trained on an object, the eye moves rapidly back-and-forth in unconscious movements called "saccades." Thus, a fixation as defined by the eye-tracking software also considers how far away each stop is from the previous one; if it is within a maximum dispersion area, the saccade and next stop are counted as part of the fixation before, making the fixation truly a group of individual stops and dwells with the saccades between them. This calculation continues similarly until the position of dwell falls outside the dispersion area and the fixation is considered to end and a new one might begin if it reaches the minimum dwell time. Duration is the total time of a fixation, based on the individual dwell times on each stop in the group plus the time of the saccades between the stops.

Participant Calibration Results

As reported by BeGaze™, all participants' calibrations (experts and novices) were within one degree of angle deviation in either direction upon

calibration, as recommended, with an average x- and y-deviation angle of 0.41º and 0.44º, respectively. No individual had a deviation angle greater than 1.0º in either direction. The tracking ratio, defined as “the number of non-zero gaze positions divided by the sampling frequency and multiplied by run duration, expressed as a percentage” (“ExperimentCenter™ Manual version 3.1,” 2012, p. 236) for all participants was greater than 75%.

Quantitative Results

In SMI BeGaze™, participant eye position data were grouped into events called “fixations” of minimum dwell duration 80ms and maximum pixel dispersion 100 pixels. The first 10 seconds of the participant viewing each visualization was called “spontaneous looking” following Libarkin et al. (n.d.). Then each participant’s data was divided by answers to the first question that was asked, the Main Idea (MI) condition. First, consider the spontaneous looking condition. Since the data are not continuous across all independent variables (for example duration is continuous, but expertise and level of scaffolding are not), I chose linear modeling to fit the data. Only left-eye data is reported throughout.

Spontaneous looking condition. For the spontaneous looking (SL) condition, eye-tracking data for the first 10 seconds for each visualization the participant viewed were entered for analysis. Each participant’s total number of fixations and duration of fixations across all five experimental visualizations were calculated by BeGaze™. The mean and median for all participants, as well as experts alone and novices alone was 29 fixations, with a range of 16 to 42 fixations for the entire group and the experts, and a slightly smaller range of 18 to 38 fixations for the novices only. See Table 21.

Durations for the overall group and experts and novices alone were produced similarly and descriptive statistics calculated, with very little variation in any measure among groups. Mean for the entire participant population was

Table 21

Number of Fixations per Participant per Visualization, 10 seconds Spontaneous Looking, All Participants, All Novices, and All Experts

Statistic	All Participants (n = 18)	Novice Participants (n = 9)	Expert Participants (n = 9)
Minimum	16	18	16
Median	29	29	29
Mean	28.81	28.53	29.09
Maximum	42	38	42

Note. One expert and one novice did not participate in the spontaneous looking because they participated before the condition was added to the procedure.

321.6ms, with a range of 82.3 to 2469.3 ms and a median of 258.6 ms. Since the distributions of the fixation durations were not normal, I used a Wilcoxon Rank Sum test to assess the means between the two groups. The test showed differences between expert and novice durations were not significant W_s (n_1 = 50, n_2 = 50) = 1193, p = .69). See Table 22.

Table 22

Durations of Fixations for All Participants, All Novices, and All Experts, SL

Statistic	Fixation Duration (ms)		
	All Participants (n = 18)	Novice Participants (n = 9)	Expert Participants (n = 9)
Minimum	82.328	82.407	82.328
Median	258.614	258.612	258.617
Mean	321.648	325.6	317.683
Maximum	2469.33	2469.33	2319.142

Note. Fixations were defined as a minimum dwell time of 80 ms.

Each participant's fixation durations were plotted as histograms and density curves to determine normality. These plots showed that some participants' histograms appeared roughly normal, but many appeared long-tailed. Figure 4 plots all individual density curves together to display the

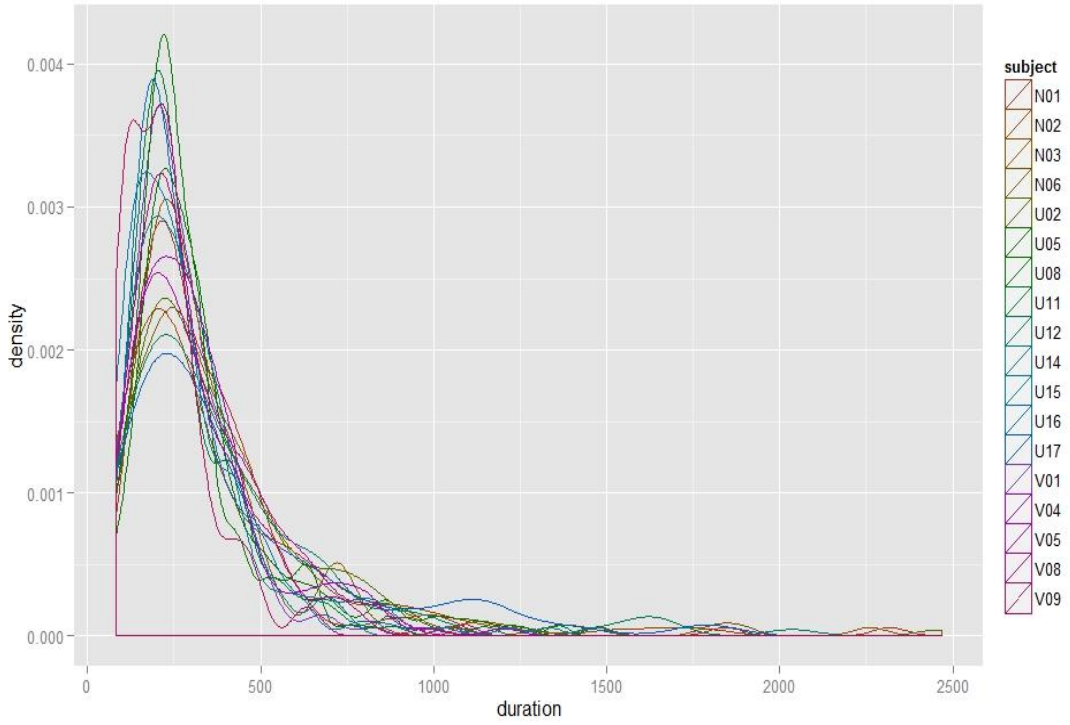


Figure 4. Density curves of individual participants' spontaneous looking fixation durations. Each color represents a different participant.

similarities and variability. Thus, I applied a logarithmic transformation to the durations and re-plotted the data. Participants' log-transformed data appeared to be less skewed, but overall, they were not vastly improved from the non-transformed data. See Figure 5.

Finally, I determined that, because of the fixation definition of a minimum duration cutoff of 80 ms, the duration data were actually best described by a truncated normal distribution, with a truncation point of 80 ms. Truncated regression models take the non-truncated data and assume the truncated data would complete a normal distribution with the non-truncated data (Robert, C. P., 1995). While analysis assuming Poisson-distribution is more common in eye-

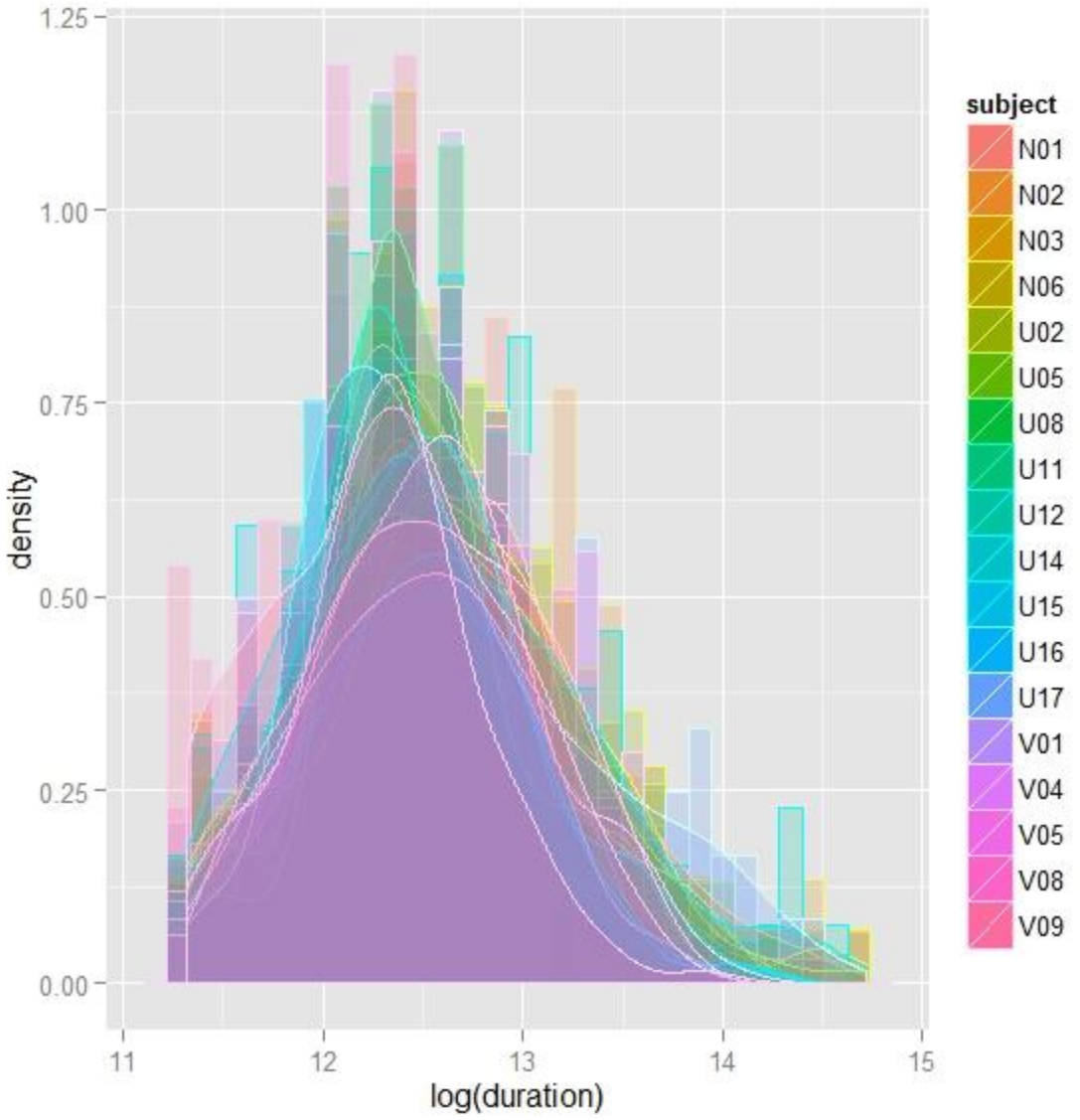


Figure 5. Log-transformed histograms and density curves of individual participants' fixation durations. Each color represents a different participant.

tracking, analysis by approximation to Gaussian curves is allowed when lambda is greater than 10 (here, 18 participants and minimum 16 fixations each) (Holmqvist et al., 2011). Poisson distributions also have equal mean and variance; the duration data had mean 321ms, standard deviation 234ms (variance 54756 ms^2).

Holmqvist, et al. cites two papers suggesting that fixation duration may not be independent: in scene viewing (Tatler & Vincent, 2008) and in visual search (Hooge, Vlaskamp, & Over, 2007), short fixations tend to be followed by short fixation durations and long fixations by long. These papers suggest there is an influence of "pre-programming" by previous fixations on subsequent ones, rather than simply a stimulus-dependent "process-monitoring" component on duration (Tatler & Vincent, 2008). However, in this study's data, after coding durations as either above, below, or equal to the median (instead of the mean as measure of central tendency due to the long-tail and truncation), only 53% of durations were found to be on the same side of the median as the previous fixation's duration, and the remaining 47% were on the opposite side. This points us to a stronger influence of process-monitoring, that is, stimulus characteristics, than on correlations between fixation durations in the present case. Further, correlations of duration and previous fixation's duration were only .1 for my data, and a joint distribution plot showed random relationship between duration and previous fixation's duration. Thus I treated the duration data as independent for analysis. Given the normality of the truncated half of the data, and the independence of fixation duration, I used truncated linear regression to model the duration data, which assumes normality for the truncated data.

Five independent variables were examined as possible predictors of fixation duration: Level of scaffolding, Trial number, Expertise, Topic, and Gender. The first model considered all these factors, without interactions. Level and trial were found to be highly significant (both $p < .00006$), gender borderline ($p = .047$), and expertise and topic not at all ($p = .12$ and $p = .26$). See Table 23. When level and trial interaction was introduced, similar patterns emerged, with the interaction significant ($p = .015$) and gender slightly more significant than previously ($p = .029$). See Table 24. Examining gender as an effect by itself proved it was not significant ($p < .47$), so it was dropped from the model at this point as a weak influence. See Table 25. Since expertise was a variable of

Table 23

Truncated Linear Regression Model with Five Independent Variables, All Participants (n = 18), SL

Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	215836.5	60165.8	3.5874	.000334 ***
Level	-26480.1	6610.1	-4.0060	6.175e-05 ***
Trial	-18056.0	2759.1	-6.5441	5.985e-11 ***
gender	49870.5	25128.6	1.9846	.047188 *
expnov	-35779.5	23270.9	-1.5375	.124166
topic	-19071.3	16936.2	-1.1261	.260139
sigma	327071.6	7544.2	43.3539	< 2.2e-16 ***

Note. The formula in R for this model is: truncreg(formula = duration ~ Level + Trial + gender + expnov + topic, data = all.participant.all.levels, point = 80000, direction = "left").

[a]Estimates are in microseconds.
*p < .05. **p < .01. *** p < .005

Table 24

Truncated Linear Regression Model with Five Independent Variables and Level and Trial Interaction Term, All Participants (n = 18), SL

Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	226159.0	60146.0	3.7602	.0001698 ***
Level	-43675.6	12218.2	-3.5746	.0003507 ***
Trial	-24290.4	4037.5	-6.0162	1.785e-09 ***
gender	56306.8	25738.1	2.1877	.0286927 *
expnov	-32828.3	23351.2	-1.4058	.1597692
topic	-17649.6	16931.2	-1.0424	.2972118
Level:Trial	3517.1	1441.3	2.4403	.0146757 *
sigma	326086.4	7499.2	43.4829	< 2.2e-16 ***

Note. The formula in R for this model is: truncreg(formula = duration ~ Level * Trial + gender + expnov + topic, data = all.participant.all.levels, point = 80000, direction = "left").

[a]Estimates are in microseconds.
*p < .05. **p < .01. *** p < .005

Table 25

Truncated Linear Regression Model with Gender as Main Effect on Fixation Duration, All Participants (n = 18), SL

Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	163195.1	35600.9	4.5840	4.561e-06 ***
gender	-14438.2	21350.7	-0.6762	.4989
sigma	303489.7	5957.3	50.9446	< 2.2e-16 ***

Note. The formula in R for this model is: truncreg(formula = duration ~ gender, data = all.participant.all.levels, point = 80000, direction = "left")

[a]Estimates are in microseconds.
*p < .05. **p < .01. *** p < .005

interest, I ran it as a lone independent variable, with level and with trial. In the first case, it was significant (p < .002), but in the second it was not (p < .16), so at this point it was also dropped from the model. See Tables 26 and 27.

Table 26

Truncated Linear Regression Model with Expertise as Main Effect on Fixation Duration, All Participants (n = 18), SL

Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	232675	12168	19.1215	< 2.2e-16 ***
expnov	-50423	16682	-3.0226	.002506 **
sigma	287722	4945	58.1844	< 2.2e-16 ***

Note. The formula in R for this model is: truncreg(formula = duration ~ expnov, data = all.participant.all.levels, point = 80000, direction = "left")

[a]Estimates are in microseconds.
*p < .05. **p < .01. *** p < .005

The final model I chose that best explains the variation in the duration data used trial and level as independent variables with highly significant main effects (p < 1.0 e-05 for both) and a significant interaction term (p < .005). See Table 28, which shows the estimates and significance levels. The expert and novice groups were later further considered separately to address specific hypotheses. At this point, I proceeded to investigate just the effects of level and trial.

Table 27

Truncated Linear Regression Model with Expertise, Level, and Trial as Main Effects on Fixation Duration, All Participants (n = 18), SL

Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	330383.9	27286.9	12.1078	< 2.2e-16 ***
Level	-34115.1	6665.5	-5.1182	3.085e-07 ***
Trial	-23342.1	2863.1	-8.1526	4.441e-16 ***
expnov	-30722.3	22208.6	-1.3834	.1666
sigma	327243.0	7252.5	45.1213	< 2.2e-16 ***

Note. The formula in R for this model is: truncreg(formula = duration ~ Level + Trial + expnov, point = 80000, direction = "left")

[a]Estimates are in microseconds.
*p < .05. **p < .01. *** p < .005

Table 28

Best Fit model: Truncated Linear Regression Model with Level and Trial Main Effects and Interaction Term on Fixation Duration, All Participants (n = 18), SL

Effect	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	330832.6	29163.7	11.3440	< 2.2e-16 ***
Level	-49487.4	11613.6	-4.2612	2.034e-05 ***
Trial	-29467.9	4074.1	-7.2330	4.725e-13 ***
Level:Trial	3903.0	1387.5	2.8130	.004909 **
sigma	325849.0	7105.9	45.8564	< 2.2e-16 ***

Note. The formula in R for this model is: truncreg(formula = duration ~ Level * Trial, point = 80000, direction = "left")

[a]Estimates are in microseconds.
*p < .05. **p < .01. *** p < .005

Unscaffolded (US) versus Fully-Scaffolded (FS) levels. First, I considered the presumably most-different two levels of scaffolding: unscaffolded and fully-scaffolded versions. I used the best fit model equation and just considered those two levels of data. See Table 29. With two levels of scaffolding, I can approximate the duration effect by level and trial linearly. See Figure 6, which shows a linear depiction; the unscaffolded version starts with a higher

Table 29

Truncated Linear Regression Model with Level and Trial Main Effects and Interaction Term on Fixation Duration, Unscaffolded versus Fully-Scaffolded Case, All Participants (n = 18), SL

Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	359697.1	31418.7	11.4485	< 2.2e-16 ***
Level[b]	-33722.7	10113.4	-3.3345	0.0008546 ***
Trial	-20601.6	3978.8	-5.1778	2.245e-07 ***
Level:Trial	2991.5	1143.3	2.6166	0.0088807 **
sigma	279134.2	8445.6	33.0507	< 2.2e-16 ***

Note. The formula in R for this model is the equation for the best fit model chosen above, but the data is a subset of the overall.

[a]Estimates are in microseconds.
[b]US case was set as Level = 0 in the data file and FS Level = 5, therefore the value of the estimate must be multiplied by 5 to get the total effect from US to FS case.
*p < .05. **p < .01. *** p < .005

average duration than the scaffolded version, demonstrating the level effect. Both lines have a negative slope, indicating the trial effect, but the flatter slope of the scaffolded case's line shows the interaction effect, namely that the difference shown due to the scaffolding levels is mitigated by the trial effect, until the trial effect actually outweighs the level effect when the lines cross.

After adjusting for the unit of coefficient estimation, the model estimates an approximate effect of 160 ms shorter duration in the scaffolded version than the scaffolded version for Trial 1. The trial effect is approximately 20 ms shorter duration per trial, with a positive interaction coefficient meaning that the sum of the two effects must be reduced by 3 ms per level and trial when they are both considered. This interaction accounts for the different slopes of the lines in Figure 6. However, none of these values are expected to be predictive, merely representative of relative differences in fixation duration under different conditions.

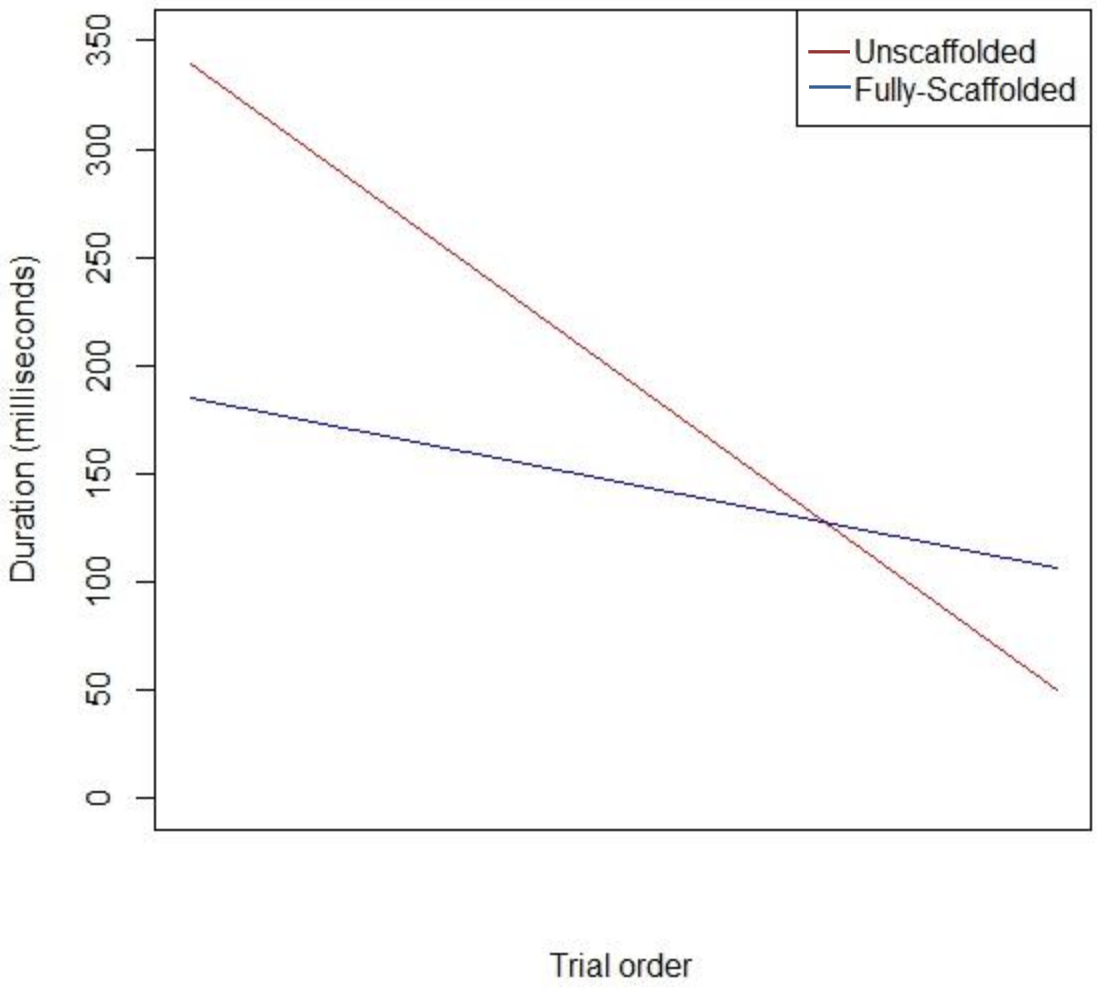


Figure 6. Linear approximation of effect size, fixation duration versus trial, Unscaffolded and Fully-Scaffolded Cases.

Thus, it appears that, when participants look at an unscaffolded visualization, regardless of their expertise level, they have longer fixation durations at the beginning of the experiment than at the end. This indicates that even in the unscaffolded case of very “scientific” visualizations, fixation duration decreases significantly with exposure, up to 300 milliseconds from the first visualization viewed to the last over the course of the experiment. Duration decrease is taken as an indicator that participants are spending more time on

meaning-making due to spending less time decomposing the given visual information (Holmqvist et al., 2011).

For the visualizations that have been scaffolded to ostensibly improve meaning-making, the fixation durations start out significantly lower than those on the unscaffolded visualizations by about 150 ms, meaning that the scaffolding improves opportunities for meaning-making. Fixation duration on scaffolded visualizations also improves with trials, though with an overall decrease over trials of roughly 60 ms, about 20% of the decrease in duration on unscaffolded visualizations due to trial. Finally, the interaction of the two effects of scaffolding (Level) and exposure (Trial) means that the two effects are not summative, but must be reduced by the interaction effect of approximately three milliseconds per trial. Thus, the scaffolded case slope is smaller and duration decreases less over the course of several trials than in the scaffolded case, to the point where eventually the trial effect wins out over the scaffolding level effect. So, given unscaffolded visualizations, one will eventually approach the fixation duration of someone looking at fully-scaffolded visualizations, but due to the higher starting point, it will take several trials to compensate for the lack of scaffolding. Finally, part of the trial effect here may be due to the tendency of participants to compare across visualizations, which was alleviated from the clinical interview question somewhat by the inclusion of three different months' worth of data in the five experimental trials, but perhaps not completely as interviews accompanying eye-tracking shown.

All scaffolding levels. The best fit model when considering all scaffolding levels involved trial number and level of scaffolding with singly-scaffolded CS, TS, and GS treated as the same level as main effects and level and trial interaction. Because I have no reason to assume the intermediate scaffolding "levels" vary linearly, the coefficients of this model are not immediately interpretable when all five levels of scaffolding are included. See Table 30.

Table 30

Best Fit Truncated Linear Regression Model with Level and Trial Main Effects and Interaction Term on Fixation Duration, Unscaffolded, Single Scaffolding, and Fully-Scaffolded Cases, All Participants (n = 18), SL

Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	350650.6	45780.9	7.6593	1.865e-14 ***
Level.dummy[b]	-59187.6	15384.5	-3.8472	0.0001195 ***
Trial	-29748.7	5819.0	-5.1123	3.182e-07 ***
Level.dummy:Trial	4503.0	1815.5	2.4803	0.0131278 *
sigma	327883.0	7202.4	45.5243	< 2.2e-16 ***

Note. The formula in R for this model is: truncreg(formula = duration ~ Level * Trial, point = 80000, direction = "left")

[a]Estimates are in microseconds.
[b]Levels set as follows: US = 0; FS = 5; GS, CS, TS all set = 3
*p < .05. **p < .01. *** p < .005

Truncated regressions were then used to investigate specific comparisons of factors based on hypotheses as shown in Table 5 in Chapter 3. Tests 1-4 of the expert--novice effect proved opposite to what I expected; there was no expert--novice effect, regardless of trial. Tests 5-9 also were opposite expectations as there was no expert--novice effect, regardless of scaffolding level. Test 10 is inconclusive as there was neither a significant expert--novice effect nor an effect of scaffolding level for novices.

Tests 11-17 for novices only showed that only the difference between levels was a trend toward significant difference (p = .08) between the unscaffolded and fully-scaffolded cases, and trial was a significant predictor in all cases for novices, p < .001 as a main effect with level and no interaction. On the other hand, level (p < .001), trial (p < .001), and level-trial interaction (p = .003) were all significant predictors for experts. It seems, therefore, that scaffolding level in the spontaneous looking condition did not alter the durations of the novice's fixations, but exposure to the visualizations over the course of the experiment did. For experts, however, scaffolding level and trial both decreased

the fixation durations. This could mean that without direction as to where to look, for example without a guiding question, novices do not have an expert-like routine to generally make meaning from the visualizations.

Tests 18 and 19 with gender as a main effect for each of the subgroups (expert and novice) showed that gender alone was not a significant effect for either subgroup in the spontaneous looking condition. When topic was considered as a main effect by itself, it was a significant predictor in the SL condition. I found a significant difference among all topics, and pairwise, each topic was significantly different from each other. See Table 31.

Table 31

Truncated Linear Regression Model with Topic as Main Effect, Pairwise Comparisons, All Participants (n=18), SL

Chlorophyll versus SST				
Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	263958.9	26432.3	9.9862	< 2.2e-16 ***
topic	-36069.1	13000.8	-2.7744	.005531 **
sigma	282838.1	7192.2	39.3256	< 2.2e-16 ***
SST anomaly versus SST				
Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	262157.6	56028.7	4.6790	2.883e-06 ***
topic	-56111.4	24921.5	-2.2515	.02435 *
sigma	301501.5	6695.3	45.0318	< 2.2e-16 ***
Chlorophyll versus SST anomaly				
Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	248031.1	29762.5	8.3337	< 2.2e-16 ***
topic	-64248.8	15461.2	-4.1555	3.246e-05 ***
sigma	309743.7	6147.6	50.3845	< 2.2e-16 ***

[a]Estimates are in microseconds.
*$p < .05$. **$p < .01$. *** $p < .005$

Durations for chlorophyll and SST anomaly were significantly longer than for SST, and SST anomaly was shorter than chlorophyll. Thus, SST was the shortest duration, chlorophyll the longest, which suggests that participants struggled most to make meaning of the chlorophyll visualization, and that SST anomaly and SST were somewhat more comprehensible, especially SST.

These results on numbers of fixations and durations revealed the variables that had significant influence on participants' eye-tracking in the spontaneous looking condition, namely level of scaffolding and trial number, and those that did not, namely expertise, gender, and topic. This gave us an overarching best fit model that included main effects of level trial and an interaction term between the two.

However, tests on specific sub-groups did reveal that novices did not show a significant effect of scaffolding level, only trial, and that there were pairwise differences of topic, and topic was significant as a main effect by itself. These results and other questions of element use were considered further by examining quantitative data on the likelihoods of looking at various areas of interest and the influence of the variables on those likelihoods.

Types of scaffolding – Areas of Interest. Fixations on each area of interest (AOI) were computed by BeGaze™ according to the 80 ms minimum and 100 pixel dispersion maximum thresholds. Across all levels, there were 2564 total fixations. The vast majority of fixations fell within the visualization areas (1182 for larger map, 575 for smaller map), and approximately 10% of fixations fell outside of any AOI, in White Space but still on screen (267 of 2564). See Table 32.

Numbers of fixations on AOIs were compared using multinomial logistic regression. This model assumes the summation of percentages of fixations on all AOIs in a trial equals 1.0, and computes the relative likelihood of looking at any given AOI. This model examines the relative risk ratio (RRR) among multiple AOIs, given independent variables, in this case scaffolding level, trial number, expertise, topic, and gender. For example, it considers whether the relative probability of looking at the title versus the map in the FS (fully scaffolded) case (numerator) is significantly different from the relative probability of looking at the title versus the map in the US (unscaffolded) case (denominator) to determine

Table 32

Number of Fixations per Area of Interest, All Participants (n = 18), SL

AOI	totals	AOI	totals
Key		Key	
chlorophyll smaller map key	19	chlorophyll larger map key	52
anomaly smaller map key	46	anomaly larger map key	64
SST smaller map key	18	SST larger map key	22
Scaffolded Title		Unscaffolded Title	
anomaly scaffolded title	73	SST unscaffolded title	8
SST scaffolded title	13	chlorophyll unscaffolded title	13
chlorophyll scaffolded title	39	anomaly unscaffolded title	33
Map		Other	
larger map[a]	1182	anomaly color cutout	55
smaller map	575	anomaly overlap	6
white space[b]	267	chlorophyll color cutout	14
out of range (no AOI)	61	chlorophyll larger map overlap	4
Total smaller map (TS, FS) fixations			806
Total larger map (US, CS, GS) fixations[a]			1378
Total Fixations			2564

[a]The larger map was used in 60% of the stimulus visualizations, while the smaller map was used in 40%, so the cases with the larger map were shown 1.5 times more often to participants, which may account for some differences in the number of fixations.
[b]White space was not divided into instances of the smaller or larger map case and so this total is for all five visualizations.

the effect of scaffolding. See Equation 1.

$$\frac{P(\text{look at title})|\text{Level FS}}{P(\text{look at map})|\text{Level FS}} \quad \frac{P(\text{look at title})|\text{Level US}}{P(\text{Look at map})|\text{Level US}} \tag{1}$$

If the FS probability is significant with a negative coefficient, then in the example, the participant is more likely to look at the fully-scaffolded title than they are to

look at the unscaffolded title versus the map. That is, increasing scaffolding increases number of fixations on the title versus map.

Below, tables are grouped by scaffolding case for discussion. First, I considered unscaffolded (US) versus geographic scaffolding (GS), using level alone, level and expertise, and level and trial as possible factors. See Table 33. For geographic scaffolding, nothing was significant except experts tended to have a smaller probability of looking off-screen in the scaffolded case, when expertise and level were factors ($p < .001$). However, overall, off-screen fixations were minimal compared to overall fixations throughout the experiments.

Next, consider color scaffolding (CS) versus unscaffolded (US) visualizations. See Table 34. In CS, I saw similar patterns to GS for the Level only and Level and Expertise cases; only Expertise in the latter case was significant for lower probability of experts looking at off-screen versus map in the scaffolded version.

Next, consider title scaffolding (TS) of title and measurement units. This also involved a change of map size, with the TS map smaller and the key on the left. Here, more variables had a significant influence on the relative risk ratios of fixations, so consider the cases separately, starting with Level only. Scaffolding level had a significant effect on probability of looking at the title and the key; more scaffolding increased the relative likelihood of looking at the title and key over the map. There was also a significant effect of increased probability of looking at white space as scaffolding increased. See Table 35. This could be simply a reflection of the smaller size of the map portion of the visualization. However, it could indicate an increased chance that participants are looking outside the main map to make meaning, and even possibly are actually looking at other elements, especially given the conservative size of the title and key AOIs. It could also indicate that participants are using peripheral vision to take in information, as per Kim et al. (2012), again possibly actually using visualization elements in that situation.

Table 33

Relative Risk Ratios for Areas of Interest, Unscaffolded versus Geographic Scaffolding, All Participants (n = 18), SL

Level as Main Effect				
Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.98/0.229***	-	-2.17/0.157***	-
Level	-0.17/0.174	0.85(0.6,1.19)	0.1/0.106	1.1(0.89,1.35)
Effect	White Space		Off-screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.45/0.179***	-	-3.49/0.293***	-
Level	-0.07/0.129	0.93(0.73,1.2)	0.05/0.2	1.06(0.71,1.56)
Level and Expertise as Main Effects				
Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.69/0.266***	-	-2.27/0.199***	-
expnov	-0.65/0.357	0.52(0.26,1.05)	0.19/0.213	1.21(0.8,1.84)
Level	-0.17/0.175	0.85(0.6,1.19)	0.1/0.106	1.1(0.89,1.35)
	White Space		Off-screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.37/0.22***	-	-2.86/0.305***	
expnov	-0.16/0.258	0.86(0.52,1.42)	-2.16/0.618***	0.12(0.03,0.39)
Level	-0.07/0.129	0.93(0.73,1.2)	0.06/0.202	1.06(0.71,1.57)
Level and Trial as Main Effects				
	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-3.57/0.409***	-	-2.35/0.236***	-
Trial	0.07/0.037	1.07(1,1.16)	0.02/0.022	1.02(0.98,1.07)
Level	-0.18/0.175	0.83(0.59,1.18)	0.09/0.106	1.09(0.89,1.34)
	White Space		Off-screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.25/0.252***	-	-3.04/0.378***	-
Trial	-0.03/0.027	0.97(0.92,1.03)	-0.07/0.043	0.93(0.86,1.01)
Level	-0.06/0.13	0.95(0.73,1.22)	0.09/0.202	1.09(0.74,1.63)

*$p < .05$. **$p < .01$. *** $p < .005$

Table 34

Relative Risk Ratios for Areas of Interest, Unscaffolded versus Colors Scaffolding, All Participants (n = 18), SL

Level as Main Effect				
	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.98/0.229***	-	-2.17/0.157***	-
Level	-0.01/0.328	0.99(0.52,1.88)	-0.18/0.235	0.83(0.52,1.32)
	White Space		Off Screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.45/0.179***	-	-3.49/0.293***	-
Level	-0.54/0.295	0.58(0.33,1.04)	-0.14/0.434	0.87(0.37,2.03)
Level and Expertise as Main Effects				
	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.76/0.27***	-	-2.18/0.201***	-
expnov	-0.46/0.338	0.63(0.33,1.23)	0.03/0.234	1.03(0.65,1.63)
Level	-0.04/0.329	0.96(0.5,1.83)	-0.18/0.236	0.83(0.53,1.32)
	White Space		Off Screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.46/0.235***	-	-2.84/0.306***	-
expnov	0.02/0.285	1.02(0.58,1.78)	-2.3/0.746**	0.1(0.02,0.43)
Level	-0.54/0.296	0.58(0.33,1.04)	-0.26/0.438	0.77(0.33,1.82)
Level and Trial as Main Effects				
	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-3.67/0.42***	-	-2.37/0.255***	-
Trial	0.08/0.038*	1.09(1.01,1.17)	0.03/0.026	1.03(0.98,1.08)
Level	0.05/0.332	1.05(0.55,2.01)	-0.18/0.235	0.84(0.53,1.33)
	White Space		Off Screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.39/0.278***	-	-3.79/0.489***	-
Trial	-0.01/0.031	0.99(0.93,1.05)	0.04/0.048	1.04(0.95,1.14)
Level	-0.54/0.295	0.58(0.33,1.04)	-0.13/0.435	0.88(0.37,2.07)

*$p < .05$. **$p < .01$. *** $p < .005$

Table 35

Relative Risk Ratios for Areas of Interest with Level as Main Effect, Unscaffolded versus Title Scaffolding, All Participants (n = 18) , SL

Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.98/0.229***	-	-2.17/0.157***	-
Level	0.55/0.088***	1.73(1.45,2.05)	0.26/0.069***	1.3(1.14,1.49)
Effect	White Space		Off Screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.45/0.179***	-	-3.49/0.293***	-
Level	0.44/0.072***	1.55(1.35,1.79)	-0.26/0.194	0.77(0.53,1.13)
*p < .05. **p < .01. *** p < .005				

For Level and Expertise as main effects, again, I found significant effects of level of scaffolding on the relative probabilities of looking at the title, key and white space compared to the probability of looking at the map. All increased as level increased, as shown in Table 36.

Table 36

Relative Risk Ratios for Areas of Interest with Level and Expertise as Main Effects, Unscaffolded versus Title Scaffolding, All Participants (n = 18) , SL

Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept) -	-2.81/0.252***	-	-2.08/0.187***	-
expnov	-0.34/0.227	0.71(0.45,1.11)	-0.17/0.203	0.84(0.57,1.26)
Level	0.55/0.088***	1.73(1.45,2.05)	0.26/0.069***	1.3(1.14,1.49)
Effect	White Space		Off Screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.37/0.205***	-	-2.85/0.312***	-
expnov	-0.15/0.199	0.86(0.58,1.27)	-2.09/0.76**	0.12(0.03,0.55)
Level	0.44/0.072***	1.55(1.35,1.79)	-0.26/0.195	0.77(0.53,1.13)
*p < .05. **p < .01. *** p < .005				

Finally, when I considered Level and Trial as main effects, Level was significant again for the relative likelihood of looking at the title, key, and white space over looking at the map. Trial was significant for the overlap area and

white space, both minimal numbers of fixations compared to overall fixations. See Table 37.

Table 37

Relative Risk Ratios for Areas of Interest with Level and Trial as Main Effects, Unscaffolded versus Title Scaffolding, All Participants (n = 18) , SL

Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-3.15/0.3***	-	-2.08/0.218***	-
Trial	0.02/0.024	1.02(0.97,1.07)	-0.01/0.021	0.99(0.95,1.03)
Level	0.54/0.088***	1.72(1.45,2.05)	0.26/0.069***	1.3(1.14,1.49)
Effect	White Space		Off-screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.13/0.225***	-	-3.13/0.432***	-
Trial	-0.05/0.021*	0.96(0.92,1)	-0.05/0.052	0.95(0.86,1.05)
Level	0.45/0.072***	1.57(1.36,1.81)	-0.24/0.195	0.78(0.54,1.15)

*$p < .05$. **$p < .01$. *** $p < .005$

Finally, I considered the AOI relative probabilities for the unscaffolded compared to the fully-scaffolded case, with the same combinations of level, expertise, and trial considered as potential main effects. See Table 38. With level as the only factor, increased scaffolding significantly increased the relative likelihood of looking at the title, key, and white space over the map, as well as the likelihood of looking off-screen. The same pattern held true with expertise in the model. In addition, expertise accounted for a lower tendency to look off-screen. Finally, the same pattern for level held true in the level and trial model. Trial also accounted for a higher probability of looking at the title, key, and white space in the scaffolded stimuli, suggesting participants altered their patterns of looking as they learned the task over the course of the experiment.

Visualization main idea condition. Next, I considered eye-tracking data for the time each participant spent viewing the visualization while they were asked and answered the question “What is the main idea of this image?” Data were entered for analysis as the main idea (MI) condition. 20 total participants

Table 38

Relative Risk Ratios for Areas of Interest, Unscaffolded versus Fully Scaffolded, All Participants (n = 18), SL

Level as Main Effect				
Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.98/0.229***	-	-2.17/0.157***	-
Level	0.25/0.055***	1.28(1.15,1.42)	0.18/0.04***	1.19(1.1,1.29)
	White Space		Off Screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.45/0.179***	-	-3.49/0.293***	-
Level	0.26/0.043***	1.29(1.19,1.41)	0.17/0.074*	1.19(1.03,1.37)
Level and Expertise as Main Effects				
Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-3.16/0.277***	-	-1.96/0.18***	-
expnov	0.33/0.257	1.39(0.84,2.3)	-0.43/0.198*	0.65(0.44,0.96)
Level	0.25/0.055***	1.28(1.15,1.43)	0.18/0.04***	1.19(1.1,1.29)
Effect	White Space		Off Screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.48/0.209***	-	-2.97/0.31***	
expnov	0.06/0.198	1.06(0.72,1.56)	-1.42/0.433**	0.24(0.1,0.57)
Level	0.26/0.043***	1.29(1.19,1.41)	0.17/0.074*	1.18(1.02,1.37)
Level and Trial as Main Effects				
Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.53/0.284***	-	-2.01/0.215***	-
Trial	-0.07/0.028*	0.94(0.89,0.99)	-0.02/0.021	0.98(0.94,1.02)
Level	0.26/0.056***	1.3(1.16,1.45)	0.18/0.04***	1.2(1.11,1.3)
Effect	White Space		Off Screen	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-1.83/0.218***	-		-3.3/0.393***
Trial	-0.1/0.023***	0.91(0.87,0.95)	-0.03/0.038	0.97(0.9,1.05)
Level	0.28/0.044***	1.32(1.21,1.44)	0.18/0.074*	1.19(1.03,1.38)
*p < .05. **p < .01. *** p < .005				

answered questions while undergoing eye-tracking, including two (one expert, one novice) who did not do the spontaneous looking condition. When these two

participants were removed from the analysis, the patterns of significance remained, so they were left in the analysis. Similarly, one expert could not be calibrated while wearing glasses, but offered to participate anyway. Removing his eye-tracking results did not affect the patterns of significance, either, so the data was left in the experiment.

Trial duration for the main idea question condition varied based on the length of answer given as described in the methods. Average was approximately 17 s. The durations were not normally distributed, so I conducted a Wilcoxon Rank Sum test to determine whether the average durations of the groups were significantly different; they were not W_s (n_1 = 50, n_2 = 50) = 1193, p = 0.69). See Table 39. The shortest trial was about one second, when the expert participant's

Table 39

Mean Trial Duration (ms), All Participants (*N* = 20)*, All Novices, and All Experts, MI*

Statistic	All Participants (*N* = 20)	Novice Participants (*N* =10)	Expert Participants (*N* =10)
Mean	16841.73	16173.88	17474.2
Maximum	50835	50835	44950
Minimum	1012	6024	1012

answer was "SST anomaly" for a later trial where she likely anticipated the question after the SL condition. The longest trial was nearly 51 seconds, where the novice participant, Kyle, hesitated and hemmed and hawed while answering, as he apparently worked to make meaning:

> Kyle: Oh, OK. ... This key indicates that ... the white - is the average. So ... it's varying from the normal average.
> Researcher: OK, so what -
> Kyle: (interrupts) So like the dark red isn't actually, uh, plus five degrees Celsius, it's, uh, ... five degrees higher than the average.
> Researcher: OK, so then what's this image about?
> Kyle: Uh, Over a month, ... how, uh ... the ocean degrees has changed ... from the average. ... Well the previous average, I guess. (eye-tracking interview)

There was also a possible influence of the length of the time taken to ask the question by the interviewer on all trials. This time was not discarded, however, as the interviewer standardized the asking as much as possible, and it is presumed that the participant began thinking about the answer as soon as the interviewer began asking the question, and thus their eye movements reflected their deductive process from the beginning of the asking.

Descriptive statistics, fixations. Each participant's total number of fixations and duration of fixations for all five experimental visualizations were calculated by BeGaze™. The mean for all participants was 42 fixations per trial, with a median of 34. Experts alone had a higher average number of fixations at 47 versus 37. As the numbers of fixations within the two groups did not appear normally distributed, I conducted a Wilcoxon Rank Sum test to determine whether the average number of fixations were significantly different; they were not, *Ws* (n_1 = 50, n_2 = 50) = 1508, p = .08. See Table 40.

Table 40

Number of Fixations per Trial, All Participants (N = 20), MI

Statistic	All Participants (*N* = 20)	Novice Participants (*N* = 10)	Expert Participants (*N* = 10)
Total	4201	1867	2334
Mean	42.01	37.34	46.68
Maximum	104	104	100
Minimum	4	13	4
Median	34	31.5	44

Note. Trial length varied due to different lengths of participants' answers.

Durations for the overall group and experts and novices alone were produced similarly and descriptive statistics calculated, with a difference in the maximum duration of almost 2.5 s longer for the novices than experts, and a difference in average duration of 60 ms longer for novices. See Table 41. Compared to the durations of the spontaneous looking case, experts had similar

Table 41

Fixation Durations (ms) for All Participants, (N = 20), All Novices, and All Experts, MI

Statistic	All Participants (*N* = 20)	Novice Participants (*N* = 10)	Expert Participants (*N* = 10)
Mean	352.6706	320.3492	393.0766
Maximum	4705	2219	4705
Minimum	81	83	81
Median	275	258	292

Note. Fixations were defined as a minimum dwell time of 80 ms.

average durations, at about 320 ms in each case. However, novices spent nearly 75 ms more on average when asked to assess the visualization main idea. I also tested the mean fixation duration of the two groups and found the data are indeed highly significantly different, *Ws* (n_1 = 1867, n_2 = 2334) = 2443391, p = 1.45 e-11.

Participant fixation duration histograms and density curves showed similar patterns to the SL condition; some participants appeared roughly normal, but many appeared long-tailed. As with spontaneous looking, correlations of duration and previous fixation's duration were only .14 for the main idea question data, and a joint distribution plot showed random relationship between duration and previous fixation's duration. Thus, duration data were treated as truncated normal and independent for analysis with truncated linear regression to model the duration data in the same manner as the spontaneous looking condition.

Regression modeling. As before, the five independent variables were examined as possible predictors of duration: Level of scaffolding, Trial number, Expertise, Topic, and Gender. Thus, the first model considered all these factors, without interactions. In this case, gender and expertise were highly significant (both $p < .004$), and topic was significant ($p = .02$). Level and trial were not

significant when participants were asked to assess the visualization main idea. See Table 42. When expertise alone was considered, it remained a significant

Table 42.

Truncated Linear Regression Model with Five Independent Variables, All Participants (N = 20), MI

Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	-18617.156	6604.621	-2.8188	.004820 **
Level	130.786	117.624	1.1119	.266179
Trial	-44.801	42.361	-1.0576	.290233
gender	6351.586	2161.892	2.9380	.003304 **
expnov	-4555.021	1563.638	-2.9131	.003579 **
topic	-1247.657	524.658	-2.3780	.017405 *
sigma	1862.177	322.081	5.7817	7.395e-09 ***

Note. The equation for this model is the same as in Table 23.

[a]Estimates are in microseconds.
*p < .05. **p < .01. *** p < .005

explanatory factor (p = .009). Since gender and topic were not expected to be predictive variables for this experiment, I selected the expertise-only model as the best fit for the data.

The expert and novice groups were again considered separately with truncated regressions as part of testing of specific hypotheses about fixation durations in the main idea question conditions (see Table 5 in chapter 3). The most interesting result was Test 10 between the novices viewing the fully-scaffolded version and the experts viewing an unscaffolded version; there was no significant difference between the novices' duration looking at fully-scaffolded visualizations and the experts looking at the unscaffolded visualizations (p > .16 for all individual variables as main effects). This suggests that novices might be able to look at fully-scaffolded visualizations similarly to experts looking at unscaffolded visualizations, an indication that scaffolding is assisting in novice meaning making. This is reflected in their improved score performance as discussed in Chapter 4.

Tests 1 and 4 again were opposite to what I expected; there was no expert--novice effect on the visualizations viewed first or last. However, Tests 2 and 3 showed no effect of trial, so this control group hypothesis was proven. Tests 5-9 remained opposite expectations, but this time due to no effect of level of scaffolding. Tests 11-17 for novices only showed no level effects in the main idea question condition.

Finally, tests 18 and 19 with gender as a main effect for each of the subgroups (expert and novice) showed that gender alone was a highly significant effect for novices ($p < .001$) and experts ($p < .007$). The only model with gender and another variable that showed significance for both (besides expertise) in the overall dataset was topic, with gender significant at $p = .001$ and topic at $p = .02$. See Table 43.

Table 43

Truncated Linear Regression Model with Topic and Gender Main Effects, All Participants (*N* = 20)*, MI*

Effect	Estimate[a]	Std. Error	t-value	Pr(>|t|)
(Intercept)	-22786.16	7117.13	-3.2016	.001367 **
gender	5998.10	1851.94	3.2388	.001200 **
topic	-1093.58	469.77	-2.3279	.019916 *
sigma	2013.60	308.64	6.5242	6.838e-11 ***

Note. The formula for this model in R is: truncreg(formula = duration ~ gender + topic, data = Q.all.participant.all.levels, point = 80, direction = "left")

[a]Estimates are in microseconds.
*$p < .05$. **$p < .01$. *** $p < .005$

Neither of these were expected to be predictor variables in this experiment, and at this point, I have no way to account for the significant factor of gender in the existing statistical data for the novices, but in the discussion I will return to the differences in other data that may provide clues to the statistical difference. The difference within the expert population could be due to differences in expertise; the female experts for the eye-tracking experiment do their research on seafloor sediments, rather than on the fluid systems that many

of the male participants investigate. In addition, since only two of the ten expert participants were female in the eye-tracking experiment, the individual differences could overwhelm the data. This could be compounded by the fact that one female participant, Maureen, had little data for the fourth visualization she was shown as the tracking system stopped capturing her. Despite a reasonable capture rate overall, dropping Maureen from the expert-only gender analysis brought the significance of gender to borderline at $p = .0503$. Topic as a significant factor is not altogether surprising given the interview results that showed temperature was a relatively familiar topic, while temperature anomaly and chlorophyll or ocean productivity were less familiar to novices, as discussed previously and further below.

When topic was considered, unlike the SL case, topic alone was not a significant predictor in the main idea question condition ($p = .12$). However, pairwise SST and chlorophyll were each significantly different from SST anomaly ($p = .044$ and $p = .041$), but not from each other ($p = .12$). Durations in the SST anomaly case were significantly longer than for SST or chlorophyll, indicating participants had more difficulty making meaning from those visualizations. This could be due to conflation between average and anomaly temperature, especially given the red-blue color scale of the anomaly condition which many participants associated with temperature, and the lack of familiarity of several participants with the word ‘anomaly’. They could be interpreting the SST part of ‘SST Anomaly’ correctly and simply ignoring what they do not know. However, the experts also tended to answer simply ‘global ocean temperatures’ without explicitly mentioning difference from normal when asked about the main idea. The difficulty encountered by participants revealed by higher fixation duration suggests, on the other hand, that especially on the part of novices, it may truly be a lack of understanding, rather than a simple failure to be explicit, on their part.

Care must be taken when interpreting these differences in topic, however, as for the novices, six were shown the anomaly condition in the eye-tracking, and

two each were shown the SST and chlorophyll visualizations, based on their previous experience in the clinical interviews and random selection for recruitment for the eye-tracking. The experts, meanwhile, broke down to three chlorophyll, four SST anomaly, and three SST, giving overall more data for the anomaly case (*n* = 10 participants) versus either of the other two (*n* = 5 each).

Types of scaffolding - Areas of Interest analysis. Areas of interest were the same as for the spontaneous looking. As before, fixations on each area of interest (AOI) were computed by BeGaze™ according to the 80 ms minimum and 100 pixel dispersion maximum thresholds. Across all levels, there were 4201 total fixations. Again, the vast majority (73%) of fixations fell within the map areas, 1963 in the larger map case and 1088 in the smaller map case, and approximately 10% of fixations fell outside of any AOI, in White Space (450 of 4201). No “out of range” fixations were reported in this condition. See Table 44. When duplicates were reported in the main idea question condition, four SST, 26 chlorophyll, and 19 SST anomaly overlap fixations were discarded out of 1088 total fixations for the smaller map (TS, FS scaffolding) main idea question data.

Numbers of fixations on AOIs were compared using multinomial logistic regression. Below, tables are shown for individually- and fully-scaffolded cases for discussion. For the main idea question experiment, all four showed slightly different patterns of significance, but all patterns in the direction of experts more likely (negative coefficients) to use the other elements of the figure (title and key) compared to the map. In all three singly-scaffolded conditions, increasing expertise meant significantly higher likelihood of looking at the title or white space as compared to the map. See Table 45.

For the fully-scaffolded case, the experts had a significantly increased relative likelihood of looking at the key or white space instead of the map, but the likelihood of looking at the title was not significantly more or less likely than the likelihood of looking at the map. See Table 46. With expertise as the only factor,

Table 44

Number of Fixations per Area of Interest, All Participants (N = 20), MI

AOI	totals	AOI	totals
Key		Key	
chlorophyll smaller map key	36	chlorophyll larger map key	82
anomaly smaller map key	74	anomaly larger map key	99
SST smaller map key	25	SST larger map key	73
Scaffolded Title		Unscaffolded Title	
anomaly scaffolded title	80	SST unscaffolded title	18
SST scaffolded title	66	chlorophyll unscaffolded title	30
chlorophyll scaffolded title	52	anomaly unscaffolded title	42
Map		Other	
larger map[a]	1963	anomaly color cutout	14
smaller map	1088	anomaly overlap	8
white space[b]	450	chlorophyll color cutout	1
Total smaller map (TS, FS) fixations			1436
Total larger map (US, CS, GS) fixations[a]			2315
Total Fixations			4201

[a]The larger map was used in 60% of the stimulus visualizations, while the smaller map was used in 40%, so the cases with the larger map were shown 1.5 times more often to participants, which may account for some differences in the number of fixations.
[b]White space was not divided into instances of the smaller or larger map case and so this total is for all five visualizations.

full scaffolding significantly increased the relative likelihood of looking at the key, and white space over the likelihood of looking at the map. The increased likelihood of the experts looking at the key may reflect some of their relative unfamiliarity with the color scales I presented. The increased chance of looking at white space may reflect the conservative size of the visualization element AOIs; participants may have been using peripheral vision rather than fixating directly on the element, in line with findings by Kim, et al. (2012). It might also reflect the

Table 45

Relative Risk Ratios for Areas of Interest with Expertise as Main Effect, Unscaffolded versus Single-Scaffolding Cases, All Participants (N = 20), MI

Geographic Scaffolding				
Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.56/0.155***	-	-2.13/0.127***	-
expnov	-1.47/0.311***	0.23(0.13,0.42)	-0.14/0.173	0.87(0.62,1.22)
	White Space			
	Coeff./SE	RRR(95%CI)		
(Intercept)	-2.38/0.142***	-		
expnov	-0.28/0.199	0.75(0.51,1.11)		
Color Scaffolding				
Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.46/0.15***	-	-1.89/0.116***	-
expnov	-1/0.268***	0.37(0.22,0.62)	-0.15/0.163	0.86(0.63,1.19)
	White Space			
	Coeff./SE	RRR(95%CI)		
(Intercept)	-2.25/0.137***	-		
expnov	-0.04/0.188	0.96(0.66,1.38)		
Title Scaffolding				
Effect	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-1.72/0.123***	-	-2.05/0.142***	-
expnov	-1.11/0.199***	0.33(0.22,0.49)	-0.16/0.184	0.85(0.59,1.23)
	White Space			
	Coeff./SE	RRR(95%CI)		
(Intercept)	-1.44/0.11***	-		
expnov	-0.73/0.159***	0.48(0.35,0.66)		
*$p < .05$. **$p < .01$. *** $p < .005$				

experts' tendencies to use more of the map than the novices do, as discussed in the section on the qualitative eye-tracking results.

In summary, the differences and similarities in number and duration of fixations among participant groups (by expertise and by gender), scaffolding levels, trials, and topics of the visualizations complement and extend the interview data. Specifically, they show what differences exist in patterns of

Table 46

Relative Risk Ratios for Areas of Interest with Expertise as Main Effect, Unscaffolded versus Fully-Scaffolded Case, All Participants (N = 20), MI

	Title		Key	
	Coeff./SE	RRR(95%CI)	Coeff./SE	RRR(95%CI)
(Intercept)	-2.02/0.125***	-	-1.81/0.114***	-
expnov	-0.34/0.178	0.71(0.5,1.01)	-0.63/0.173***	0.53(0.38,0.75)
	White Space			
	Coeff./SE	RRR(95%CI)		
(Intercept)	-1.53/0.102***	-		
expnov	-0.59/0.152***	0.55(0.41,0.75)		

*$p < .05$. **$p < .01$. *** $p < .005$

attention as measured by the time spent looking at the various versions of the stimuli. Compared to novices, the experts generally spend less time per fixation and have more fixations per visualization when specifically asked to answer questions about the visualizations, indicating greater meaning-making (Holmqvist et al., 2011), in line with interview results. The quantitative data give specific insights into the role of experience, as evidenced by the significant influence of trial on lowering the durations of fixations, and of scaffolding, as evidenced by similar decreased durations in the fully-scaffolded versus unscaffolded cases. These were especially evident in the case of spontaneous looking.

However, there are some questions that the quantitative eye-tracking data did not resolve, especially as regards the use of particular elements of the visualizations that the differences in duration based on scaffolding suggest. Thus, I explored the scan paths, that is, the individual fixations, their order, and durations in a visual overlay on the stimulus visualizations to elucidate those differences as well as other expert-novice differences that may not have affected overall quantitative measures.

Qualitative Results

I used qualitative visualizations produced in BeGaze™ to investigate more deeply the statistical findings of the eye-tracking around scaffolding in the spontaneous looking condition. BeGaze™ generates visualizations of scan path, with fixation-centered circles which vary in size based on duration and are connected by lines indicating the paths between them. I used SST as the base visualization and mocked all scan paths as if they were on that visualization; that is, data from a participant who looked at versions of the chlorophyll and anomaly visualizations were overlaid onto the SST visualization in order to plot all participants on the same visualization. Scaffolding levels were collapsed as well, so that trial effect was minimized through randomization. That is, regardless of order the participant was shown a particular version, say fully-scaffolded, the scan paths of the fully-scaffolded case were considered together. This allows us also to look at universal viewing patterns within the map AOI in particular, which are not revealed by the statistical analysis.

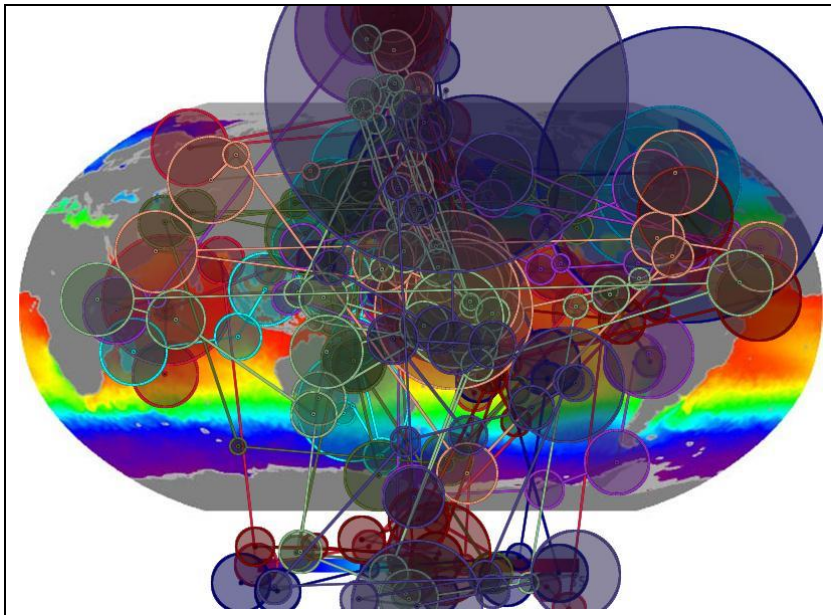
The first comparison was between the unscaffolded and fully-scaffolded visualizations. Novices seem to spend a significant portion of their first ten seconds, indicated by the preponderance of large circles, looking at the Western Hemisphere, which is on the right-hand side of this Pacific-ocean-basin visualization. The remaining circles in the novice scan path visualization are smaller. Larger circles and longer fixation times indicate confusion, suggesting perhaps that novices are struggling to understand even the unusual geography of the Pacific-Ocean-basin centered projection used here. This focus is despite the colored data portion being centered in the visualization, suggesting participants must deliberately look right in order to orient themselves. In fact, besides the colors, which were familiar to many participants as temperature indicators, the shapes of the continents, particularly of North and South America were probably the only things in the unscaffolded condition that novices recognized and could latch onto to orient themselves (as explored in the

interview data write up in chapter 4), so their bias is not surprising. Interviews suggested that participants, especially novices, are more familiar with the Western and Northern Hemisphere patterns of the oceans as well.
There are also a large number of large circles on the title, again indicating confusion among novices as to the meaning of the abbreviated (“SST”) or jargon-filled (chlorophyll, anomaly) titles. Time spent looking at the title leaves little time to even begin to look at the data-filled portions of the map, namely the oceans, though several, but not all, do notice and fixate on the key. As argued previously, time spent looking at the title may have taken time away from time spent making meaning from the map itself.

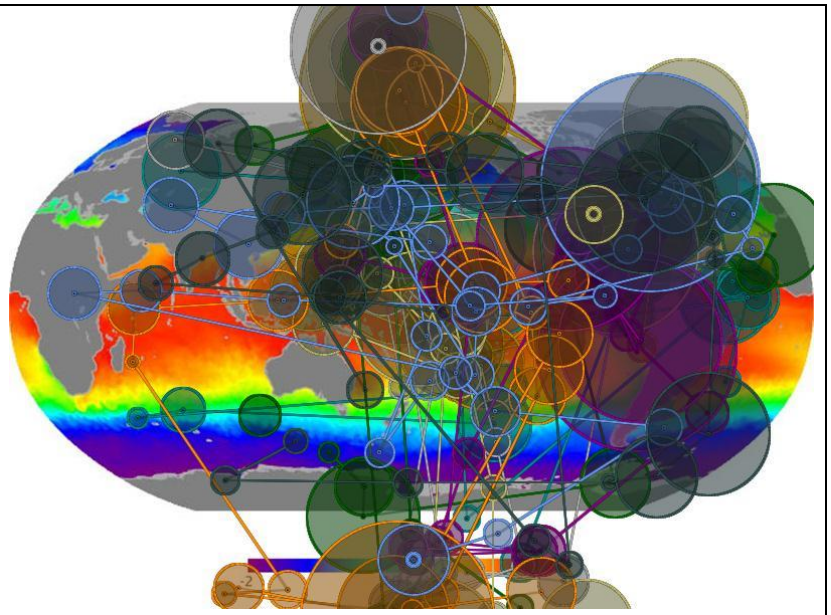
On the other hand, the experts have more uniformly sized and uniformly distributed circles, with more focus on the data in the ocean, aside from two participants who had unusually large fixations in the Northern hemisphere. See Figure 7.

Novice scan paths seem much more like those of the experts when the fully-scaffolded visualizations are examined. See Figure 8. As the figure indicates, novices spend more time focusing on the ocean, and perhaps more time on the title and key. Also, tellingly, the fixations on the title are smaller, suggesting easier, more familiar language, if not more meaning-making as evidenced by the interviews.

Next I examined the effects of individual types of scaffolding on the scan paths, comparing them to the unscaffolded version scan paths for the novices. Here again, novices have a tendency toward the Western Hemisphere, though the tendency seems slightly reduced in the color scaffolding case. See Figure 9. This could be an indication of the familiarity of the colors of the unscaffolded visualization, which in all cases are the same “rainbow” depicted in the unscaffolded visualization, albeit with different meanings. If, as the interviews also suggest, participants are in fact accustomed to associating that color scale with temperature, they may find themselves in the color-scaffolded case taking

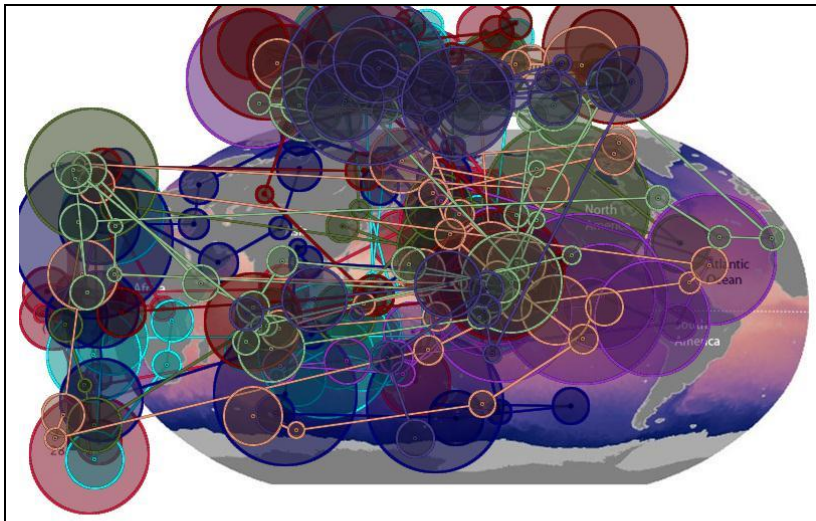


Experts

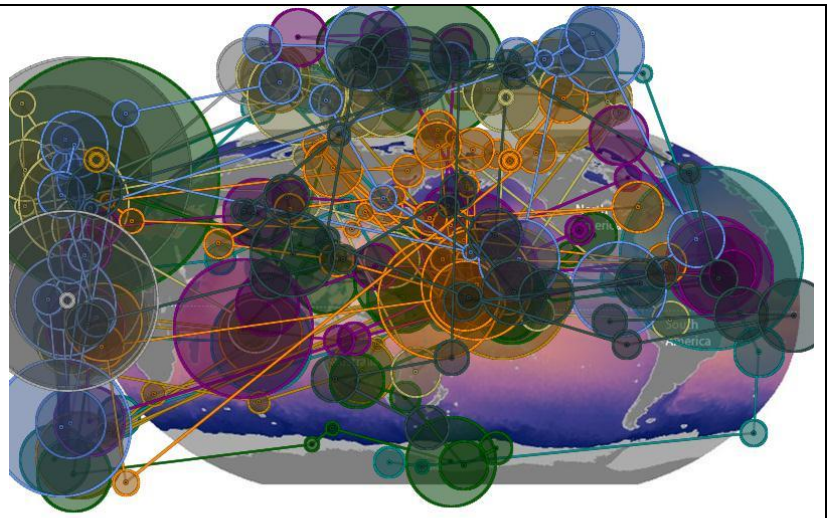


Novices

Figure 7. All experts' versus all novices' eye-tracking scan paths, Unscaffolded, SL. Each color represents an individual expert or novice participant. The same colors depict the same participants throughout the figures.

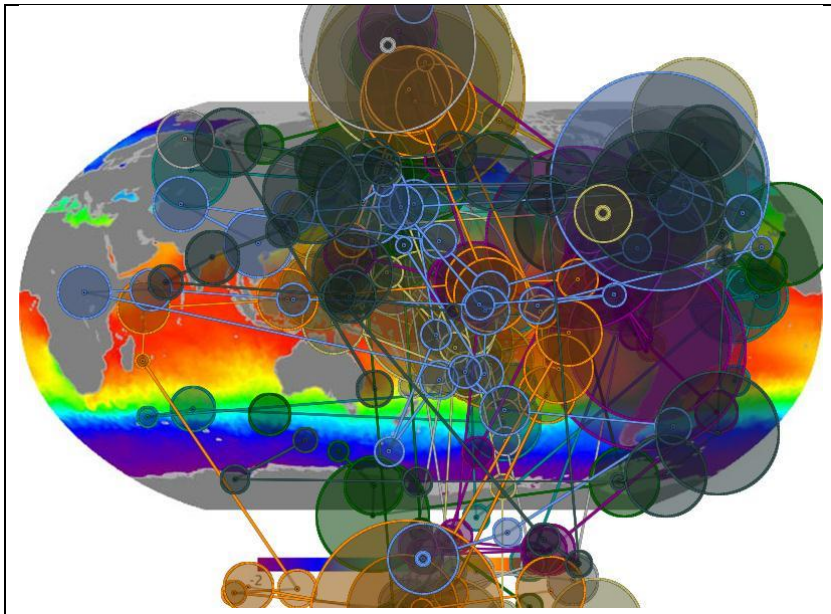


Experts

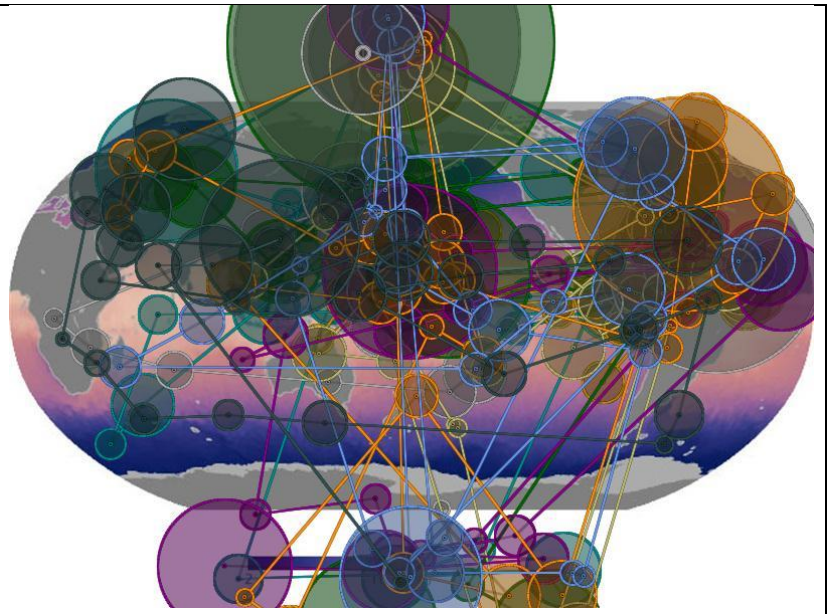


Novices

Figure 8. All experts' versus all novices' eye-tracking scan paths, Fully-Scaffolded, SL. Each color represents an individual expert or novice participant. The same colors depict the same participants throughout the figures.



Unscaffolded



Color Scaffolding

Figure 9. All novices' eye-tracking scan paths, Unscaffolded versus Color Scaffolding, SL. Each color represents an individual novice participant. The same colors depict the same participants throughout.

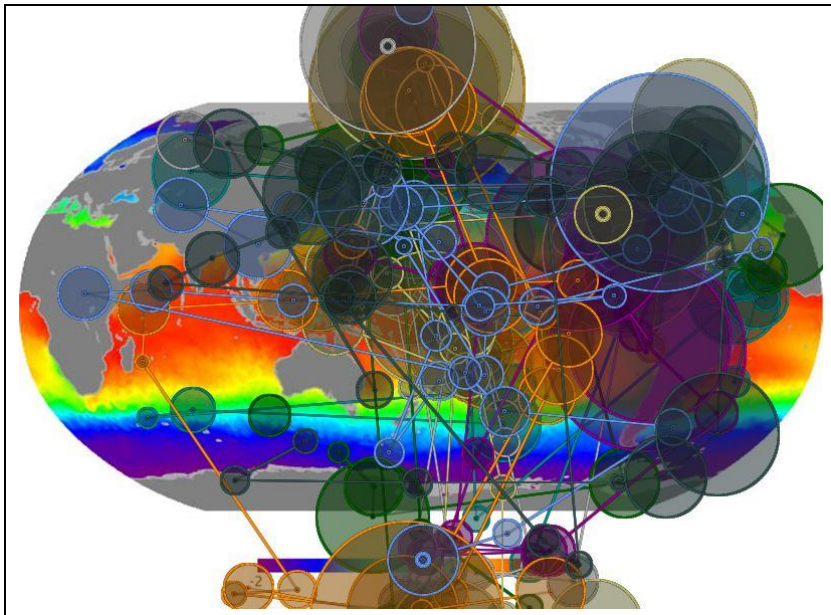
more time to try and make meaning from the patterns. Thus, in the unscaffolded case for chlorophyll and SST anomaly, instead of either accepting the rainbow color scale as temperature but in unfamiliar patterns or dismissing the pattern as unintelligible, the novices might be spending time sorting through the various conflicting associations from their prior knowledge and cultural experiences by looking longer and more comprehensively at the patterns.

When offered the unscaffolded visualization plus geographic labels, novice scan path patterns were also different. See Figure 10. The fixations on the Western Hemisphere became shorter, and more fixations appeared on the Eastern Hemisphere, especially on the Indian Ocean basin where virtually none occurred in the unscaffolded case.

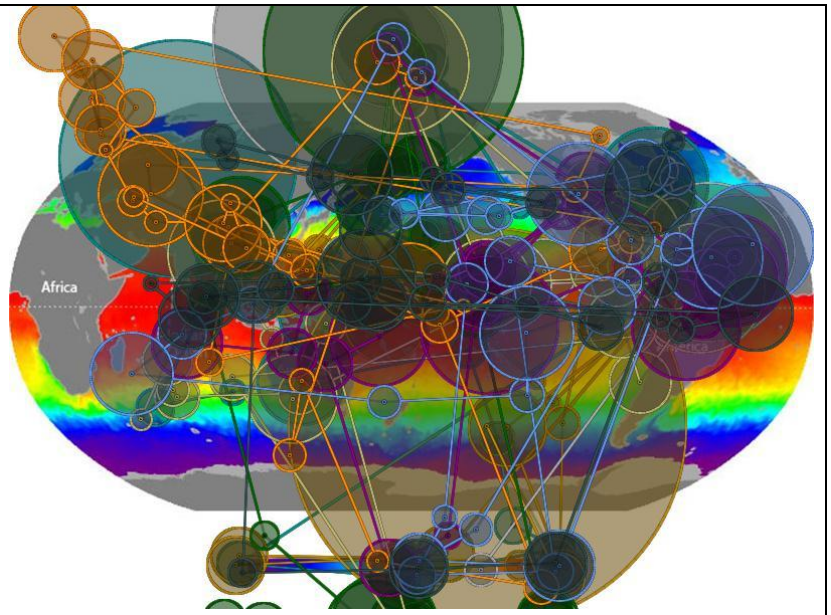
When titles and measurement units were scaffolded, there seemed to be heavier key use, including some confusion represented by the larger circles. There were also more but smaller fixations on the title, and fewer fixations on the Western Hemisphere. However, virtually all of the Southern Hemisphere was overlooked in this case, possibly due to a lack of familiarity with the area, or simply a lack of time to consider the unfamiliar patterns given more opportunity to make meaning from the key or title. See Figure 11.

Experts, on the other hand, had relatively similar scan paths in both the unscaffolded and fully-scaffolded cases, in accordance with their familiarity with not only these types of illustrations, but the orientation, emphasis on the ocean, and meaningfulness of the title and key. See Figure 12. These scan paths by experts also give clues to the significant statistical differences of likelihood of looking at white space. In both cases, they tended to look in the area between the top of the map and the title. See Figure 13 for an example of a single expert's scan path as illustration.

Since the AOIs were drawn conservatively, these fixations could be either indicating that they were looking at the no data areas of the northern hemisphere or looking at the title somewhat obliquely (Kim et al., 2012). As participants were

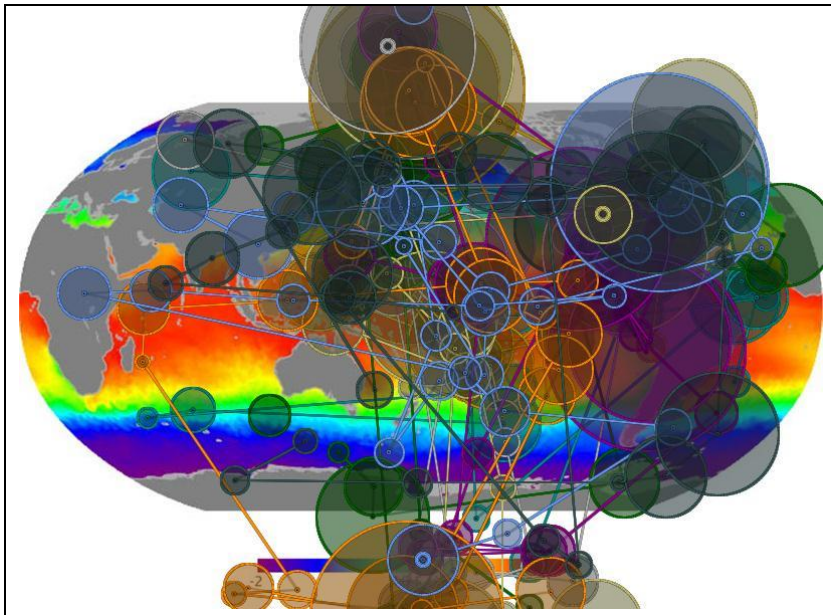


Unscaffolded

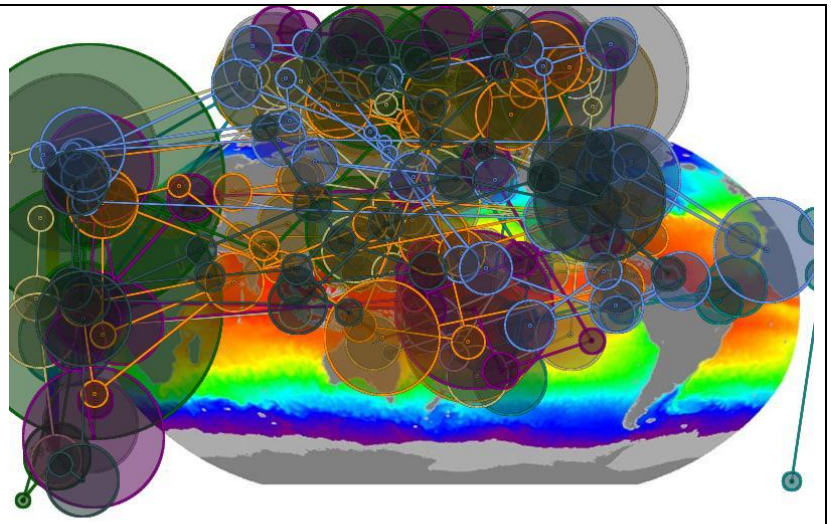


Geographic Labels Scaffolding

Figure 10. All novices' eye-tracking scan paths, Unscaffolded versus Geographic Scaffolding, SL. Each color represents an individual novice participant. The same colors depict the same participants throughout.

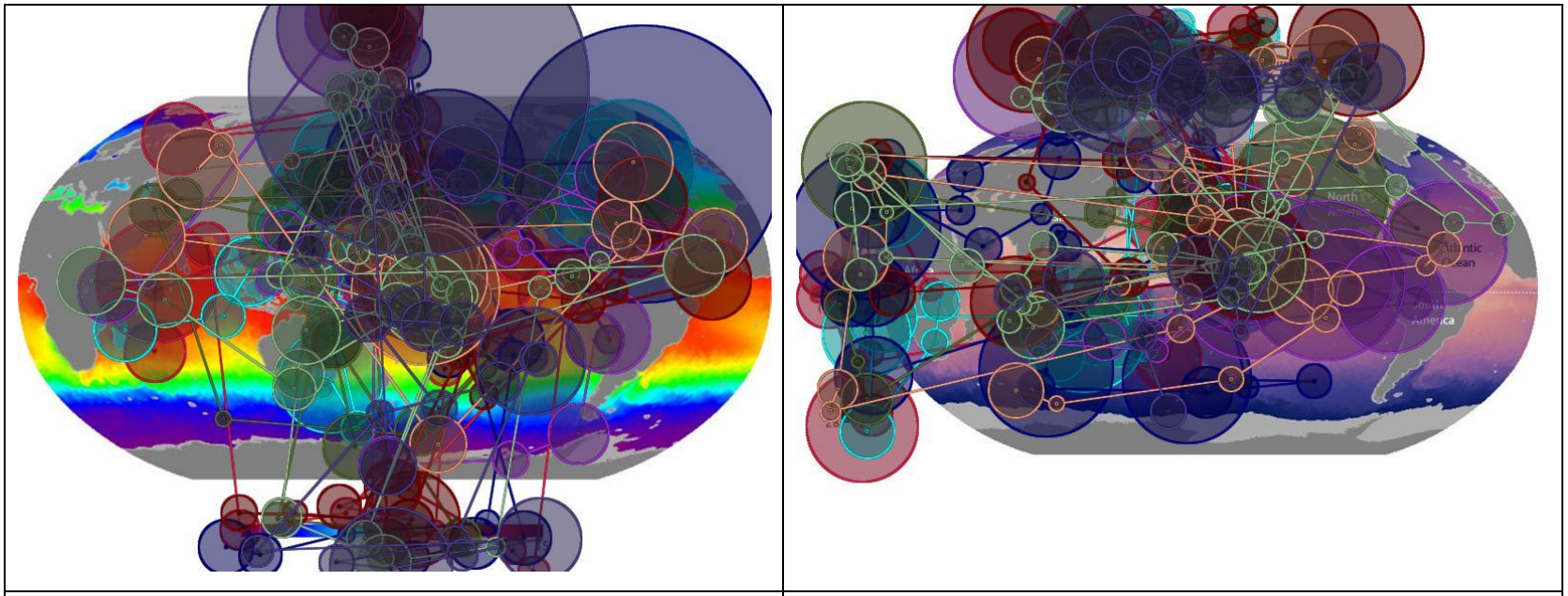


Unscaffolded



Title Scaffolding

Figure 11. All novices' eye-tracking scan paths, Unscaffolded versus Title Scaffolding, SL. Each color represents an individual novice participant. The same colors depict the same participants throughout.



Unscaffolded	Fully-Scaffolded

Figure 12. All experts' scan paths, Unscaffolded versus Fully-Scaffolded, SL. Each color represents an individual novice participant. The same colors depict the same participants throughout.

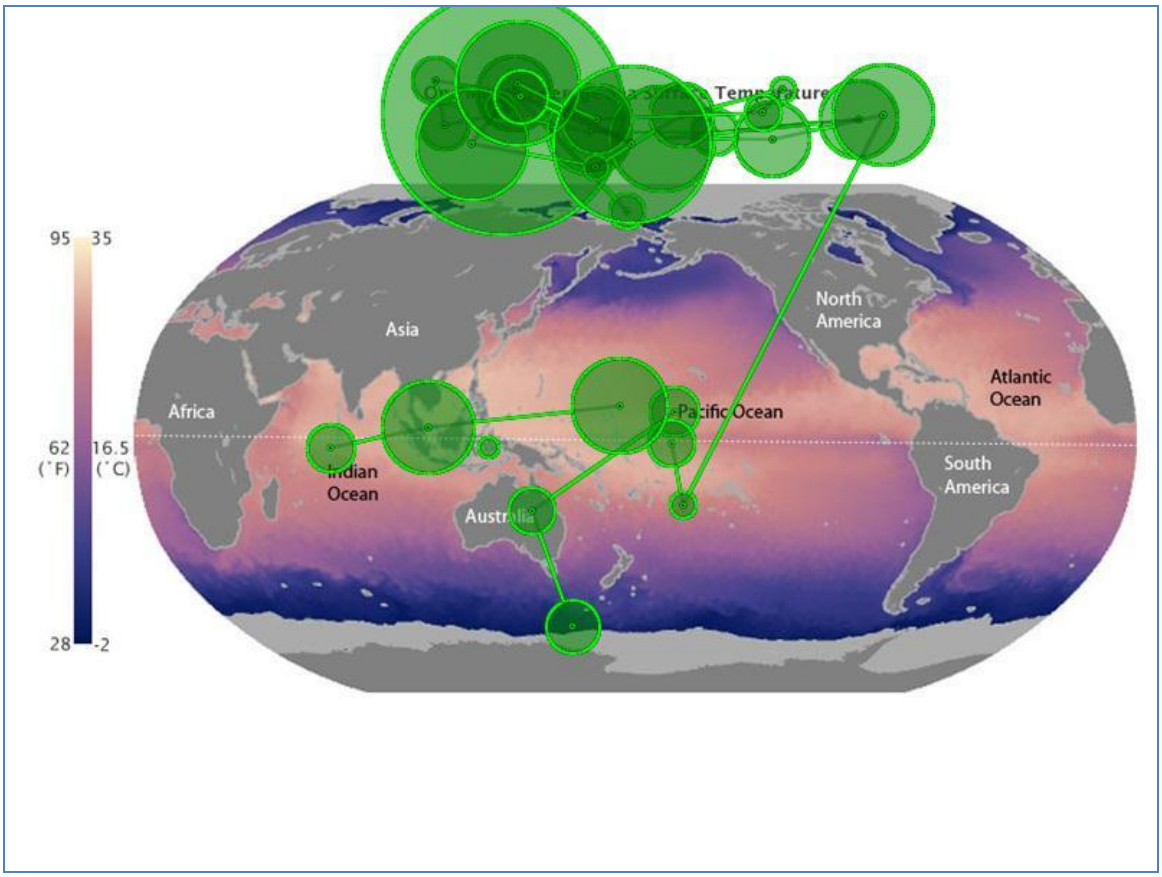


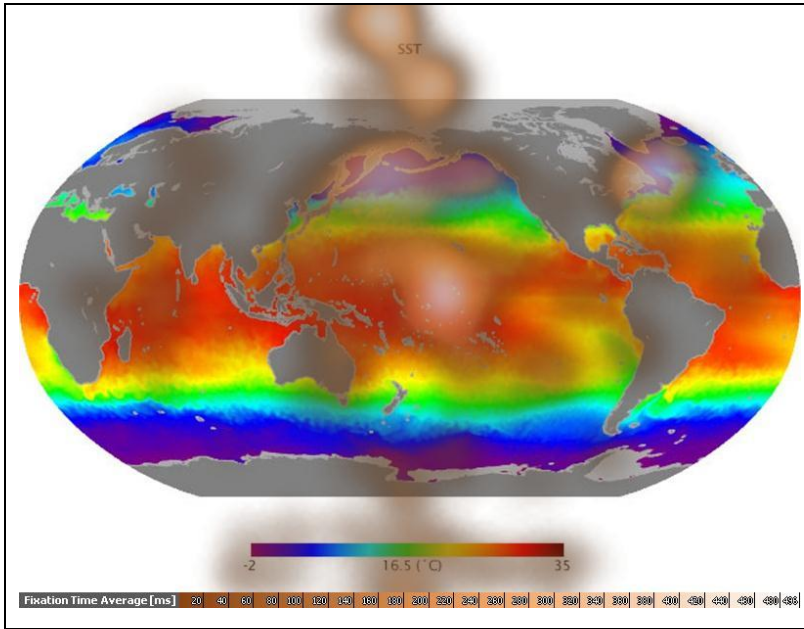
Figure 13. One expert's scan path, Fully-Scaffolded, SL. Note the location of fixations below the title but above the top of the map.

asked about the meaning of the grey in the visualization in both the clinical interviews and the interviews that accompanied eye-tracking, they may have been anticipating those questions, especially in later trials. Additionally, participants were asked in both experiments about the season of the year that was depicted, which also called on them to judge data coverage at the poles, and again these participants could have been anticipating those questions. Thus, it seems premature to assign those fixations to either the title or the map without further information to resolve those two possibilities. The increased statistical likelihood in the scaffolded cases could point to participants actually looking at the title, however, given its sheer length difference between the cases ("SST" versus "One Month Average Sea Surface Temperature" for example).

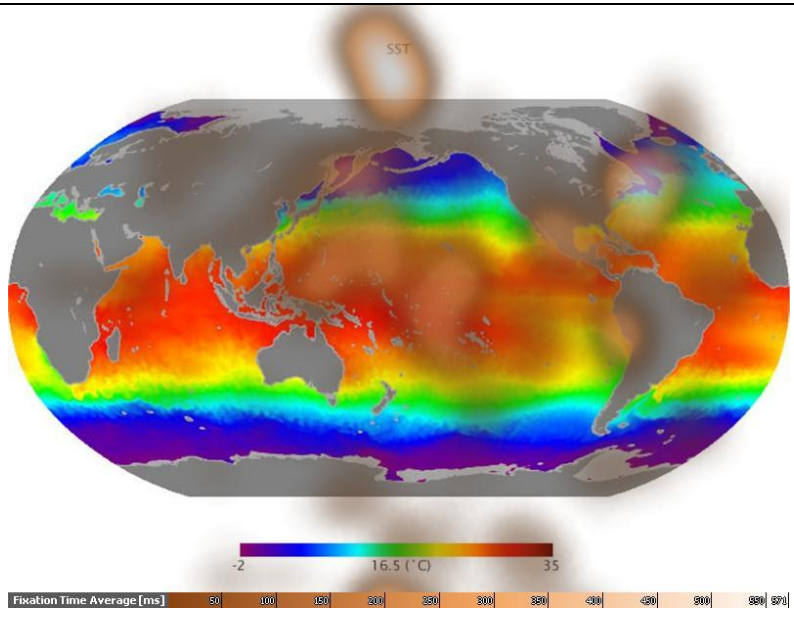
BeGaze™ also allows the production of heat maps, visualizations that highlight overall fixation duration summed across participants. See Figure 14. This visualization shows experts used the key more (more coverage), the Indian (more coverage) and Pacific (brighter white) Ocean basins more, and the title (dimmer white) slightly less than novices in the unscaffolded condition. From these visualizations, it is also apparent that the long fixations by users in both groups on the upper right part of the map could have been on the Gulf Stream pattern curling off of the northeast coast of the United States, which in the SST anomaly case (viewed by 10 of the 20 participants), had both cold and warm anomalies right next to each other and were very salient in the visualizations, based on the interview data.

The heat maps of novices in the unscaffolded versus geographic scaffolding case indicate they definitely use the geographic labels, based on the brighter spots around the ocean basin names and Asia and North America labels in the right-hand portion of the figure. See Figure 15. For the color-scaffolding case, the heat maps of novices did not differ very obviously from the unscaffolded visualizations. See Figure 16.

The title and measurement unit scaffolding heat map also revealed some subtle features of the novice differences from the unscaffolded case. See Figure 17. Namely, the right-hand part of the key was used quite a bit more than the left-hand side in the scaffolded case. In the scaffolded SST anomaly visualizations, the right-hand side contained the words “higher”, “average”, and “lower” in addition to the Fahrenheit temperature deviations (+9º F, 0º F, and -9º F respectively). In the chlorophyll case, the numerical values on the left-hand side were matched with the words “high”, “medium”, “low”, and “very low” on the right-hand side. Interviews confirmed that participants were using the units in particular to make meaning from the visualizations, especially in the scaffolded

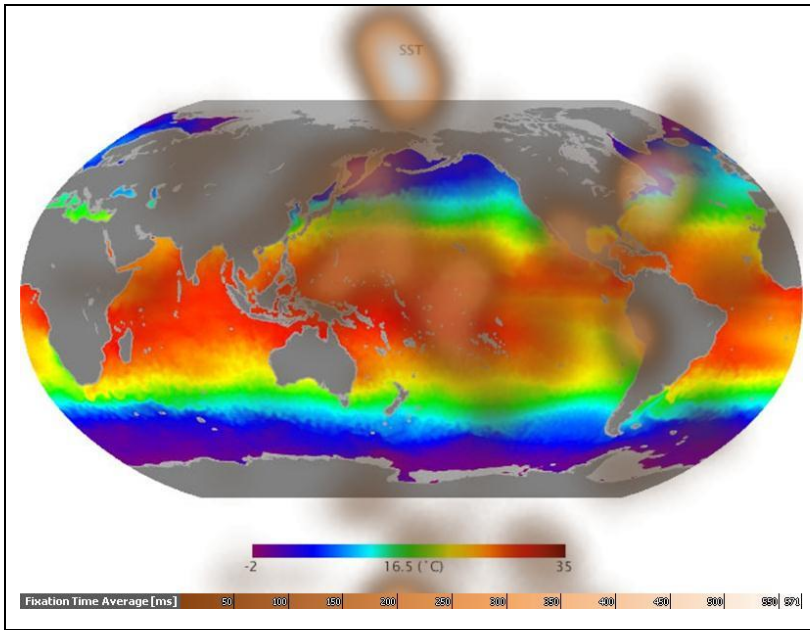


Experts

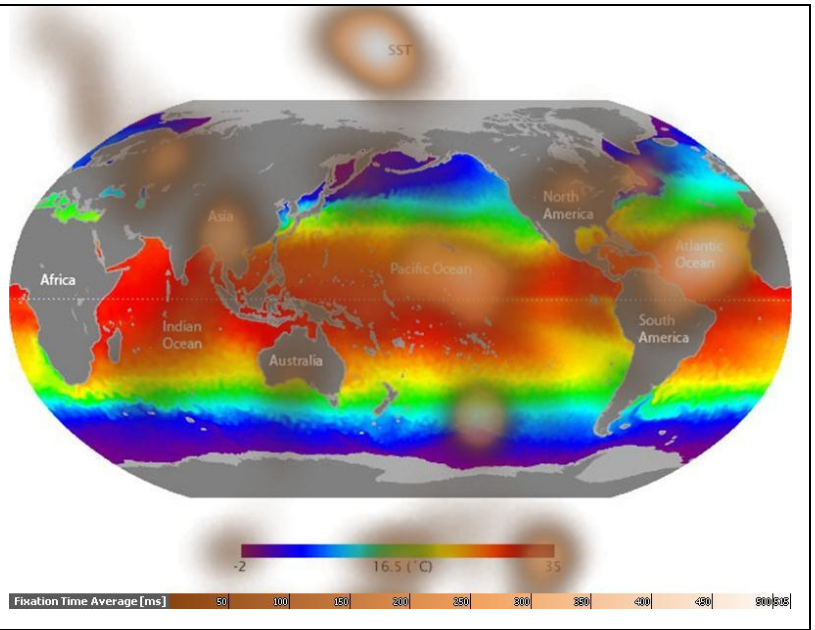


Novices

Figure 14. All experts' and novices' eye-tracking heat maps, Unscaffolded, SL. Dark brown is shortest average fixation time; white is longest average fixation time.

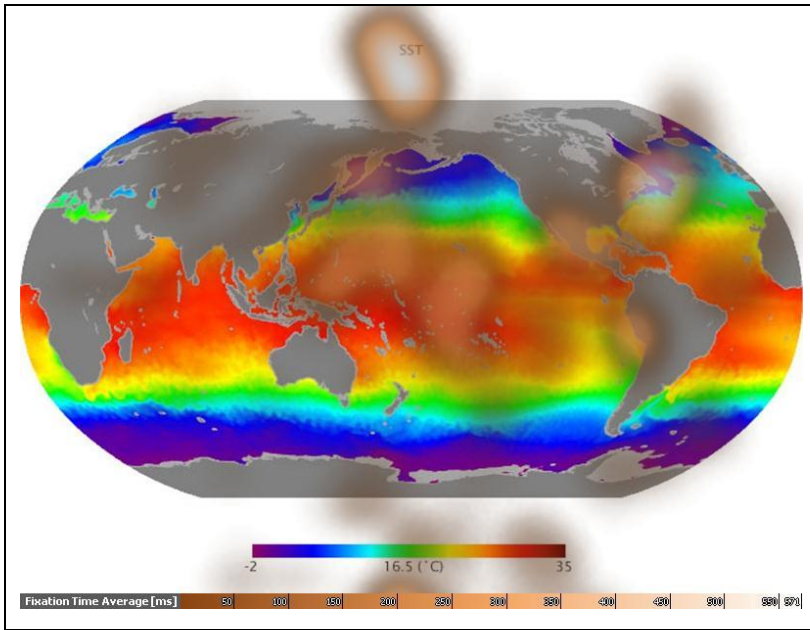


Unscaffolded

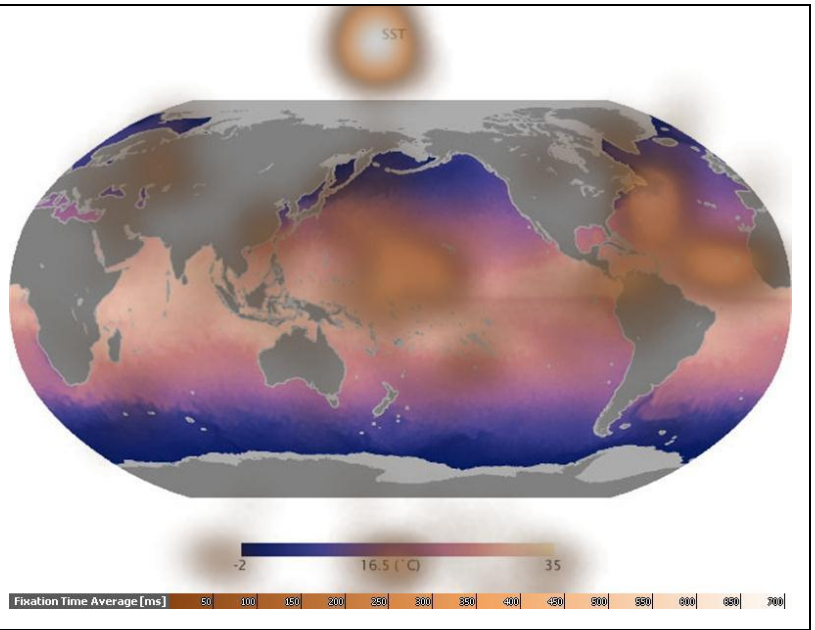


Geographic Scaffolding

Figure 15. All novices' eye-tracking heat maps, Unscaffolded versus Geographic Scaffolding, SL

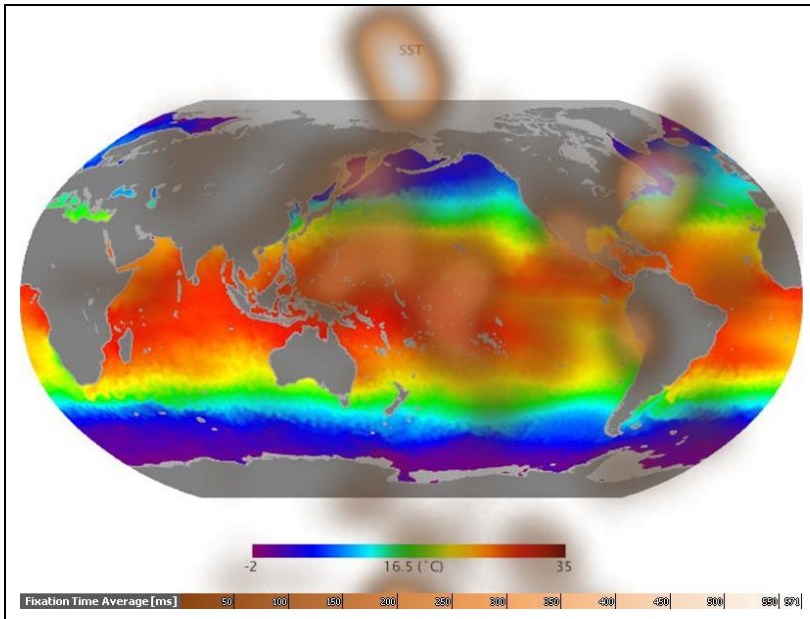


Unscaffolded

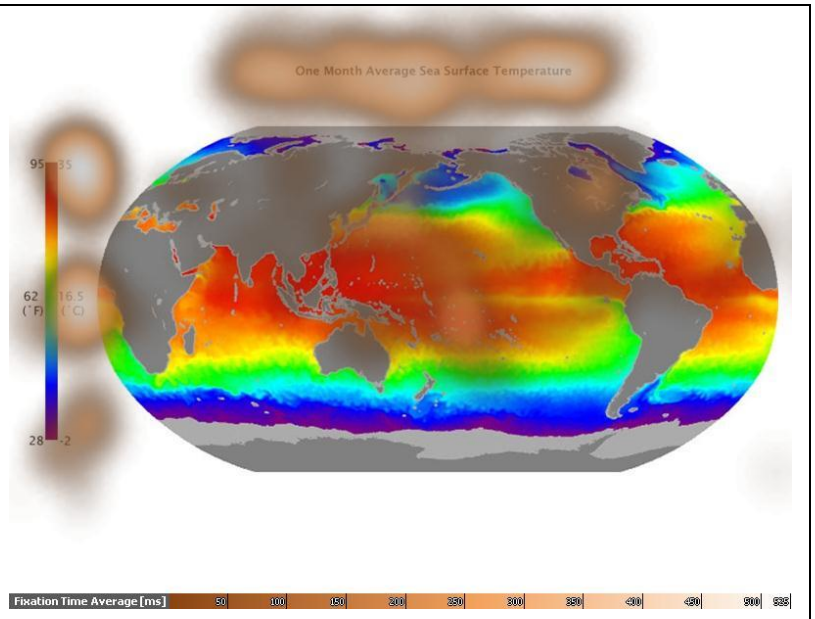


Color Scaffolding

Figure 16. All novices' eye-tracking heat maps, Unscaffolded versus Color Scaffolding, SL



Unscaffolded



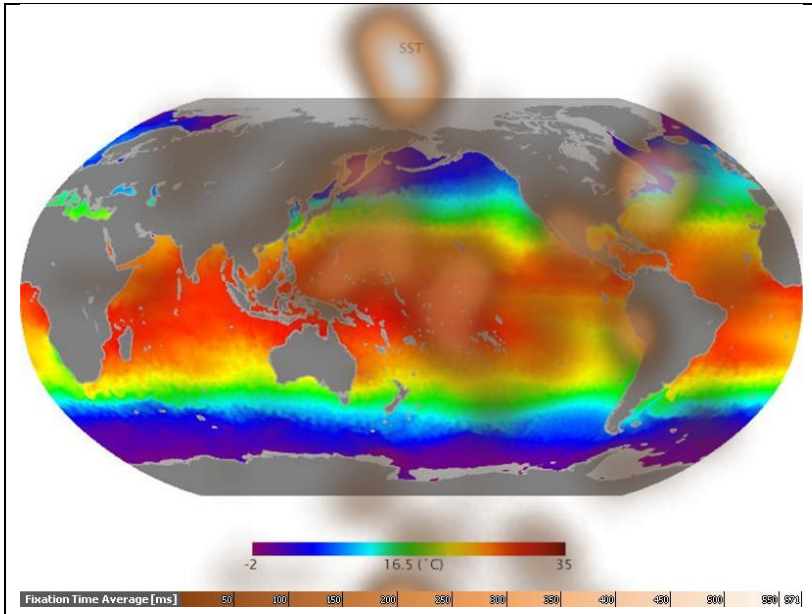
Title Scaffolding

Figure 17. All novices' eye-tracking heat maps, Unscaffolded versus Title Scaffolding, SL

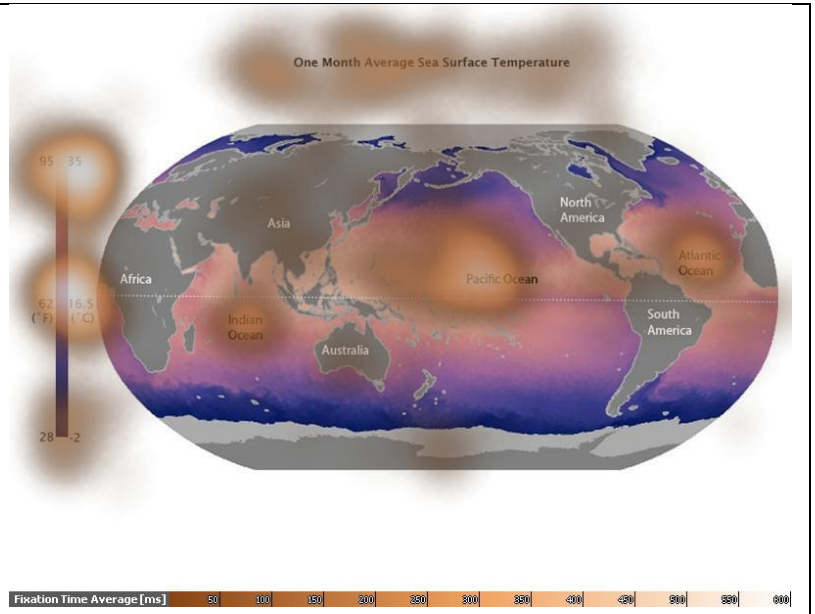
chlorophyll visualizations where they referenced the “high” or “low” values from the key.

Overall, comparing the heat maps of the fully-scaffolded case to the unscaffolded case reveals that the novices start to use the title more comprehensively but with less dwell time on any given part that would indicate confusion. They also use the key more, particularly the right side with the scaffolded versions, and generally do away with the Western Hemisphere bias in favor of other elements that are present to help them make meaning. See Figure 18.

Heat maps for the experts comparing the unscaffolded and fully-scaffolded cases were relatively similar, but offered some interesting comparisons with the interview data. See Figure 19. For the title, in the scaffolded case experts were somewhat biased toward looking at the left-hand portion, which for all topics indicated the one piece of information that was most difficult for them to judge: the time span represented. This emphasizes the need for explicit information for all audiences, regardless of expertise, in order for them to be able to spend less time guessing at the important but agreed-upon details and more time making deeper meaning from the patterns that are the new data under consideration. Being explicit in this manner could also, in fact, enable them to question the particular given details. One expert participant offered an anecdote at the end of the interview in which a peer-reviewed publication had a figure with unclear details about an anomaly from which the participant was unable to ascertain an important piece of information that was needed to properly assess the merit of the results and relevance to other work. Other experts noted the helpfulness of the “high” and “low” labels in the chlorophyll case; especially as those who didn’t work directly with chlorophyll didn’t immediately remember what the appropriate range for the actual numerical values was, unlike temperature, where 35 degrees Celsius and 95 Fahrenheit were not questioned as appropriate high temperatures at the equator.

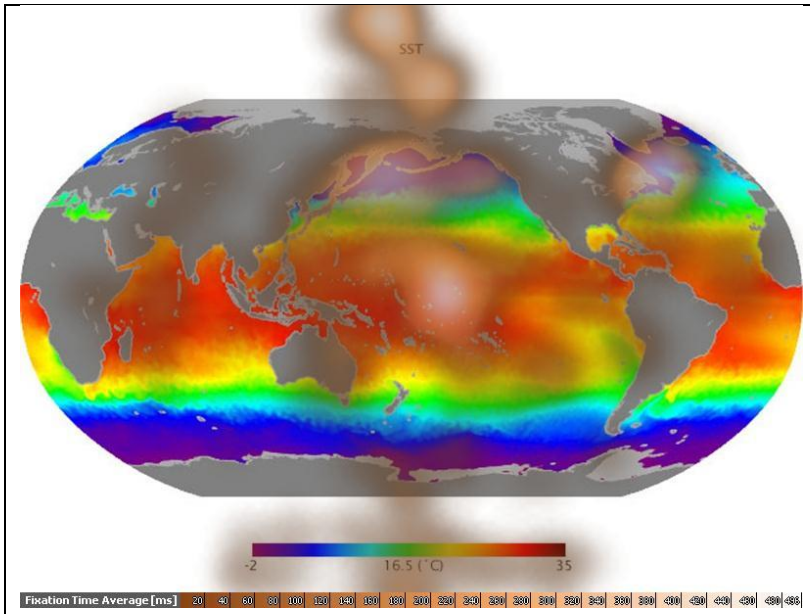


Unscaffolded

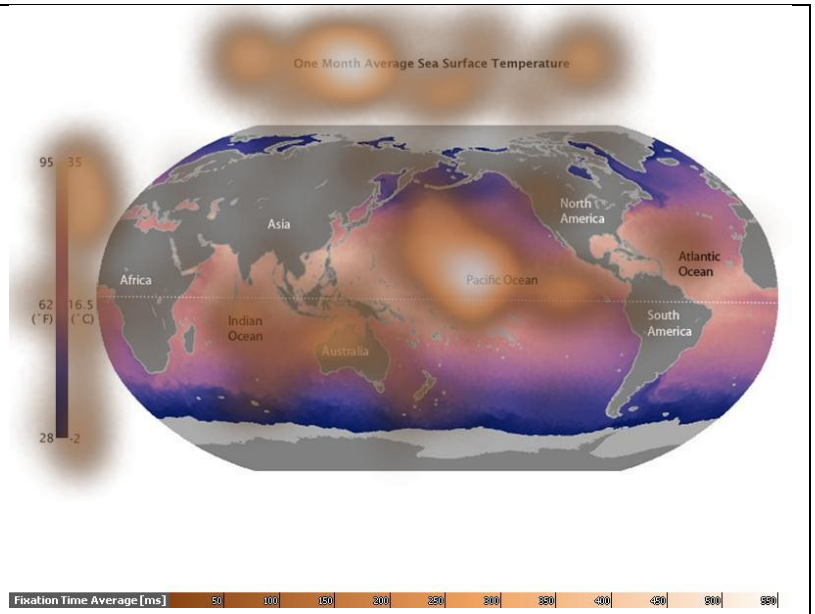


Fully-Scaffolded

Figure 18. All novices' eye-tracking heat maps, Unscaffolded versus Fully-Scaffolded, SL



Unscaffolded



Fully-Scaffolded

Figure 19. All experts' heat maps, Unscaffolded versus Fully-Scaffolded cases, SL

To consider the timing of use of the areas of interest, BeGaze™ facilitated production of sequence charts that showed when and for how long each participant fixated on any given AOI during the 10 seconds. Again, all visualizations regardless of trial number or topic were collapsed by scaffolding level. From these, it is apparent that in the unscaffolded case, several of the novices did not look at the title at all, at least directly, or looked only briefly. See Figure 20. Because the AOIs were drawn conservatively, looking at the

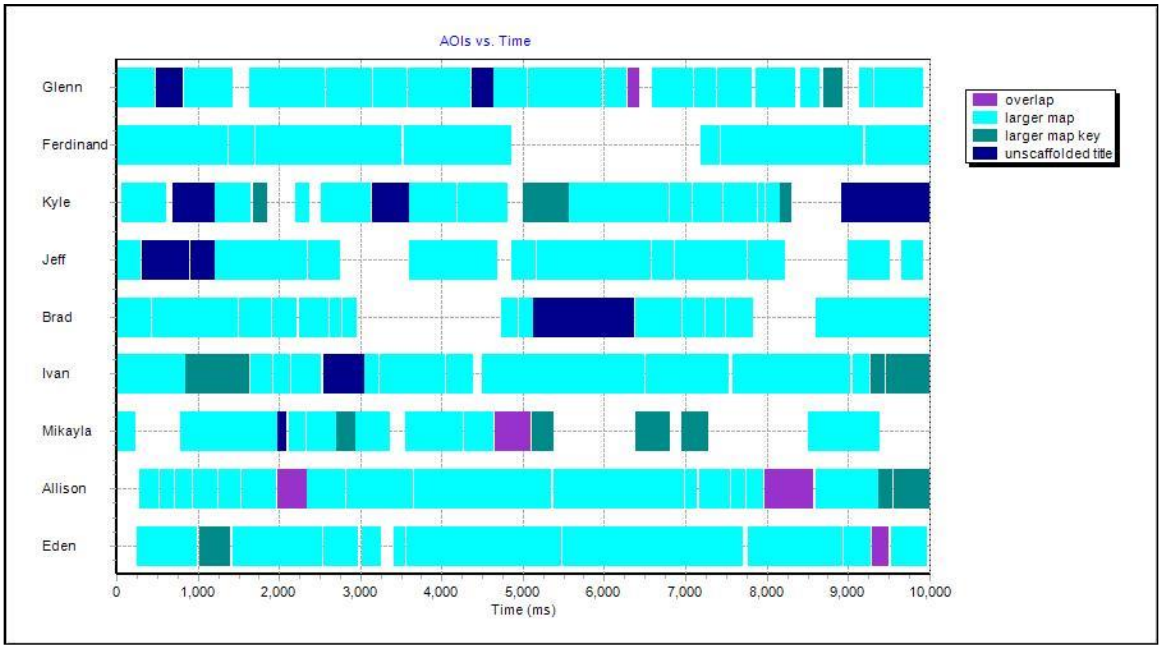


Figure 20. All novices, Areas of Interest versus time, Larger Map cases (US, CS, GS), SL

individual scan paths reveals that in the case of some participants, such as Mikayla (novice), the white space fixations are near the other visualization elements. One set in that case is below the key and is likely the white space fixations between five and seven seconds shown in the figure. However, some of the other fixations are between the title and the map, in which case it is impossible to know whether the participant is reading the title or looking at the top of the map, where there was no data in the oceans due to satellite coverage, a specific element of the visualizations that was asked about in both the interview

experiment and the interviews that accompanied the eye tracking. Overall, though, the novices seemed to spend very little time on the elements besides the map in the unscaffolded case. Also note the relative infrequency of looking at the AOI designated “overlap” in the novices (purple); this provides more evidence that that area of the map (namely, Africa) was not relied heavily on for meaning-making in the visualizations, and the spillover of the key onto that area in the smaller map case of the TS and FS scaffolding levels is not interfering with potential data use from that area of the world.

On the other hand, experts tend to look a little longer at the other elements, and particularly, six of nine looked at the title within the first two seconds. See Figure 21. Novices alter their viewing patterns in the fully-

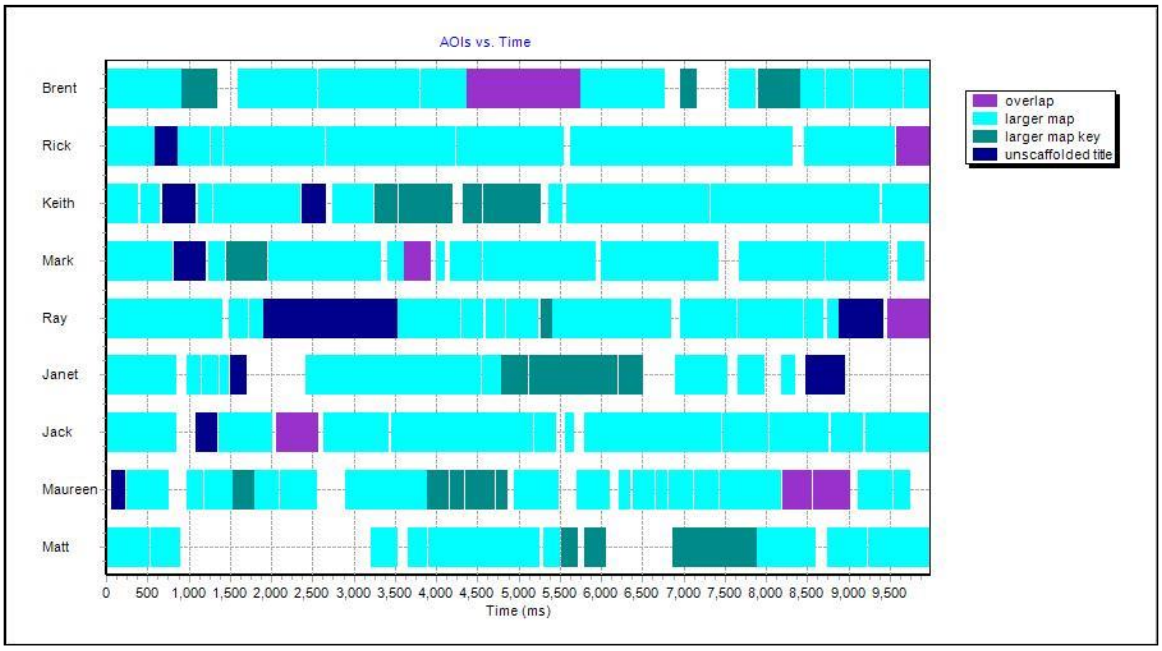


Figure 21. All experts, Areas of Interest versus time, Larger Map cases (US, CS, GS), SL

scaffolded case to rely more fully on the other supporting elements. See Figure 22. However, they seem to move to the key earlier on in the trial. Experts

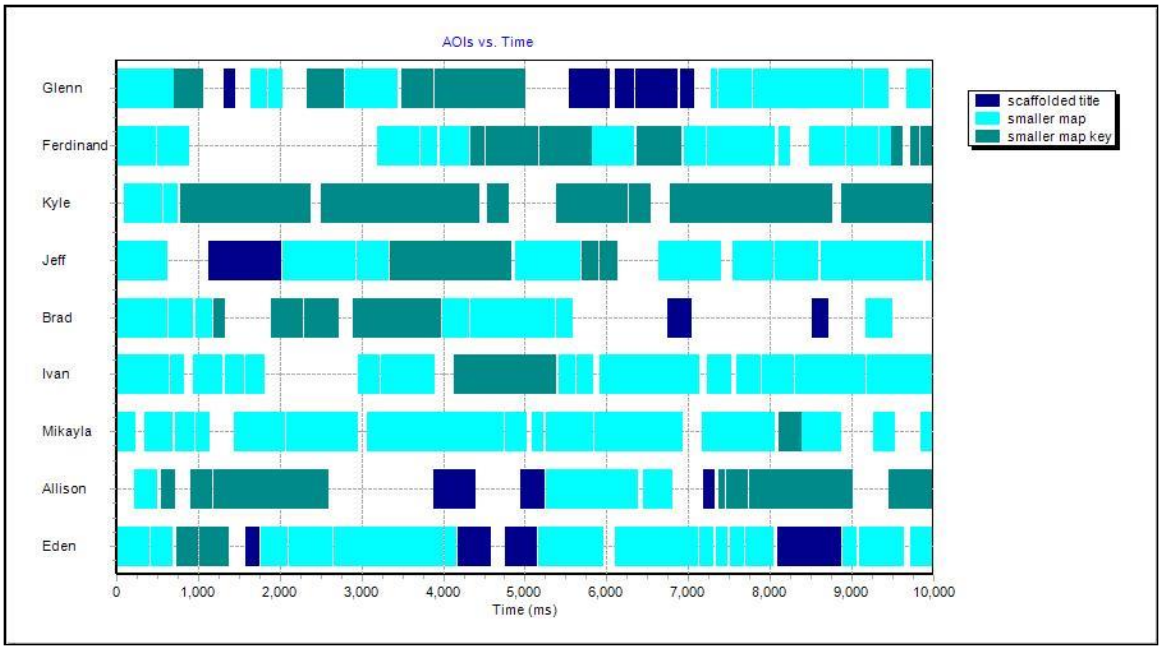


Figure 22. All novices, Areas of Interest versus time, Smaller Map cases (TS, FS), SL

continue to use the title first in the fully-scaffolded case, but also increase their use of the key. See Figure 23.

These two timing-and-use patterns of experts and novices in the fully-scaffolded conditions are in line with interview data which suggest both that novices relied heavily on the measurement units in their meaning-making (which falls into the key AOI), and that while they might have understood the words of the title better, or simply recognized it was there more frequently, it might not have meant much more to them than the abbreviations or jargon.

Overall, the qualitative eye-tracking results demonstrated how the scan paths of the expert and novice participants differed, and how the novices especially differed in their scan paths among the various types and levels of scaffolding. The experts generally looked at more of the visualizations than novices, as evidenced by their more widespread fixations. The experts especially were better able to make meaning using the titles and keys, evidenced by the

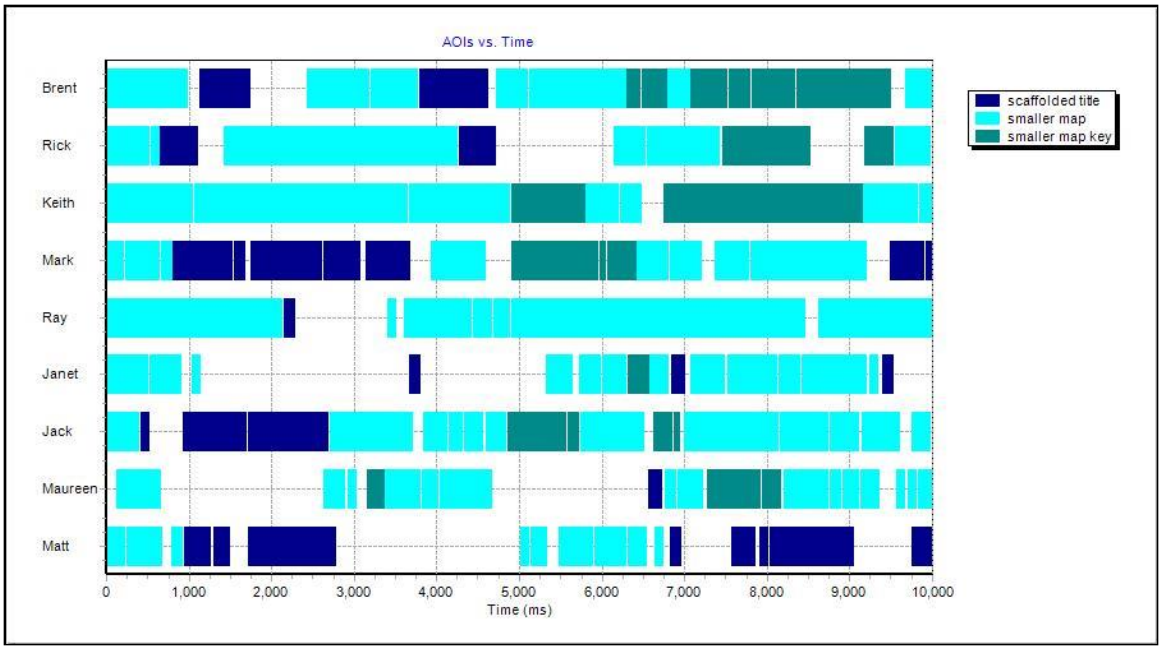


Figure 23. All experts, Areas of Interest versus time, Smaller Map cases (TS, FS), SL

shorter durations of fixations on these elements. Novices, meanwhile, were able to use the scaffolding in various ways to become more expert-like in their looking; they changed the amount of the visualization, especially the map portion, that they looked at as scaffolding was added, and their fixation durations on the title and key got shorter as they made more meaning. The differences between the two participant groups in the unscaffolded case was confirmed by the order in which the groups looked at various parts of the visualization: the novices and experts had different patterns in the unscaffolded case, with the experts tending to look at the supporting elements earlier and for longer overall, and novices showing changes that moved toward similarity to the expert patterns as scaffolding was added.

Summary

The quantitative and qualitative eye-tracking results overall supported and provided context to the interview results. The quantitative results confirmed

differences between the experts and novices in meaning-making by showing differences in fixation durations based on different scaffolding levels and different levels of experience with the visualizations (namely, the effect of trial in the spontaneous looking condition). Further, the differences between the variables affecting the spontaneous looking and main idea question conditions give insight into the influence of a guiding task on focusing attention for meaning-making. These differences suggest that without direction, there is no learned way of looking that the experts employ differently from the novices, but when asked to complete a specific, highly-familiar task, differences in attention based on experience are evident. Finally, the qualitative eye-tracking data reveal *how* those patterns of looking are different in terms of elements of the visualization used and in what order, as well as specifically confirming what elements remain sources of confusion, especially for novices, based on longer dwell times.

Discussion and Conclusion

Educators increasingly look to cognitive science research on learning to provide guidance on improving educational methods and tools based on how the human brain works (Bruer, 2006; Cocking et al., 2000). Direct triangulation of the methods of multiple disciplines gives data that points us to ways to not only design new tasks for experiments with eye-tracking, but also improve acquisition of the skills that could benefit communication. This dissertation used complementary methods to investigate specific elements of experts' meaning making to compare to novices' meaning-making in order to better marry the two in the creation of visualizations that are both more useful to experts and more accessible to novices.

The data presented here clearly reinforce the idea that expert and novice academic scientists use different meaning-making strategies when processing information in global visualizations of ocean data. Professional scientists and the general public differ in meaningful ways in terms of both the sociocultural expectations and the knowledge resources they bring to bear on interpreting spatially-based visualizations of scientific data. Experience using and creating the visualizations provides a further level of sophistication with the academic process that separates expert and novice scientists from the everyday scientists in the outreach audience.

Thus, the groups often arrive at different conclusions about the visualizations, meaning that these visualizations fail Lotman's (1988) first purpose of communication texts, that of sharing information. Sometimes, these visualizations fail to communicate amongst scientists themselves, especially between scientists of different disciplines or even sub-specialties of the same discipline. This dissertation examined the strategies used by expert scientists, novice scientists and outreach audiences in order to improve these visualizations for both shared purposes: transmission (Lotman's first function of texts), in this

case of academic scientific meaning, and new-meaning making (Lotman's second function of texts) by a multicultural group of everyday and academic science visualization readers.

In this discussion, I will first lay out the findings as related to the four research questions raised and discuss the limitations of the findings and their comparison to previous work. Following that, I lay out implications for development of visualizations for communicating among everyday and academic science audiences. I address the ways that educators, both informal and formal, can assist the enculturation of novice academic scientists into academic scientific culture, and how other professionals who may not be immediately seen as integral to this effort may also assist. Finally, I conclude with ideas for future work.

Research Question 1

What are the "bottlenecks" preventing novice meaning-making from data visualizations?

In the laboratory, novices struggled with almost every aspect of the unscaffolded visualizations. They averaged no better than partially correct on the majority of the questions related to the visualizations: main idea, evidence of the main idea in the data, meaning of no color areas, time span depicted, season depicted, conditions in the Equatorial Pacific, and normalcy of the conditions in the Equatorial Pacific. The two exceptions were measurement unit, which averaged completely correct, and color meaning, which was approaching complete correctness (0.81). However, certain participants did struggle on these two questions. They struggled particularly with the visualizations on chlorophyll and SST anomaly as compared to the average SST, which was arguably the most culturally familiar to them in their everyday lives. Novices were not reliably able to determine the difference between a logarithmic and linear scale bar and

understand the implications for the data representation. They also very rarely explicitly demonstrated they understood the difference between the anomaly and average SST visualizations, especially because they were unfamiliar with Celsius. Even if they identified zero degrees as freezing in the anomaly visualization, they did not question that five (Celsius) or nine (Fahrenheit) degrees below zero would mean that much of the ocean was frozen in the visualization depicted, suggesting they were likely also unfamiliar with normal ocean surface temperatures. This was also reflected in their confusion about what was considered normal or average in the visualizations, aside from concerns about the inconsistent use of those terms.

More than one could not immediately orient to the visualization, though they reliably pointed out the equator. They confused the land with the ocean, worked to understand the Pacific Ocean centered projection with the Americas on the right-hand side, or confused the geologic epoch with an earlier one where the continents were at least in different places with relation to one another, if not joined, or partially submerged as compared to modern-day continents. At least one admitted to being unable to identify the particular ocean basins before the geographic names were provided.

Novices also missed several supporting elements of the visualizations, especially in the first trial. In the cases where novices could report accurate titles and measurement units, essentially reading them off the stimulus, in some cases the interviews showed they could not appropriately make use of the information in them to make scientific meaning about the visualizations. Similarly, they could report patterns in the data, even recognizing similar patterns across trials, but not put them together with other elements to make broader meaning. They missed patterns such as the east-west difference along the equator in all three topics, likely due to scientific inexperience; they noticed a north-south difference which had meaning to them based on sunlight, but without ocean knowledge, they missed the other pattern that held significance for them (Kastens & Ishikawa,

2006). When they did make scientific meaning from the visualization, they could not extend the patterns and context to support multiple stories about the visualization, such as currents revealed by temperature distributions in the average SST visualization.

Novices were highly unfamiliar with the concept of what and how the visualizations portrayed the data. Some did not understand the visualizations were portraying averages over time. Many referenced either seeing these visualizations for the first time or being asked to consider these questions about the visualizations for the first time. When they critiqued the visualization elements as hard to use, many also blamed themselves for being unable to use them. Novices felt they lacked any information to make meaning on particular questions almost four times as often as experts. Of the few that recognized or guessed the data could be from satellite, none could explain how the satellite worked. On the other hand, several suggested that this data was collected manually by researchers or machines in the ocean.

They could not decode culturally-specific jargon in the form of words, abbreviations or colors. They only reliably decoded units which they had been using for many years in their lives or in recent academic scientific contexts such as milligrams per meter cubed as a concentration unit in chemistry. Novices could occasionally apply the idea of concentration or measurement of a substance to the chlorophyll visualizations; chlorophyll and investigations of substances in the ocean are in fact chemical oceanography, an overlap of the two disciplines in an applied context. They rarely encoded jargon, either, beyond words they were likely exposed to in biology or chemistry, but when asked further what photosynthesis was, for example, they could not accurately explain chlorophyll's role or the outcome of the process itself.

Novices reported missing the supporting elements of the visualization or struggling to decode them. This was evidenced in their interviews, and in the eye-tracking by the longer duration of their glances (in the main idea condition),

smaller number of fixations, and their less comprehensive patterns of looking at the overall visualization elements, especially within the map element itself. Evidence of their increased proficiency at the task with scaffolding also appeared in the interview scores, changing eye-tracking patterns, and shorter durations of fixations on the scaffolded visualizations; the effectiveness of scaffolding for low-content non-expert users may be attributable to reductions in cognitive load by externalizing of some of the task (Kalyuga, Ayres, Chandler, & Sweller, 2003). In particular, scaffolding here might lower load so novice users can spend cognitive resources not on orienting geographically or on decoding jargon but on viewing patterns and reconciling discrepant information to make meaning. Alternatively, the influence of visual salience (Parkhurst, Law, & Niebur, 2002) of distracting hues representing middle values in the unscaffolded, rainbow-colored visualizations may have given way to knowledge-driven control (Henderson et al., 2007) in the color-scaffolded visualizations. This would be indicative of increased meaning-making, and is evidenced in the shorter durations, though I did not directly investigate the salience distraction in the scan paths. Finally, the decrease in duration over the course of several trials is in line with my findings of guiding questions or simple experience with the visualizations improving comprehension.

Both the expert and novice groups struggled to match particular values to particular colors, consistent with work by Steffke and Libarkin (2012). Participants could match the extreme values with the color scale in the key when they were able to locate the key, although several reported only high or low, but not both, values when asked for the extremes. However, given the experts' reported tendencies to use visualizations more for seeing larger patterns than for judging exact values and the variability of climate, let alone weather, this may not be an issue. In particular, the broader communication aim of visualization use is not necessarily to teach specific values, either, but rather to convey patterns ("Data

Visualization | SAS," n.d.; Ware, 2008), so these are not intended to be tools for memorizing factual numerical knowledge.

While limited in scope, the *in situ* results show similar patterns. In the science center, visitors with limited formal scientific training had trouble discerning the visualization main idea, color meaning, time span, high value locations, and season. They seemed to improve with practice, indicating that a lack of familiarity with these types of visualizations and the types of questions I asked about the visualizations.

Research Question 2

What strategies do experts use to make meaning from the visualizations, and how do reported conscious strategies correlate with their perceptual and physiological strategies?

Experts can put all the pieces of data together that they gather from the visualizations and their background in ways that novices cannot or do not. The experts draw on more extensive specific factual knowledge as evidenced in the interviews. More importantly, though, cognitive research tells us they organize that knowledge differently and use it more efficiently (Bransford et al., 2000; Kastens & Ishikawa, 2006; G. A. Miller, 1956). They know what is important and how to use it through years of experience looking at these visualizations and having enculturated academic scientific routines of transforming patterns (Lave & Wenger, 1991) and accommodating discrepant information (Dunbar, 2000) on top of simply assimilating new information into their much more well-developed theories of the academic scientific world. In particular, scientists were confident about their thinking processes but often wrong about some details such as how exactly the seasons occur and occasionally likened the line of questioning to a candidacy exam, which also aims to test a graduate student's breadth of knowledge in the field. That is, they were comfortable with knowing what they did

not know, what was relevant or not, and what missing details from the visualization or their background knowledge that they could look up but that did not affect their overall reasoning.

Experts expressed a high level of familiarity throughout the experiment, and drew extensively on that specific background experience as well as the knowledge gained from that experience. Many have seen these types of visualizations, and some have seen visualizations representing the particular topics and content portrayed many times, through research and teaching that they do professionally. Most were aware the data was from satellite and overall they had better explanations for how the satellite worked or at least why the satellites were used, namely to collect sufficient amounts of data as to be depicted this way. A third of the participants even create visualizations based on satellite imagery. Moreover, the experts rarely expressed that the questions asked about the visualizations were novel, though it may have been some time since they had had to answer such questions.

The interview data and eye-tracking complemented each other in the investigation of this research question as well. The experts revealed their comfort and familiarity with the task verbally; their short fixations and sweeping comprehensive scanning patterns were consistent with comfort and familiarity with both the task and the information presented. Short fixations generally indicate better comprehension, though some specific exceptions indicate that expertise can lead to longer fixations where experts take more information in at a time with a larger visual span, as in the cases of goalkeepers and chess masters (Holmqvist et al., 2011). In my task, experts may need to consider finer detail in the small visual field than either goalkeepers or chess masters who are more concerned with taking in a big picture, especially when asked about details such as the season and location of the extreme values. Longer fixations by experts in the scaffolded visualizations in the spontaneous looking condition, more similar to novices, may also have resulted due to the influence of color scales and titles

that were unfamiliar enough to experts as to counteract somewhat their normal pattern of “taking it all in,” or to the lack of supporting information with which experts can confirm the conclusions they reach from absorbing the overall patterns. In addition, work with pilots indicates that longer fixations occur on elements that require information extraction rather than just information check, which may also increase the experts’ fixations in the direction of the novices in the spontaneous looking condition, especially on visualizations that have unfamiliar titles and color scales (Holmqvist et al., 2011). Cognitive load changes with the introduced scaffolding mean that eventually, scaffolding may make a task harder for experts, by increasing their load when the scaffolding provides extraneous, redundant (for them) guidance that has to be processed alongside their internalized knowledge and skills (Kalyuga et al., 2003).

On the matter of the interference of the psychophysics of perception with expert performance, it is likely that experts have learned to overcome or ignore perceptual salience that conflicts with data salience. For example, one expert noted, “Green stands out to me” in the anomaly visualizations, but recognized that that made the color scale simply annoying rather than causing her significant difficulty in meaning-making. This is supported by research suggesting that visual salience is less important than knowledge-driven control when “meaningful scenes are viewed during active viewing tasks” (Henderson et al., 2007, p. 541). This learned ability may be due to the reduced cognitive load experts experience overall in processing the visualizations. If experts’ processing is not held up by decoding other elements of the visualizations such as jargon, they may be able to overcome the cognitive demands of visual perceptual salience conflict, whereas it is one more thing that novices experience as overwhelming much cognitive demand. This is supported by the evidence of longer fixations by novices in the main idea question condition, when they are trying to make specific meaning.

Research Question 3

How can these strategies be compensated for and ultimately explicitly modeled for application to new visualizations?

These non-science-major novice scientists have started becoming apprentices on the way to achieving the status of academic scientists, using jargon, units or abbreviations, and similar strategies of visualization use to an extent. However, in their unscaffolded forms, these visualizations are not helping enculturate them further. Novice performance improved, at least on information retrieval, with scaffolding. More scaffolding is needed to support putting the information together completely. Since these stimuli stripped away the normal captioning that accompanies a figure in a report or web page, the surrounding explanation in a textbook, or the exhibit kiosk text or narration in a science center, those resources can add back in information about the season, time span, satellite source and interpolation process of the visualizations. In addition, the figures themselves could be annotated to highlight important features that either align with or conflict with normal visualizations. The addition of visualizations representing normal conditions alongside anomalous conditions of interest and the presentation of explicit reference to the typicality or unusualness of the visualizations presented could also be tested for effectiveness at compensating for the experts' extensive experience or education.

Scientists, for their part, who produce and speak about this data with their peers, could recognize that even their peers who are in slightly different areas of the discipline have likely forgotten some of what might be considered shared knowledge that doesn't have to be made explicit in the visualizations. Orienting themselves to remember when they were learners or how specialized and knowledgeable about particular sub-disciplines they are could help them be more sensitive to being explicit for audiences that need the most help. However, they do need to take care not to over-scaffold visualizations and actually interfere with

meaning-making in higher knowledge audiences (Cook et al., 2008). By being sensitive to the everyday scientific knowledge that non-academics draw on as well, scientists can begin to incorporate that knowledge into their visualization design. This would move toward a truly multicultural (academic and everyday) scientific-visualization-reader community, which validates both ways of knowing and corroborates them as meaningful (Fayden, 2005).

I can look to current methods of graduate training in oceanography for more suggestions on modeling for novices. Many of the experts reported using these visualizations in context in research, through collecting data and producing visualizations themselves. Learning in context has proven more effective than teaching abstracted knowledge in myriad forms: apprenticeship in the workplace (cf. Lave & Wenger, 1991), exploration in out-of-school learning environments (cf. Bell et al., 2009), enculturation in family and community groups (cf. Paradise & Rogoff, 2009), or cognitive apprenticeship and situated, contextualized learning in formal schools (Brown, Collins, & Duguid, 1989). Being able to teach about content is also an indication of at least burgeoning mastery (Shulman, 1986); peer-to-peer tutoring or work in small groups, methods encouraged in the learning of science writ large, could be investigated for efficacy here.

Exposure, modeling the routines of the expert, and practice are all supported by evidence in this study as means for enculturating novices into a visualization reading culture. Simple exposure to *meaningful*, scaffolded visualizations with interpretation by a more knowledgeable other could provide the first step, as novices reported familiarity with the use of these visualizations from TV or from school, where presumably a teacher or reporter explained the meaning to them. There is definitely a demand for such visualizations and a growing infrastructure of both technology and education professionals ready to support its use in attractive and perhaps more motivating ways (Haley Goldman et al., 2010). Modeling could be used to take the next step towards internalizing a socially-learned skill of interpreting visualizations. Carefully-scripted video

showing a scientist explaining what he or she sees in the deliberately-constructed visualization could illustrate the routines experts have to think about the patterns they expect (Dunbar, 2000). A video showing expert eye-tracking paths might demonstrate the visualization features most important to inspect (Jarodzka, Scheiter, Gerjets, & Gemballa, 2008). Finally, practice with interpretation, especially scaffolded in a group context or with a more-knowledgeable other, provides further exposure, further knowledge, and further modeling either by the user herself or the more-knowledgeable other. This, too, could be undertaken in multiple learning venues, from the formal to the informal.

To build effective interventions for broader audiences to enculturate them into the community of visualization readers, educators must start by considering where the non-scientists are, as addressed in the first research question. In this experiment, novices proved to have a fair amount of academic scientific information; they were familiar with temperatures and seasons, even though they couldn't always accurately reason about why the seasons occur. What they lacked was familiarity with the information in an oceanographic context, exactly the sort of information that becomes shared in an academic-only scientific culture and thus removed from visualizations as "given." Enculturated, assumed, shared practices must be made explicit. This can be done through information conveyed in the data itself, through color and other representational choices, through the supporting elements such as more markings on the scale both in numbers and in unlabeled but division-marking "tick marks" and descriptive titles and legends, and through supporting information presented in a kiosk, in the body of a textbook, or as a narration with an animation.

Novices are already demonstrating active sense-making through comparison to visualizations within and outside the experiment and the use of interview probes as guiding questions (Magnusson & Palincsar, 1995). Both of these lead them to reprocess information throughout the course of the experiment. While researchers interested in accurate task performance often

throw out that data on training their participants to perform a task as irrelevant, Vygotsky (1978) notes that it is precisely that evidence of microgenesis, learning as it happens, that is of interest. Drawing on that evidence in the data here, educators could offer "normal" visualizations for comparison to scaffolded use of normal patterns for those less familiar with the academic or oceanographic context of temperatures or chemical distributions, etc. In addition, they could design guiding questions, another proven method used by educators and interpreters (Magnusson & Palincsar, 1995) to focus the learners on particular parts of the task (Wood et al., 1976), such as reconciling discrepant information (Dunbar, 2000) or examining global rather than local patterns, as part of the scaffolding process.

Research Question 4

How can methods that are based in diverse epistemologies be triangulated to provide insight into these research questions and larger problems?

By using a real-world task, triangulation among clinical interviews, eye-tracking with interviews, and *in situ* eye-tracking and interviews allowed me to not only verify findings from one method with another, but also gain detailed insight on the findings by fitting the pieces together. Interviews alone revealed the combination of different types of knowledge and different types of visualization-reading skills between the scientific expert and novice groups, rather than an overarching influence of one or the other. However, eye-tracking allowed me to more subtly probe the visualization-reading skills in a way that participants could not or did not voice. Namely, I found that novices and experts did not differ statistically when they were just asked to look at the visualization as a whole, but differed in their looking patterns within the map part of the visualization and had statistically different fixation durations when faced with the specific meaning-making task of deciphering the visualization main idea. Data from each

experiment filled in gaps in the findings revealed by the others, and they were directly comparable due to the use of the same visualizations throughout.

Eye-tracking assumes that the participant's overt attention, as measurable by the recording device, is an indicator of his covert attention as well. However, the field acknowledges that this may not always be the case; astronomers in particular are known for detecting dim objects by directing their attention to areas outside the foveal point on which their gaze is fixed (Duchowski, 2007). Thus, the ability to record not only gaze position with the eye-tracker, but also conscious attention through the interviews allows us more insight into whether and how participants used particular visualization elements by combining these techniques.

While technical problems prevented full use of the *in situ* eye-tracking, the interviews there were in line with the findings in the laboratory, namely that academic science novices are not yet facile at making academic scientific meaning from these types of visualizations but that they could improve with some of the interventions suggested above. The *in situ* results also confirm the need to test visualizations with real populations and real exhibit situations, however, as the science center visitors performed worse than the laboratory population, suggesting undergraduates may be more enculturated into academic science than the general public. This could be due to the increased amount of information offered particularly undergraduate classes today versus when an older population studied science in school (Raloff, 2010). Many of the laboratory participants were also either currently studying or had very recently studied related science disciplines, such as chemistry, that could have contributed to that disciplinary knowledge being at the forefront of their minds, in a way that was less likely for the *in situ* participants.

Moreover, it points to the need to actually design visualizations with the real exhibit situations foremost in mind, and perhaps reverse the experimental process. It would behoove us to start with development of the *in situ* situation and

design visualizations around the limitations of the exhibit and test what balance of visualization principles and lighting and other real-world constraints are optimal. Afterwards, I can then step back to the laboratory to understand the deeper meaning-making processes, rather than starting with an ideal visualization and working toward the exhibit.

Limitations

Stimulus visualizations. Despite my efforts to make the most-broadly meaningful visualizations possible, even expert participants had trouble interpreting some versions. This means my visualizations were still not as ideal as they could be. Both expert and novice participants expressed difficulty picking exact values in the SST and chlorophyll cases where the full range of 0-256 brightness, hue, and saturation values was used. At the ends of the spectrum, the values were too low (too little brightness) or too high (too much brightness) to be distinguishable. The range of hues also seemed too broad (too many hues in too small of a range) to provide detail. This may be akin to problems users had in judging values from colors in visualizations found by Steffke et al. (Steffke & Libarkin, 2012). Perhaps blocks of color to represent ranges of values instead of absolute continuity through the spectrum are warranted. However, since it may be that the patterns, rather than the absolute values, are most important especially for a majority of outreach and communication purposes, continuity may be preferable to being able to determine absolute values. This is a subject for future research, alongside more in-depth analysis of how professional scientists use imagery in their research work, not just their communication of results.

In addition, experts expressed some reservations about the titles, colors, and data shown in the visualizations, particularly of the chlorophyll or ocean productivity set. Some experts had reservations about what data was not masked

in the chlorophyll cases that gave false high values of chlorophyll when they were actually sediment or reflection from ice. I was also specifically testing whether the blue to yellow-white scale would provide meaning that would allow both global land and ocean chlorophyll data to be depicted on the same visualization. When showing the ocean chlorophyll alone as I was doing, this color scale did not seem to provide much cultural context. However, two hues of green scaled through white may be sufficient to contrast ocean and land chlorophyll on the same visualization, but this will require additional testing. Finally, the choices of the words “productivity” or “microscopic ocean plants” for the scaffolded version of the title were remarked upon as inaccurate by several expert participants. The choice of the word “average” in the titles and “normal” and “typical” in the interview questions were also not clearly defined to many participants, and were often conflated. Therefore, there may need to be a tradeoff between what is perfectly scientifically accurate and what conveys the most meaning to general audiences, especially given other constraints of time, money, and space in communication products. These results suggest that there will always be a role for prototyping imagery both in its development stage outside of final context as well as in its proposed final context. This is also the case as I did not, of course, test the entire range of oceanographic spatially-based visualizations here, let alone the entire range of spatially-based visualizations of data for Earth science or science writ large.

Because I wanted to compare across topics and scaffolding levels, there might have been more factual carryover that impacted score accuracy from one visualization to another than if they were different seasons or time spans. This could be the cause for decreased duration in the eye-tracking spontaneous looking condition with increasing number of trials; however, I saw no trial effect in the main idea question condition, suggesting less factual carryover and truer microgenetic development. However, there was more reported recognition in the clinical interviews than in the eye-tracking, suggesting that either the variation in

seasons indeed prevented strict answer parroting from earlier or the shallower probing did not encourage participants to reveal such extensive recognition.

These visualizations were originally developed to constraints for use in a functional magnetic resonance imaging experiment, and may not have been optimal for viewing on a larger computer screen due to lower resolution “blown up” to the larger monitor. This may have especially influenced the legibility of the font in the titles and measurement units. Again, prototyping with users and pilot testing for experimental use may minimize these problems. Pilot users here did not report problems with legibility, though some of the participants did. This could be an age effect as pilot participants were graduate students and generally younger on average than the study population of experts in particular.

Study sample. Although I attempted to contact a random sample of oceanography researchers from the university, researchers may have self-selected to participate if they were interested in or work with visualizations of data, particularly given the 1.5 hour time commitment to participate in the interview. The summer timing of the experiments might also have conflicted with those oceanographers doing fieldwork. This could limit the ability to generalize broadly to all oceanographers with Ph.D.’s, rather than those who are more intensely involved with especially global satellite visualization use. However, while the overall participant population did report generally frequent to very frequent use of visualizations, interviews revealed fewer than half actually use global data, or even satellite data in the visualizations they work with. Further, even the expert participants who expressed little familiarity with the specific subject matter in the visualizations or with satellite data had greater ability to make meaning than the majority of the novice population. I also cannot speculate how these results might apply to oceanographers at other institutions where there may be a more- or less-heavy concentration of researchers using spatially-based visualizations or satellite data.

The study was also by-and-large restricted to a participant population of undergrads in a presumably narrow age range at a public four-year research university in the Pacific Northwest of the United States. Compared to the total public, four-year research university demographics, the overall university population from which my sample was drawn is slightly less likely to be female (48% versus 51%), less diverse (15% United States minorities versus 27%), roughly the same age (87% under 25 versus 84% under 23), and more likely to attend full-time (85% versus 70%). As a land-grant institution, there is a heavy emphasis on science, and even though these participants were not majoring in science, they had all completed three years of high school science at a minimum.

There also may be an effect of age; the laboratory experiment novices all appeared to be of relatively traditional age (under 25) for undergraduates with the exception of the one novice who did not attend college. However, if age were a contributor to expertise by simple fact of length of life experience, I would expect better performance in the *in situ* population, who presented as older than the novices, by and large. This study found the opposite, albeit likely significantly confounded by exhibit lighting issues. Whether or not the scaffolded visualizations are a sufficient starting point for transmission of even some academic scientific knowledge for those persons, or as a starting point for enculturation into a group of visualization readers, however, remains an open question.

Methodology. The clinical interviews were semi-structured. This means participants were not all asked exactly the same questions or shown all the visualization versions, for several reasons. First, the experts often talked at great length about their reasoning and the information they could or could not get from the visualizations, meaning there was not enough time to show the full ten visualizations in the allotted time. In addition, some participants from both groups recognized the data patterns as the same from one visualization to the next, so it

was deemed unnecessarily repetitive to either show them multiple further versions if their answers did not change or to continue asking them questions they had answered accurately in previous versions. For this reason, the number of scores per visualization, per topic, and per scaffolding level varied among participants. The accuracy scores were computed as participant averages and then those averages were averaged together to minimize the effects of the variable number of scores, either of fewer or more high or low scores based on the particular questions asked.

In the eye-tracking, titles were longer in the scaffolded case, which made them naturally a larger AOI to hit, which may have affected some of the relative likelihoods reported in Chapter 5. However, with the longer duration of fixations for novices in the unscaffolded cases reflecting longer processing time, one would actually expect a higher relative likelihood of novices hitting the title AOI in the unscaffolded case, which I did not see. In addition, the increased relative likelihood of looking at the title shown in the scaffolded versions here mirrored the performance in the interviews, both clinical and accompanying eye-tracking, namely that the titles were better understood in the scaffolded cases, suggesting more use due to better comprehension rather than simply more area at which to look. AOIs for the key actually encompass both the color scale and the measurement units, both of which interviews revealed that users were using, somewhat differently to make meaning. They were also scaffolded separately, as the colors changed in the CS case and the units in the TS case. In the future, I will consider separate AOIs for these.

The map portions of the visualizations were also smaller in the FS and TS scaffolded cases, providing a smaller AOI to hit and also more white space overall in the visualization. This may account for the increased relative likelihood of participants to look at white space in these cases versus unscaffolded cases, though I also saw an effect of trial. Thus, the change could be attributable to microgenetic development as participants learned the task during the

experiment. Different stimuli designed to address these issues will be needed to tease apart the causes.

Comparison to previous work. The eye-tracking pattern findings here are somewhat at odds with other expert-novice research on global visualizations (Libarkin et al., n.d.). In that study, experts were observed to use the visualizations more systematically and novices more randomly. However, their study used visualizations with data on both land and the ocean, and gaze patterns reported focused on the land, almost to the exclusion of the ocean in both cases. While they do not report the specific disciplinary expertise of their participants, nor the questions asked, if those experts were geographers or geologists predominantly as opposed to oceanographers, this could account for the discrepancy. Alternatively, the lack of accompanying information besides a color key in Libarkin, et al.(n.d.), may have led to more orienting behavior based on the familiar continental forms, in line with the novice findings in the unscaffolded case especially. The single expert and novice pathways depicted by Libarkin, et al., do align with the relative use of the key demonstrated by the experts and novices in this study.

On the other hand, Grant and Spivey (2003) found that highlighting a critical area of a diagram increased the number of participants who correctly inferred the presence of a tumor in a radiograph. More importantly, they found a difference in eye-tracking patterns among those who solved the problem correctly and those who did not. While Grant and Spivey did not run the experiment, it stands to reason that eye-tracking patterns of unsuccessful unaided problem solvers would shift in the aided condition to more closely resemble the patterns of the successful unaided problem solvers. This is consistent with findings in this dissertation of shifts in novice eye-tracking patterns from scaffolded to unscaffolded versions. Further, guiding problem solvers' eye movements before giving them a problem substantially increases the

chances of the participant successfully solving the problem, ostensibly by modeling the eye movements of someone close to solving a problem (Thomas & Lleras, 2007).

Our inexperience with the eye-tracking system and the discrepancies of some of the findings versus others in the fields, particularly as regards independence of the fixations (Hooge et al., 2007; Tatler & Vincent, 2008) leaves room for further work on increasing precision and accuracy of these methods on real-world tasks. Despite its hundred-year existence, eye-tracking remains a relatively young science, accompanied by many of the same issues that plague other high technology neuroscience based methods, especially the disconnect between the expertise of the engineers who build the devices and the scientists who wish to use them (Duchowski, 2007). Use of eye-tracking as a more neuroscience-based test to weigh in on real-world learning tasks is probably still not mature enough to fill all holes in methodology (Masson, Potvin, Riopel, Brault Foisy, & Lafortune, 2012), but by starting to use it I can start to realize those holes, first steps to actually using it. Particularly in science learning, fMRI represents a highly underutilized partnership of cognitive science and education research (Masson et al., 2012). The recent pronouncement by the U.S. President Barack Obama of his Human Brain Mapping initiative (Memmott, 2013) will undoubtedly spur the research across the board in this mapping of all stripes. Balancing the use of the technology in the more controlled and well-understood laboratory setting and the real-world situation of the formal or informal learning setting will also be a challenge to pursue in the future.

Generalizability

Data for this study are by-and-large very selective, especially on sub-questions such as differences among visualization main ideas. However, they have far-reaching implications.

Implications about participant populations. These results suggest there are at least three groups of scientists here: the academic scientists, typified by the expert participants in this study, the everyday scientists without formal science training, exemplified by the science center visitors, and a group in between, the laboratory novice group, composed mostly of undergraduates pursuing majors outside of academic science. The limited sample size, especially in the *in situ* experiment means I must be cautious about extending these findings to whole populations. However, given that educators are trying to reach all levels of academic science learners, I can take these results as evidence of *at least* some of the population struggling with unscaffolded visualizations of this ilk. The least-represented population here is those who lack even a high school degree; I speculate that those persons may struggle further to make meaning from the non-scaffolded visualizations, certainly, than the undergraduate participants if not the science center visitors here, given a presumed even smaller body of academic scientific knowledge and cultural experience on which to draw. At the same time, there may be an intermediate level user such as the current undergraduate population to whom educators might offer slightly less or different scaffolding, presuming that population to be more able to solve parts of a visualization meaning-making task independently.

Implications for visualizations. This research also revealed specific issues of scaffolding that I had not previously considered. All colors, including areas represented as no data, need to be labeled, and explained further if necessary; novice participants especially but also some experts were not clear on the implications of two shades of grey, neither of which were labeled in the stimuli used here.

Word choice for scaffolded titles, guiding questions, and additional information is important and should be tested with audiences for clarity. Here, the use of “average” was unclear, especially when coupled with what participants

used as guiding questions containing “normal,” “typical,” “usual or unusual,” or “expected.” Scale bars need tic marks and, to the extent possible without introducing visual clutter, should include values that make any mental math in dividing up the color scale easier for users.

For all audiences, visualizations need to include data on the methods of data collection. Specifically, they should include: a) the data source, such as satellite or ship-board, b) the time span over which data was collected and averaged, and c) the particular time period depicted, such as “Average of All Decembers, 1997-2007,” or “Averages of all data points, 1979-2010,” with further definition provided in a caption or other supporting text. Additionally, the amount and type of interpolation or smoothing done to the visualizations should be mentioned in visualizations for higher-level audiences.

Color choice may provide meaning based on cultural associations, but the colors I used here did not apparently do so. At the same time, those cultural associations of colors may interfere with meaning-making if they do not match the data depicted. Critiques offered by users indicated that the single color gradation or spiraling through hue, saturation, and brightness equally offered them less opportunity to distinguish particular values, especially at the low and high ends of the scale. Depending on the purpose of the visualization, perhaps a compression of the color scale range or use of successive blocks of color instead of a completely continuous scale is warranted. However, introduction of arbitrary divisions of color may introduce artificial divisions in the data patterns. Using fewer hues could assist in offering information as text; color “words” are not always similarly perceived by different users, so text that refers to “yellow-green” versus “green-yellow” can be ambiguous; for true clarity, text should have color swatches to match the visualization when color names need to be provided. Fewer hues also facilitate translation of the visualizations to black-and-white for users without access to color printers or color blind users.

Areas of no data must also be colored with care; users here reported difficulty distinguishing the light gray of no data and the white of high values in the chlorophyll visualizations. This perceptual illusion is due to principles of gestalt perception that have evolved to help us process the visual world efficiently, but that sometimes backfire and cause illusory perceptions that actually interfere with meaning making (Koffka, K., 1935).

More information in a visualization is not a bad thing, especially if it is in different modes, which complement each other and "[divide] the semiotic labor" (Kress, 2010, p. 1). Especially for more novice users, redundant information could point to ways to make meaning among the different modes (Cook et al., 2008). Use elements in a coordinated, deliberate fashion. The data representation itself, the supporting title, measurement units, key, including tick marks and value labels, and if possible, accompanying text can provide the information multiple times and in multiple modes so that there are more chances for novice users to quickly coordinate their everyday understanding with that of the academic scientist and move on to the process of meaning-making. Indicating features of interest or referred to in the text with arrows, circles, or other annotations in the data representation can draw attention visually as well as textually and even auditorially in particular settings where the visualizations are accompanied by text that is narrated instead of or in addition to written. This will also help to overcome textual literacy barriers for either those non-fluent in reading or those not fluent in the particular national language used in the visualization.

There is no substitute for trying out things with real people and in real situations, even if one also follows all best practices for design; one expert was confused by the physical proximity of the "5" and "9" on opposite sides of the key when it was presented vertically and offered the values in both Celsius ("5") and Fahrenheit ("9"). This likely could be easily clarified by moving the values or increasing the font, but it would be an expensive textbook reprint if it proved to

confuse many users in a class just for want of thorough testing. Lighting issues in the exhibit for the *in situ* case were exacerbated by the design of visualizations for a well-lit laboratory setting. Particularly for spherical digital displays that motivated this research, users and the exhibit may ironically limit the "global" nature of the data: the data gap at poles will not readily be seen when the visualizations are presented in the traditional pole-up (or even tilted-pole) orientation, and the potential affordance of the full 360-degree experience may not be realized if users do not walk around more than one side of the exhibit.

Earth science data may have some advantages over data from other domains when it comes to visualization based on the general population's familiarity with the globe, at least in the United States. The recommendations above all hold true for visualizations also for data from other scientific or even non-scientific domains. For data which may be even more abstract and with fewer existing characteristic connections than Earth science data, those recommendations could be generalized. First, background given information and foreground the information the visualizer is trying to convey by offering clarity and even redundancy so a viewer of a visualization can orient easily. For example, adding geographic labels to the image allows the user to quickly confirm what users likely suspect to be familiar shapes as continents, rather than taking extra time to wonder about something familiar that is presented in a novel context. In a more abstract visualization of a protein structure, the visualizer might wish to describe a particular portion of the overall structure that interacts with other molecules, but show it in context of the protein backbone. Here, the visualizer could simplify parts of the structure that are not the interaction site, by labeling the backbone as such so that viewers do not spend time wondering if that part is indeed the backbone, or if their prior understanding is different.

Color in an Earth science data visualization conveys differences in the data being depicted. The color not only draws the human eye to focus attention on important parts of the visualization, but also provides detail in the form of

shading and hue. While Earth science data in at least some cases may have a cultural association with color, such as temperature as red and blue, or plants as green, this association may not exist in other domains. Yet color in a protein visualization, could draw the eye to the interaction site and provide detail of interest by representing different elements, atoms, or electronegativity with different shades or hues, while the backbone is greyed out and backgrounded. As with Earth science data, different hues are best used when different categories are represented, such as different elements. On the other hand, to depict differences in values in one category such as electronegativity, employ different shading of the same hue, or at least a smooth transition equally among hue, saturation, and brightness as employed in the sea surface temperature average visualizations in this study.

The pieces of supporting information must also be clear and complete. This may be easier to accomplish than in the case of a domain where the audience might be assumed to have a level of familiarity with the subject; if the visualization designer has no assumptions about what the audience knows, it is more likely the designer will test titles, labels, and measurement units or key information for clarity with an actual sample of the audience. What words are jargon in labels and where to start to find acceptable substitutes in the case of data from domains other than Earth science may be less apparent; “backbone” of a protein structure may still not be clear to a completely novice audience. That means that testing with a small population of the target audience is even more essential when the topic is more abstract than concrete, more obscure than familiar. Drawing attention to particular parts of the visualization with arrows, circles, or other shapes could provide an additional layer of information to highlight important portions of the data, especially for the most novice audiences.

Finally, visualizations in other domains would likely benefit from the new considerations discovered here. Namely, what is normal in other domains is apt to be unfamiliar, so providing comparison images and highlighting specific

similarities and differences could aid meaning making. Captions that include information on how the data was collected, and the assumptions and methods of the modeling undertaken should also be provided for the user to understand how the visualization relates to the real world.

Implications for informal science education. In the fully-scaffolded cases here, most of the novices still struggled to fluently grasp the academic scientific meaning. They did demonstrate better pattern recognition, especially when prompted with guiding questions, in line with Chapman's (2000) finding that interventions designed to improve chemistry student pattern recognition not only strengthened that skill but also taught molecular structure. These visualizations were not as fully-scaffolded as they would be in educational or other *in situ* naturalistic contexts that are more ecologically valid than the laboratory setting studied here. That means there is more potential for enculturating skills through these tasks with increased supporting information. What remains to be seen is how transferrable those pattern recognition skills are, across scientific disciplines and across domains, that is, from map-based visualizations such as these to other representations of data in scatterplots, line graphs, or even as text, and even further to other disciplines and to situations outside of recognized educational interventions, such as true "in the wild" learning and decision-making.

Designing proper, effective interventions must be a team effort. Visualizers can lend their graphic design experience and technical capability. Educators and education researchers and outreach professionals can lend expertise at designing pedagogical materials and interventions and specific visualization instances based around their specific contexts such as three-dimensional versus flat-screen exhibit, mobile versus traditional web page, or informal versus formal education institution. Scientists who work with the data can lend expertise to help

balance the need for scaffolding skill development with the integrity and accuracy of the scientific content provided to do so.

Implications for formal science education. The current emphasis in formal education on standardized testing of content knowledge, especially in science, conflicts with the picture of expertise presented in this study as based in practice. It is not just generic disciplinary knowledge that transfers for the experts here but rather skill from uses, practices, and purposes. In order to possess enough facts to make meaning from even the three different topics presented in this experiment, one would have to be specialized in so many different sub-sub-disciplines of oceanography that it is impractical. Further, content knowledge at some level is probably irrelevant to the larger task here. Novices may be unable to judge the reason for the seasons, but ultimately to an extent that detail is irrelevant for the broader meaning that scientists want to transmit here in the context of climate change; one could understand that the Earth's ocean is changing, as well as the how and perhaps even the why, without being able to determine whether the Earth is warmer in summer because it is closer to the sun or because of the planet's tilt.

This lack of emphasis on factual correctness is consistent with the constructivist philosophy that learning is not only content acquisition, but also meaning making (Hein, 1998). Meaning making is not retrieval, which even novices proved they could do adequately in the scaffolded cases here. Meaning making is *constructing*, putting all the pieces retrieved from prior knowledge, prior experience, and the tools at hand together coherently in a way relevant to the task on which one is working. The human mind is an adept pattern recognizer and transformer (Gee, 1999), but making use of those patterns in proper context must be learned. Gee suggests this learning of analysis is best done in "secondary institutions" of school or work, whereas acquisition is best

accomplished in “functional, repeated use contexts” such as the home and community.

Pedagogical techniques that reformers are encouraging around contextualized science learning, peer-to-peer and student-centered, practice-based learning (Greene & Land, 2000; Magnusson & Palincsar, 1995; “The Next Generation Science Standards,” n.d.) could all be used in schools for enculturating novices into the academic-visualization-reading community. At the same time, as with Chapman’s work (2000), teaching these skills can provide context for increasing factual knowledge.

In the formal education classroom, students could benefit from a more long-term, in-depth training in reading more varied types of visualizations, creating those visualizations, and developing content knowledge through using visualizations in context. The progression of scaffolding and removal from exposure, to modeling by experts, to practice described previously could be accomplished through integration into science curricula over the years while simultaneously offering content acquisition. To that end, integrated, interdisciplinary curricula are a step in the right direction. Dealing directly with data provides an authenticity with science research and science itself that is a goal of inquiry-based curricula as well (“The Next Generation Science Standards,” n.d.).

Implications for education and outreach professional development. Increasing use of these visualizations in all types of learning settings will require skilled educators to help guide novice users. People are increasingly exposed to these types of data-driven visualizations, in weather reports and documentaries as reported by participants here, but also in schools based on teacher interest. However, to move through the levels of expertise acquisition and scaffolding removal, educators and communicators of all manner will need specific professional development around facilitating this development. Integration of this

professional development and ultimately integration of visualization-reading into the curriculum naturally will support a broader effort to teach science as an active, changing discipline rather than as a history lesson.

How can I facilitate a true unity of everyday and academic scientific culture? “Educator” must also be considered broadly in this effort and specific effort must be made to reach out to some professionals not traditionally considered educators. For example, meteorologists may be one of the primary users of visualizations that the public sees through their weather reports. The traditional routes of publishing and presenting findings from these and similar studies within the fields of science education and informal outreach will not be enough to reach audiences that actually create visualizations, by and large. Educators can, of course, share findings as often as they can with the producers they work with, but a different strategy also could help. First, the multidisciplinary nature of this work will allow us to publish in more neuroscience and psychology journals, although those also are unlikely to get into the hands of visualizers in a straightforward manner. A more viable approach might therefore be to work with organizations such as the National Oceanic and Atmospheric Administration, who are coordinating production and dissemination of many of the visualizations through their Science on a Sphere™ program, of which at least a few meteorologists are already a part, through collaborations with their local science centers or universities, for example. The National Oceanic and Atmospheric Administration has already published a document on best practices for docents who are interpreting visualizations on spherical digital display systems, and are working on similar documentation for creating visualizations. Working with the visualizers and outreach professionals at those larger organizations, including those at the National Aeronautics and Space Administration who created the visualizations for this dissertation, is a direct line to at least some of those visualization creators. They in turn will be able to work with scientists. However, I can go even further and suggest publishing in disciplinary journals that are

starting to focus on discipline-based educational research, and scientific society publications or university education associations aimed at professional development for university professors that might publish a research brief with an education and outreach bent. Targeting graduate students and young researchers might prove especially fruitful. Finally, given the permeation of these visualizations through the news media, that seems to be an untapped audience that could support this mission, first at the meteorological level with the ubiquitous weather map in typically rainbow color scale, through work with professional societies, publications, and meetings there. Second, I should reach out to the scientific journalism community at large, who work hard to translate words into compelling stories but like educators, may lack the resources to do the same with the accompanying visualizations. Reaching broad-based science funding and advocacy organizations such as the American Association for the Advancement of Science and the National Science Foundation could eventually fold these recommendations into their broader impacts trainings and advice for grantees.

Future Work

This dissertation offers several ideas for future work. First and foremost, *in situ* work must be expanded, with design of visualizations optimized for these settings but balanced with design principles revealed by laboratory experiments and design theory. The idea of gender differences in viewing imagery warrants at the least a more gender-balanced and deliberate sampling and experimental procedure to verify the findings on the small sample here.

Methods of meaning-making by expert visualization readers could be probed in more detail. The investigation of saccades to understand the role of memory in the selection of information for viewing, for example, could be investigated (Henderson et al., 2007). Examining the particular cases in my eye-tracking experiments where experts struggle to make meaning and their

particular patterns of visualization use alongside their verbal reports could reveal more information about where expert meaning making breaks down, and detail about how they deal with that break down or attempt to reconcile conflicting or unexpected information, develop hypotheses, and reason in general. More importantly, the data gathered here could lead to better task design for eye-tracking to investigate just that specific information.

More explicit information about expectations of data patterns drawn from everyday science knowledge will reveal more about the disconnects and alternate translations between everyday and academic scientists. While participants were asked to verbally describe their expected patterns of data here, drawings could prove even more powerful at explicitly conveying details that may be alternatively applied in new contexts.

Despite the thorniness of the definition of ‘model’ in science, visualizations are at their core, models, as are all data and their representations. Specific decisions and assumptions are made that impact the interpretation of data, whether it is presented in a text, a graph, or a spatially-based visualization. Better understanding of how learners understand the concept of model and the constraints and affordances provided by the decisions and assumptions that go into their production can lead to better design of visualizations and all communication tools for all types of science audiences.

Conclusion

Expert academic scientists are failing to communicate their results with novice academic scientists because the world does not, as yet, have the complete Rosetta Stone for academic scientific visualizations to translate visualizations for academic scientific language learners. While the elements of the academic visualizations may seem like everyday elements, the ways in which they are put together and the addition of a few specialized elements put the academic visualizations in a different register, just as academic register differs

from everyday speech (Halliday & Hasan, 1976). Designing products to communicate among these and other varied stakeholder groups is non-trivial and does not just mean adding more contextual information in jargon-free language.

The balance between too little and too much detail to provide for widely variable public audiences remains. The integration of text and images can work well for learners, especially those with high existing prior knowledge (Cook et al., 2008). Low prior knowledge users may need more explicit information about comparing the multiple representations, however, in order to be able to effectively integrate text and images, as is suggested by the results in this dissertation. On the other hand, the danger of confusing high-knowledge audiences with too much information is a real one, especially with this relatively unfamiliar task (Kalyuga et al., 2003). Some exhibit and outreach developers have explored the effectiveness of multiple-entry-points to the same content to alleviate this problem (Mikulak, 2009; National Earth Science Teachers Association, n.d.).

There may always be problems with communicating via visualizations, regardless of the audience. James Hansen, climate change researcher, published an article that observers noted had a new color added to the color scale to accommodate the extent of warming due to climate change (2011). The Rutgers Coastal Ocean Observation Lab web site represents the temperatures of the Chesapeake Bay on the same color scale year round, despite the fact that temperatures vary so greatly that red is 66 degrees Fahrenheit in early April but the same shade is 85 degrees Fahrenheit in September (Lichtenwalner, 2010; “Sea Surface Temperature - IMCS Coastal Ocean Observation Lab,” n.d.). Are users confused by this potential confound of color scales? Is a better solution to use such a broad range that most of the range is unutilized throughout the year? Or are these products adequately communicating the academic science that the visualization creators intend? These, too, are questions ripe for further investigation.

Epilogue: Reflection on a Pragmatic Framework

Pragmatism is used primarily to get results to answer a particular question, not necessarily to make disparate theories fit together neatly. As this research was exploring meaning making in a new context, the use of multiple frameworks provided a starting point, allowing research into this new context without assuming which of the single perspectives might best explain the phenomenon. The pragmatic approach used in this dissertation had several advantages and disadvantages that affect the power and limitations of my findings. First and foremost, I was able to explore very practical questions about the differences in two populations and the effects of a particular intervention, but I did not necessarily advance understanding of any of the individual frameworks on which I drew. I had triangulation as an integral part of the study, but I also had to limit the depth of the data I collected the data from each method. I used an experimental task designed to be used in each method, but had to compromise such that the task was not necessarily ideal for any method. I was able to measure the phenomenon of meaning making from visualizations somewhat thoroughly though the phenomenon itself is small in scope. Here, I reflect on what the choice of pragmatism meant for the fashioning and implications of this dissertation.

The use of pragmatism allowed me a great deal of freedom in how I examined the problem of visualization construction for meaning making by everyday scientists, but did not satisfy the conditions of any given perspective entirely. I was able to investigate a broad array of possible influences on the meaning making task from each of the frameworks. However, to a neuroscientist, the task may seem too high-level, whereas to a socioculturalist, the laboratory setting may have been too divorced from context to be truly relevant.

The experiments revealed a great deal of information about the task of meaning making from visualizations of Earth science data that are scaffolded in

particular ways based on the assumptions of the three frameworks. I found that the scaffolds did improve academic meaning making for the everyday scientists. However, the everyday scientists still had room for improvement and by using all three frameworks together, it is not immediately clear which of the strategies might be best for scaffolding further improvement. This also meant I did not necessarily find insights that advance any of the theoretical understanding of the particular frameworks.

By using multiple methodologies, I also built triangulation into the fabric of the project. I was able to confirm findings from one experiment with another as when I saw low scores in interviews and long fixation durations both indicating a lack of academic meaning making. I also got insight into meaning of the findings of one experiment using findings from another. For example, when there were no significant differences in the quantitative spontaneous looking fixation durations between experts and novices, the different patterns in eye-tracking scan paths clarified that qualitative differences did exist.

Moreover, I was able to triangulate the different perspectives of my framework by using the same task with methods and traditions from each. In the same way that the results confirmed and clarified each other, the triangulation of perspectives can do the same. While both groups drew on prior knowledge and experience as expected in a constructivist framework, the particular prior experiences especially were truly socioculturally based. Deliberately considering both possible frameworks allowed me to collect evidence to start to distinguish between them. While I have not yet delved into fixations on individual colors within the visualizations, the fact that fixation patterns changed with different colors and the lack of answers in the interviews suggesting that eyes were drawn to yellow both point to the probability that cultural influences outweigh perceptual ones, at least with a certain level of experience. The results further point to ways to design future experiments to tease out more subtle influences as well; experimental tasks built without the familiar anchors of the continents and with

less realistic data could be designed in order to probe whether perceptual influences win out in the face of minimal experience with which to compare new data.

However, I did make tradeoffs in order to accomplish these things. I collected many types of data, but did not go very far in depth on any of them. In the interviews, for example, I had to cut off the in-depth probing in several cases in order to present multiple versions of the visualizations. On the other hand, in order to probe on some questions, I did not show all versions to all participants. In the eye-tracking experiment, in order to perform the interviews, I sacrificed having all of the trials be exactly the same length.

The task, designed to accommodate all three perspectives, meant that it was not ideal for examining any individual framework. As mentioned, the cultural scaffolding provided by the familiar context of the Earth may have masked the influence of the perceptual constraints in the data. At the very least, the tight controls typically designed for a neuroscientific investigation were missing from this analysis. On the other hand, by attempting to control for learning effects and effects of the different topics during the experiment by randomizing presentation of the stimuli meant that I was unable to probe participants on all of the topics or all of the versions of scaffolding in a true sociocultural or constructivist manner.

In the end, “simple and robust [did not] go hand in hand” (A. Lobben, personal communication, April 29, 2013). Despite the complexity of the experimental design, the use of multiple frameworks led me to collect a great deal of data of various types that I am only just beginning to make meaning from myself. As they are the result of a task designed for all three frameworks, despite the limitations of that compromise, I can examine the data from any individual framework in the future as well. The power afforded by the design in this initial exploration of the context outweighs the limitations at this point and gave me a strong foundation on which to build using any or all of the frameworks. Finally, this is one of the first examples of which I am aware of an attempt to reconcile a

learning theory framework with a neurocognitive tradition by using a meaning making task with a neuroscience methodology. This reconciliation is simply not possible without a bit of consideration of two disparate frameworks; if true progress is to be made on bringing the findings of neuroscience and the constraints and affordances of the physical human brain to bear on practical learning research, this complexity and collaboration must be embraced.

Bibliography

Abdullaev, Y. G., & Posner, M. I. (1997). Time course of activating brain areas in generating verbal associations. *Psychological Science*, *8*(1), 56–59.

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, *16*(3), 183–198. doi:10.1016/j.learninstruc.2006.03.001

Allen, S. (2004). Designs for learning: Studying science museum exhibits that do more than entertain. *Science Education*, *88*(Suppl. 1), S17–S33. doi:10.1002/sce.20016

Alley, M., Schreiber, M., Ramsdell, K., & Muffo, J. (2006). How the design of headlines in presentation slides affects audience retention. *Technical Communication*, *53*(2), 225–234.

Allum, N., Sturgis, P., Tabourazi, D., & Brunton-Smith, I. (2008). Science knowledge and attitudes across cultures: a meta-analysis. *Public Understanding of Science*, *17*(1), 35–54. doi:10.1177/0963662506070159

Bailer-Jones, D. M. (2003). When scientific models represent. *International Studies in the Philosophy of Science*, *17*(1), 59–74.

Baker, R., & Dwyer, F. (2000). A meta-analytic assessment of the effect of visualized instruction. *International Journal of Instructional Media*, *27*(4), 417–426.

Bell, P., Lewenstein, B., Shouse, A. W., & Feder, M. A. (Eds.). (2009). *Learning science in informal environments: People, places, and pursuits*. Washington, DC: National Academies Press.

Bernard, H. (2005). *Research methods in anthropology : qualitative and quantitative approaches* (4th ed.). Lanham MD: AltaMira Press.

Bevlin, M. E. (1977). *Design through discovery* (3rd ed.). New York: Holt, Reinhart, & Winston.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., & Wilderman, C. C. (2009). Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report. *Online Submission*. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED519688

Borun, M., & Dritsas, J. (1997). Developing Family-Friendly Exhibits. *Curator: The Museum Journal*, *40*(3), 178–196. doi:10.1111/j.2151-6952.1997.tb01302.x

Bransford, J., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How People Learn : Brain, Mind, Experience, and School* (Expanded ed. --.). Washington D.C.: National Academy Press.

Brewer, C. A., Hatchard, G. W., & Harrower, M. A. (2003). ColorBrewer in Print: A Catalog of Color Schemes for Maps. *Cartography and Geographic Information Society*, *30*(1), 5–32. doi:10.1559/152304003100010929

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher*, *18*(1), 32–42.

Bruer, J. T. (2006). On the implications of neuroscience research for science teaching and learning: Are there any? A skeptical theme and variations: The primacy of psychology in the science of learning. *CBE Life Science Education*, *5*(2), 104–110.

Bukach, C. M., Gauthier, I., & Tarr, M. J. (2006). Beyond faces and modularity: the power of an expertise framework. *Trends in Cognitive Sciences*, *10*(4), 159–166. doi:10.1016/j.tics.2006.02.004

Burch, M., Konevtsova, N., Heinrich, J., Hoeferlin, M., & Weiskopf, D. (2011). Evaluation of Traditional, Orthogonal, and Radial Tree Diagrams by an Eye Tracking Study. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2440–2448. doi:10.1109/TVCG.2011.193

Casey, K. S., & Cornillon, P. (1985). *NSIPP AVHRR Pathfinder and Erosion Global 9km SST Climatology*. NASA JPL Physical Oceanography DAAC.

Chakravartty, A. (2010). Informational versus functional theories of scientific representation. *Synthese*, *172*(2), 197–213.

Chapman, O. L. (2000). Learning Science Involves Language, Experience, and Modeling. *Journal of Applied Developmental Psychology*, *21*(1), 97–108. doi:10.1016/S0193-3973(99)00053-2

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–61.

Cheng, P. C.-H., & Peebles, D. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. (Special Section). *Human Factors*, *45*(1), 28+.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices*. *Cognitive Science*, *5*(2), 121–152. doi:10.1207/s15516709cog0502_2

Cocking, R. R., Mestre, J. P., & Brown, A. L. (2000). New developments in the science of learning: Using research to help students learn science and mathematics. *Journal of Applied Developmental Psychology*, *21*(1), 1–11.

Compton, P. E., Grossenbacher, P., Posner, M. I., & Tucker, D. M. (1991). A cognitive-anatomical approach to attention in lexical access. *Journal of Cognitive Neuroscience*, *3*(4), 304–312.

Conroy, E. (1998). *The symbolism of color: (1921)*. Kila, MT: Kessinger Pub.

Cook, M., Wiebe, E. N., & Carter, G. (2008). The interpretation of cellular transport graphics by students with low and high prior knowledge. *International Journal of Science Education*, *30*(2), 239 – 261.

CoVis Information. (n.d.). Retrieved January 12, 2012, from http://www.covis.northwestern.edu/info/

Cremin, L. A. (1990). *American education: The metropolitan experience, 1876-1980*. New York: Perennial Library.

Da Costa, N., & French, S. (2000). Models, theories, and structures: Thirty years on. *Philosophy of Science*, *67*, S116–S127.

Dallas, D. (2006). Café Scientifique—Déjà Vu. *Cell*, *126*(2), 227–229. doi:10.1016/j.cell.2006.07.006

Data Visualization | SAS. (n.d.). Retrieved April 19, 2013, from http://www.sas.com/data-visualization/overview.html

Davidson, R. J. (2003). Affective neuroscience and psychophysiology: Toward a synthesis. *Psychophysiology*, *40*(5), 655–665.

Dean, D. J. (n.d.). The Arthur H. Robinson Map Library at the University of Wisconsin-Madison. *The Robinson Projection*. Retrieved February 11, 2013, from http://www.geography.wisc.edu/maplib/robinson_projection.html

Develaki, M. (2007). The model-based view of scientific theories and the structuring of school science programmes. *Science & Education*, *16*(7), 725–749.

Driver, R. (1995). Constructivist approaches to science teaching. In *Constructivism in Education*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Duchowski, A. T. (2007). *Eye Tracking Methodology* (Second.). London: Springer Science+Business Media.

Dunbar, K. (2000). How Scientists Think in the Real World: Implications for Science Education. *Journal of Applied Developmental Psychology*, *21*(1), 49–58. doi:10.1016/S0193-3973(99)00050-7

Eberbach, C., & Crowley, K. (2009). From everyday to scientific observation: How children learn to observe the biologist's world. *Review of Educational Research*, *79*(1), 39 –68. doi:10.3102/0034654308325899

Edelson, D. C. (1997). Realising authentic science learning through the adaptation of scientific practice. In B. Fraser & K. Tobin (Eds.), *International handbook of science education* (pp. 317–331). Dordrecht, NL: Kluwer Academic Publishers.

Edelson, D. C., & O'Neill, D. K. (1994). The CoVis collaboratory notebook: supporting collaborative scientific inquiry. Presented at the National Educational Computing Conference, Boston, MA.

Edelson, D. C., Pea, R. D., & Louis D. Gomez. (1996). Constructivism in the collaboratory. In B. G. Wilson (Ed.), *Constructivist learning environments: Case Studies in Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications, Inc.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. doi:10.1037/0033-295X.100.3.363

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: evidence of maximal adaption to task constraints. *Annual Review of Psychology*, *47*, 273+.

ExperimentCenter Manual version 3.1. (2012, March). SensoMotoric Instruments.

Falk, J. H., & Dierking, L. D. (2000). *Learning From Museums: Visitor Experiences and the Making of Meaning*. Lanham, MD: AltaMira Press.

Falk, J. H., & Dierking, L. D. (2010). The 95 Percent Solution. *American Scientist*, *98*(6), 486. doi:10.1511/2010.87.486

Falk, J. H., & Needham, M. D. (2011). Measuring the impact of a science center on its community. *Journal of Research in Science Teaching*, *48*(1), 1–12. doi:10.1002/tea.20394

Falk, J. H., & Needham, M. D. (2013). Factors Contributing to Adult Knowledge of Science and Technology. *Journal of Research in Science Teaching*, *50*(4), 431–452. doi:10.1002/tea.21080

Faughn, J. S., & Serway, R. A. (2003). *College Physics* (6th ed.). Canada: Thomson, Brooks/Cole.

Fayden, T. (2005). *How children learn: getting beyond the deficit myth*. Boulder, CO: Paradigm Publishers.

Filippini-Fantoni, S., Jaebker, K., Bauer, D., & Stofer, K. (2013). Capturing visitors' gazes: Three eye tracking studies in museums. Presented at the Museums and the Web, Portland, OR. Retrieved from http://mw2013.museumsandtheweb.com/paper/capturing-visitors-gazes-three-eye-tracking-studies-in-museums/

French, J., Ekstrom, R., & Price, L. (1963). *Kit of reference tests for cognitive factors*. Princeton, NJ: Educational Testing Service.

Frigg, R., & Hartmann, S. (2006, February 27). Models in science. In (E. N. Zalta, Ed.)*Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*. Retrieved from http://plato.stanford.edu/entries/models-science/

Gauthier, I., & Tarr, M. J. (2002). Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 431–446. doi:10.1037/0096-1523.28.2.431

Gee, J. P. (1999). *An Introduction to Discourse Analysis: Theory and Method*. London: Routledge.

Giere, R. (1999). Using models to represent reality. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-Based Reasoning in Scientific Discovery*. New York: Kluwer/Plenum.

Giere, R. (2010). An agent-based conception of models and scientific representation. *Synthese*, *172*(2), 269–281.

Gigerenzer, G., Hertwig, R., Van Den Broek, E., Fasolo, B., & Katsikopoulos, K. V. (2005). “A 30% chance of rain tomorrow”: How does the public understand probabilistic weather forecasts? *Risk Analysis*, *25*(3), 623–629. doi:10.1111/j.1539-6924.2005.00608.x

Gilhooly, K. J., Wood, M., Kinnear, P. R., & Green, C. (1988). Skill in map reading and memory for maps. *The Quarterly Journal of Experimental Psychology*, *40A*, 87–107.

Gläscher, J., Rudrauf, D., Colom, R., Paul, L. K., Tranel, D., Damasio, H., & Adolphs, R. (2010). Distributed neural system for general intelligence revealed by lesion mapping. *Proceedings of the National Academy of Sciences*, *107*, 4705–4709.

Global Maps. (n.d.). Retrieved April 17, 2013, from http://earthobservatory.nasa.gov/GlobalMaps/

Glossary. (n.d.). Retrieved April 17, 2013, from http://chl.erdc.usace.army.mil/glossary

González, N., Moll, L. C., & Amanti, C. (Eds.). (2005). *Funds of knowledge: theorizing practice in households, communities, and classrooms*. Mahwah, N.J: L. Erlbaum Associates.

Goodwin, C. (1994). Professional Vision. *American Anthropologist*, *96*(3), 606–633. doi:10.1525/aa.1994.96.3.02a00100

Graham, L. (2002). *Basics of Design: Layout and Typography for Beginners*. Albany, NY: Delmar.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, *14*(5), 462–466. doi:10.2307/40064168

Grecucci, A., Giorgetta, C., van't Wout, M., Bonini, N., & Sanfey, A. G. (2012). Reappraising the ultimatum: an fMRI study of emotion regulation and decision making. *Cerebral Cortex*, *23*(2), 399–410. doi:10.1093/cercor/bhs028

Greene, B. A., & Land, S. M. (2000). A Qualitative Analysis of Scaffolding Use in a Resource-Based Learning Environment Involving the World Wide Web. *Journal of Educational Computing Research*, *23*(2), 151–179. doi:10.2190/1GUB-8UE9-NW80-CQAD

Greenfield, P. M. (1999). A theory of the teacher in the learning activities of everyday life. In B. Rogoff & J. Lave (Eds.), *Everyday Cognition* (2nd ed., pp. 117–138). Cambridge, MA: Harvard University Press.

Grenier, R. S. (2009). The role of learning in the development of expertise in museum docents. *Adult Education Quarterly*, *59*, 142–157.

Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching*, *28*(9), 799–822. doi:10.1002/tea.3660280907

Gutwill, J. P., & Allen, S. (2010). Facilitating family group inquiry at science museum exhibits. *Science Education*, *94*(4), 710–742. doi:10.1002/sce.20387

Hacking, I. (1983). *Representing and Intervening*. Cambridge: Cambridge University Press.

Haley Goldman, K., Kessler, C., & Danter, E. (2010, September). Science on a Sphere: Cross-site summative evaluation. Institute for Learning Innovation. Retrieved from http://www.oesd.noaa.gov/network/SOS_evals/SOS_Final_Summative_Report.pdf

Halliday, M. A. K. (1975). *Learning How to Mean--Explorations in the Development of Language.* London: Edward Arnold. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED105507

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Harrower, M., & Brewer, C. A. (2011). ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. In *The Map Reader* (pp. 261–268). John Wiley & Sons, Ltd. Retrieved from http://dx.doi.org/10.1002/9780470979587.ch34

Hein, G. E. (1998). *Learning in the museum*. London ; New York: Routledge.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye Movements: A Window on Mind and Brain*. Oxford, UK: Elsevier Ltd.

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems. *Journal of the Learning Sciences*, *16*(3), 307–331.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Weijer, J. van de. (2011). *Eye Tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Hooge, I. T. C., Vlaskamp, B. N. S., & Over, E. A. B. (2007). Saccadic search: On the duration of a fixation. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye Movements: A Window on Mind and Brain* (pp. 581–595). Oxford, UK: Elsevier Ltd.

Hughes, H. C., Kitterle, F., & Nozawa, G. (1996, May). Global precedence, spatial frequency channels, and the statistics of natural images. *Journal of Cognitive Neuroscience*, *8*(3), 197+.

Jarodzka, H., Scheiter, K., Gerjets, P., & Gemballa, S. (2008). In the eyes of experts: Teaching dynamic features in biology by modeling experts' eye movement strategies to novices. In *Cre8ing a learning world: Proceedings of the 8th International Conference for the Learning Sciences, ICLS '08* (Vol. 1). Presented at the 8th International Conference for the Learning Sciences, ICLS '08, Utrecht, The Netherlands.

Justi, R., & Gilbert, J. (2000). History and philosophy of science through models: some challenges in the case of "the atom." *International Journal of Science Education*, *22*(9), 993–1009. doi:10.1080/095006900416875

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The Expertise Reversal Effect. *Educational Psychologist*, *38*(1), 23–31. doi:10.1207/S15326985EP3801_4

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of neural science* (4th ed.). New York: McGraw-Hill, Health Professions Division.

Karp, J., & Leblang, J. (2004). *Making models: Summative evaluation report April–September 2004*. Cambridge, MA: Program Evaluation and Research Group at Lesley University.

Kastens, K. A., & Ishikawa, T. (2006). Spatial thinking in the geosciences and cognitive sciences: A cross-disciplinary look at the intersection of two fields. In C. A. Manduca & D. W. Mogk (Eds.), *Earth and mind: How geologists think and learn about the earth* (pp. 53–76). Geological Society of America.

Kate. (2011, April 5). Climate change breaks NASA's temperature charts. *Climate Safety*. blog. Retrieved April 15, 2013, from http://climatesafety.org/climate-change-breaks-nasas-temperature-charts/

Kim, S.-H., Dong, Z., Xian, H., Upatising, B., & Yi, J. S. (2012). Does an Eye Tracker Tell the Truth about Visualizations?: Findings while Investigating Visualizations for Decision Making. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2421–2430. doi:10.1109/TVCG.2012.215

Kincheloe, J. L., & Berry, K. S. (2004). *Rigour and complexity in educational research: Conceptualizing the bricolage*. Berkshire, England: Open University Press.

Koffka, K. (1935). *Principles of Gestalt psychology*. Oxford, UK: Harcourt, Brace.

Kress, G. R. (2010). *Multimodality: a social semiotic approach to contemporary communication*. London ; New York: Routledge.

Ladson-Billings, G. (1995). Toward a Theory of Culturally Relevant Pedagogy. *American Educational Research Journal*, *32*(3), 465–491. doi:10.3102/00028312032003465

Larkin, J. H., & Simon, H. A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, *11*(1), 65–100. doi:10.1111/j.1551-6708.1987.tb00863.x

Lave, J., & Wenger, E. (1991). Legitimate peripheral participation in communities of practice. In *Situated learning: Legitimate peripheral participation.* (pp. 90–117). Cambridge: Cambridge University Press.

Lehr, J. L., McCallie, E., Davies, S. R., Caron, B. R., Gammon, B., & Duensing, S. (2007). The Value of “Dialogue Events” as Sites of Learning: An exploration of research and evaluation frameworks. *International Journal of Science Education*, *29*(12), 1467–1487. doi:10.1080/09500690701494092

Leontyev, A. (2009). *The Development of Mind: Selected Works of Aleksei Nikolaevich Leontyev*. Pacifica, CA: Marxist Internet Archive.

Lesgold, A. M., Rubinson, H., Feltovich, P. J., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing X-Ray pictures. In M. T. H. Chi, R. Glaser, & M. Farr (Eds.), *The Nature of Expertise* (pp. 311–342). Hillsdale, New Jersey: Erlbaum.

Levi-Strauss, C. (1966). *The savage mind*. Chicago: The University of Chicago Press.

Libarkin, J. C., Clark, S. K., & Simmon, R. (n.d.). Preliminary Findings from Eye Tracking Study of novice and expert interactions with Global Temperature Maps. Retrieved from https://www.msu.edu/~libarkin/research_eye.html

Lichtenwalner, S. (2010, July 6). Sea Surface Temperature. *Visual Ocean*. blog. Retrieved April 15, 2013, from http://coseenow.net/visual-ocean/2010/07/sea-surface-temperature/

Light, A., & Bartlein, P. J. (2004). The end of the rainbow? Color schemes for improved data graphics. *EOS Transactions of the American Geophysical Union*, *85*(40), 385–391.

Linn, M. C., & Peterson, A. C. (1986). A meta-analysis of gender differences in spatial ability: Implications for mathematics and science achievement. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through meta-analysis.* Baltimore: The Johns Hopkins University Press.

Lobben, A. K. (2007). Navigational map reading: Predicting performance and identifying relative influence of map-related abilities. *Annals of the Association of American Geographers*, *97*(1), 64–85. doi:10.1111/j.1467-8306.2007.00524.x

Lobben, A. K., Lawrence, M., & Olson, J. M. (2009). fMRI and human subjects research in cartography. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *44*(3), 159–169.

Lobben, A. K., Olson, J. M., & Huang, J. (2005). Using fMRI in cartographic research. In *Proceedings of the 22nd International Cartographic Conference* (p. 10). Presented at the International Cartographic Conference, A Coruna, Spain.

Lotman, Y. (1988). Text within a text. *Soviet Psychology*, *XXVI*(3), 32–51.

Lucy, J. (2011). Language and cognition: The view from anthropology. In V. Cook & B. Bassetti (Eds.), *Language and Bilingual Cognition* (pp. 43–68). New York: Psychology Press.

Lurija, A. (1979). *The making of mind : a personal account of Soviet psychology*. (M. Cole, Ed.). Cambridge Mass: Harvard Univ. Press.

Lurija, A. (1981). *Language and Cognition*. (J. V. Wertsch, Ed.). Washington, DC: V. H. Winston & Sons.

MacEachren, A. M., & Kraak, M.-J. (2001). Research challenges in geovisualization. *Cartography and Geographic Information Science*, *28*(1), 3–12.

Magnusson, S. J., & Palincsar, A. S. (1995). The Learning Environment as a Site of Science Education Reform. *Theory into Practice*, *34*(1), 43–50. doi:10.2307/1476543

Magnusson, S. J., Palincsar, A. S., & Templin, M. (2004). Community, culture, and conversation in inquiry based science instruction. *Scientific Inquiry and Nature of Science*. Retrieved from http://dx.doi.org/10.1007/978-1-4020-5814-1_7

Masson, S., Potvin, P., Riopel, M., Brault Foisy, L.-M., & Lafortune, S. (2012). Using fMRI to study conceptual change: Why and how? *International Journal of Environmental & Science Education*, *7*(1), 19–35.

McCallie, E. (2010). *Argumentation among publics and scientists: A study of dialogue events on socio-scientific issues* (Doctoral Dissertation). King's College London, University College London, London.

Memmott, M. (2013, April 2). Obama Says $100 Million Will Be Invested In Brain-Mapping Initiative : NPR. *NPR.org*. Retrieved April 23, 2013, from http://www.npr.org/blogs/thetwo-way/2013/04/02/176013635/obama-says-100-million-will-be-invested-in-brain-mapping-initiative

Middendorf, J., & Pace, D. (2004). Decoding the disciplines: A model for helping students learn disciplinary ways of thinking. *New Directions for Teaching and Learning*, *2004*(98), 1–12.

Mikulak, S. E. (2009, June 10). *The development and evaluation of an interactive exhibit to support real-time water quality data interpretation by the public at an informal education setting* (Master's thesis). Oregon State University, Corvallis, OR. Retrieved from http://ir.library.oregonstate.edu/xmlui/handle/1957/12002?show=full

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *101*(2), 343–352.

Miller, J. D. (2004). Public Understanding of, and Attitudes toward, Scientific Research: What We Know and What We Need to Know. *Public Understanding of Science*, *13*(3), 273–294.

Miller, J. D. (2010). Adult Science Learning in the Internet Era. *Curator: The Museum Journal*, *53*(2), 191–208. doi:10.1111/j.2151-6952.2010.00019.x

Miller, J. D., & Pardo, R. (2000). Civic scientific literacy and attitude to science and technology: A comparative analysis of the European Union, the United States, Japan, and Canada. In M. Dierkes & C. von Grote (Eds.), *Between understanding and trust: The public, science and technology* (pp. 81–130). Abingdon, Oxon: Routledge.

Miller, S. (2001). Public understanding of science at the crossroads. *Public Understanding of Science*, *10*(1), 115 –120. doi:10.1088/0963-6625/10/1/308

Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. In M. S. Morgan & M. Morrison (Eds.), *Models as Mediators: Perspectives on Natural and Social Sciences* (pp. 10–37). Cambridge: Cambridge University Press.

National Academies Press (U.S.). (2006). *Learning to think spatially*. Washington, D.C: National Academies Press.

National Earth Science Teachers Association. (n.d.). Windows to the Universe. Retrieved April 16, 2013, from http://www.windows2universe.org/

National Research Council (U.S.). (1996). *National Science Education Standards : observe, interact, change, learn.* Washington DC: National Academy Press.

Nelson, A. G. (2006, July). Science on a Sphere: Formative evaluation report. Science Museum of Minnesota's Department of Evaluation and Research in Learning.

Nelson, A. G., & Ellenbogen, K. M. (2006). *Science On a Sphere Front-End Evaluation Report*. Minneapolis, MN: Science Museum of Minnesota.

Newberg, A., Alavi, A., Baime, M., Pourdehnad, M., Santanna, J., & d' Aquili, E. (2001). The measurement of regional cerebral blood flow during the complex cognitive task of meditation: a preliminary SPECT study. *Psychiatry Research: Neuroimaging*, *106*(2), 113–122. doi:10.1016/S0925-4927(01)00074-9

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, *108*(2), 291–310. doi:10.1037/0033-295X.108.2.291

Olson, J. M., Lobben, A., & Huang, J. (2005). An experimental study comparing two map tasks using functional magnetic resonance imagery. Presented at the Association of American Geographers, Denver, CO.

Paradise, R., & Rogoff, B. (2009). Side by side: Learning by observing and pitching in. *Ethos*, *37*(1), 102–138.

Park, C. C. (2007). *A dictionary of environment and conservation*. Oxford ; New York: Oxford University Press.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123. doi:10.1016/S0042-6989(01)00250-4

Patton, M. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks Calif.; London: Sage.

Pedretti, E. G. (2004). Perspectives on learning through research on critical issues-based science center exhibitions. *Science Education*, *88*(S1), S34–S47. doi:10.1002/sce.20019

Phipps, M., & Rowe, S. M. (2010). Seeing satellite data. *Public Understanding of Science*, *19*(3), 311 –321. doi:10.1177/0963662508098684

Piaget, J. (1929). *The Child's Conception of the World*. New York: Harcourt.

Piaget, J. (1967). l'Épistémologie et ses variétés. In *Encyclopédie de la Pléiade. Logique et connaissance scientifique*. Paris: Gallimard.

Piaget, J., & Inhelder, B. (1956). *The child's conception of space*. Routledge & K. Paul.

Portides, D. (2007). The relation between idealisation and approximation in scientific model construction. *Science & Education*, *16*(7), 699–724.

Posner, G. J., & Gertzog, W. A. (1982). The clinical interview and the measurement of conceptual change. *Science Education*, *66*(2), 195–209. doi:10.1002/sce.3730660206

Products. (2008). Global Imagination. Retrieved from http://globalimagination.com/products.html

Project 2061 (American Association for the Advancement of Science). (1993). *Benchmarks For Science Literacy*. New York: Oxford University Press.

Ragauskayte, Y. (n.d.). Becoming a university professor. *Journal of Young Investigators: Science Career Center*. Retrieved April 10, 2013, from http://legacy.jyi.org/SCC/Article.php?articleNum=87

Raloff, J. (2010, March 13). Science literacy: U.S. college courses really count. *Science News*, *177*(6). Retrieved from http://www.sciencenews.org/view/generic/id/56517/description/Science_literacy_US_college_courses_really_count

Random.org - True Random Number Service. (n.d.). Retrieved April 16, 2013, from http://www.random.org/

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372.

Reese, W. J. (1999). *The origins of the American high school.* New Haven; London: Yale University Press.

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science*, *12*, 49–56.

Rennie, L., & Stocklmayer, S. M. (2003). The communication of science and technology: Past, present and future agendas. *International Journal of Science Education*, *25*(6), 759–773. doi:10.1080/09500690305020

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, *5*, 121–125.

Roediger, H. L., Dudai, Y., & Fitzpatrick, S. M. (2007). *Science of Memory: Concepts*. New York, NY: Oxford University Press.

Roschelle, J. (1995). Learning in interactive environments: Prior knowledge and new experience. In J. H. Falk & L. D. Dierking (Eds.), *Public institutions for personal learning: Establishing a research agenda* (pp. 37–51). Washington, DC: American Association of Museums.

Roth, W.-M., Lawless, D. V., & Masciotra, D. (2001). Spielraum and Teaching. *Curriculum Inquiry*, *31*(2), 183.

Rowe, S. M. (2002). The role of objects in active, distributed meaning-making. In S. G. Paris (Ed.), *Perspectives on object-centered learning in museums* (pp. 17–32). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Rowe, S. M., Stofer, K. A., & Barthel, C. (in preparation). Supporting meaning making with complex visualizations.

Rowe, S. M., Stofer, K., Barthel, C., & Hunter, N. (2010). *Hatfield Marine Science Center Magic Planet Installation Evaluation Findings*. Corvallis, OR: Oregon Sea Grant.

Rowe, S. M., Stofer, K., Bullick, S., & O'Brien, S. (2011, August 29). GEO SSI project Phase 1 evaluation results. Oregon State University.

Russell, S. H., Hancock, M. P., & McCullough, J. (2007). Benefits of undergraduate research experiences. *Science*, *316*, 548–549.

Scaife, M., & Rogers, Y. (1996). External cognition: how do graphical representations work? *International Journal of Human-Computer Studies*, *45*(2), 185–213. doi:10.1006/ijhc.1996.0048

Schacter, D. L. (1996). *Searching for memory: the brain, the mind, and the past*. New York: Basic Books. Retrieved from http://books.google.com/books?id=ZYMhfK1qwBYC

Sea Surface Temperature - IMCS Coastal Ocean Observation Lab. (n.d.). Retrieved April 15, 2013, from http://marine.rutgers.edu/mrs/sat_data/?nothumbs=0&product=sst

Selvakumar, M., & Storksdieck, M. (2013). Portal to the Public: Museum Educators Collaborating with Scientists to Engage Museum Visitors with Current Science. *Curator: The Museum Journal*, *56*(1), 69–78. doi:10.1111/cura.12007

Shen, B. S. P. (1975). Science literacy and the public understanding of science. In *Communication of Scientific Information* (pp. 44–52). Basel: Karger.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Slocum, T. A., Blok, C., Jiang, B., Koussoulakou, A., Montello, D. R., Fuhrmann, S., & Hedley, N. R. (2001). Cognitive and usability issues in geovisualization. *Cartography and Geographic Information Science*, *28*(1), 61–75.

Smithson, J. (2006). File:Rubin2.jpg. In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/wiki/File:Rubin2.jpg

Steffke, C., & Libarkin, J. (2012, November 4). *Guiding symbology and display selection to produce more effective images for conveying information*. Poster presented at the Geological Society of America, Charlotte, NC.

Steinberger, M., Waldner, M., Streit, M., Lex, A., & Schmalstieg, D. (2011). Context-Preserving Visual Links. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2249–2258. doi:10.1109/TVCG.2011.183

Tarr, M. J., & Gauthier, I. (2000). FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, *3*(8), 764.

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, *2*, 1–18.

Thayer, H. S. (Ed.). (1982). *Pragmatism, the classic writings: Charles Sanders Peirce, William James, Clarence Irving Lewis, John Dewey, George Herbert Mead*. Indianapolis: Hackett Pub. Co.

The Next Generation Science Standards. (n.d.). Retrieved April 19, 2013, from http://www.nextgenscience.org/next-generation-science-standards

Thomas, L. E., & Lleras, A. (2007). Moving eyes and moving thought: On the spatial compatibility between eye movements and cognition. *Psychonomic Bulletin & Review*, *14*(4), 663–668. doi:10.3758/BF03196818

Tobin, K. G. (2005). Urban science as a culturally and socially adaptive practice. In K. G. Tobin, R. Elmesky, & G. Seiler (Eds.), *Improving Urban Science*

Education: New Roles For Teachers, Students, And Researchers (pp. 21–42). Rowman & Littlefield.

Treagust, D., Chittleborough, G., & Mamiala, T. (2002). Students' understanding of the role of scientific models in learning science. *International Journal of Science Education*, *24*(4), 357–368. doi:10.1080/09500690110066485

U. S. Census Bureau, D. I. S. (2008). 2008 National Population Projections. Retrieved May 7, 2013, from http://www.census.gov/population/projections/data/national/2008.html

Van Gog, T. (2006). *Uncovering the problem-solving process to design effective worked examples* (Doctoral Dissertation). OpenUniversiteitNederland, Heerlen, The Netherlands.

Vygotsky, L. (1978). *Mind in society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Ware, C. (2004). *Information visualization: Perception for design*. San Francisco, CA: Morgan Kaufman Publishers.

Ware, C. (2008). *Visual thinking for design*. Burlington, MA: Morgan Kaufmann. Retrieved from http://www.engineeringvillage.com/controller/servlet/OpenURL?genre=book&isbn=9780123708960

Webvision. (2011, March 30). The primary visual cortex. Retrieved June 3, 2011, from http://webvision.med.utah.edu/book/part-ix-psychophysics-of-vision/the-primary-visual-cortex/

Weiner, L. (2006). Challenging deficit thinking. *Educational Leadership*, *64*, 42–45.

Wertsch, J. V. (1994). The primacy of mediated action in sociocultural studies. *Mind, Culture, and Activity*, *1*(4), 202–208. doi:10.1080/10749039409524672

Wertsch, J. V. (2007). Mediation. In H. Daniels, M. Cole, & J. V. Wertsch (Eds.), *The Cambridge Companion to Vygotsky*. New York, NY: Cambridge University Press.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. doi:10.1111/j.1469-7610.1976.tb00381.x

Zeidler, D. L., Sadler, T. D., Simmons, M. L., & Howes, E. V. (2005). Beyond STS: A research-based framework for socioscientific issues education. *Science Education*, *89*(3), 357–377. doi:10.1002/sce.20048

Zhang, J., & Norman, D. A. (1994). Representations in Distributed Cognitive Tasks. *Cognitive Science*, *18*(1), 87–122. doi:10.1207/s15516709cog1801_3

APPENDICES

Appendix 1 - Stimulus Visualizations

All three unscaffolded and fully-scaffolded visualizations are presented here. Visualizations for the geography, color, and title scaffolding have a single element scaffolded as shown in the fully-scaffolded case, but are not presented here.

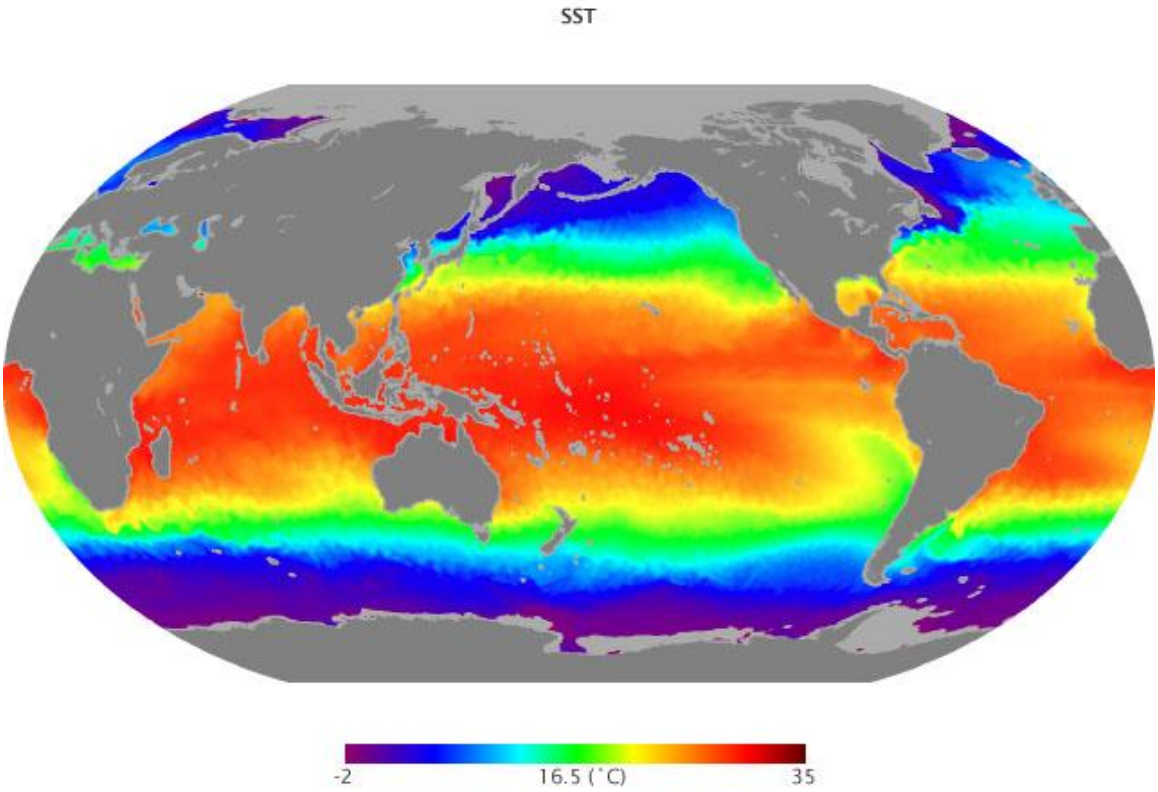


Figure A1. Unscaffolded (US) SST (sea surface temperature) visualization. This figure illustrates a visualization without any culturally-familiar supporting information added.

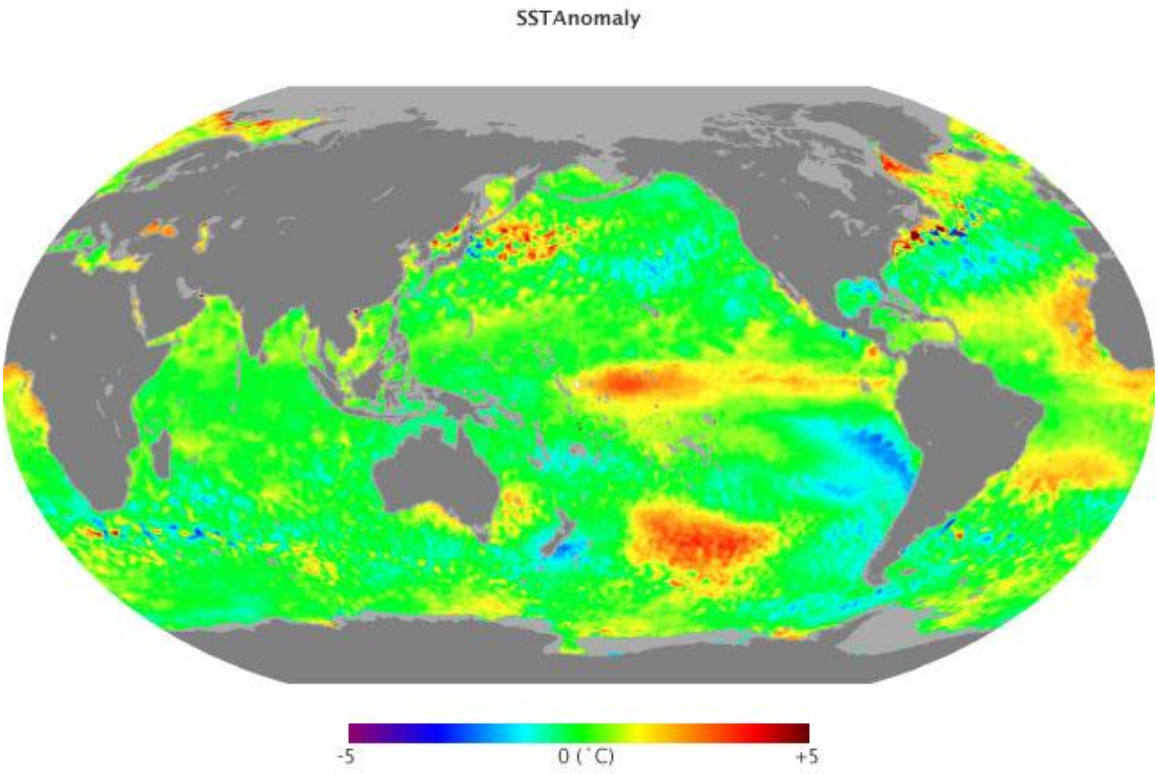


Figure A2. Unscaffolded (US) SST anomaly visualization

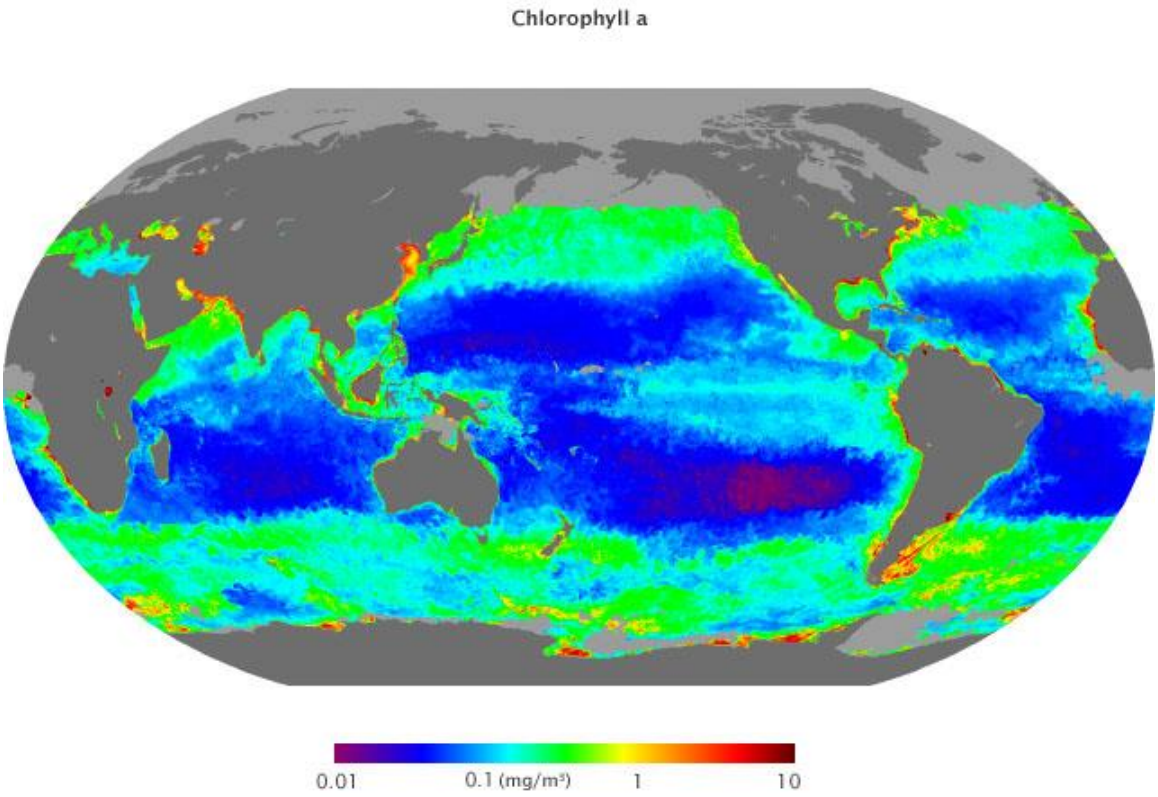


Figure A3. Unscaffolded (US) chlorophyll visualization

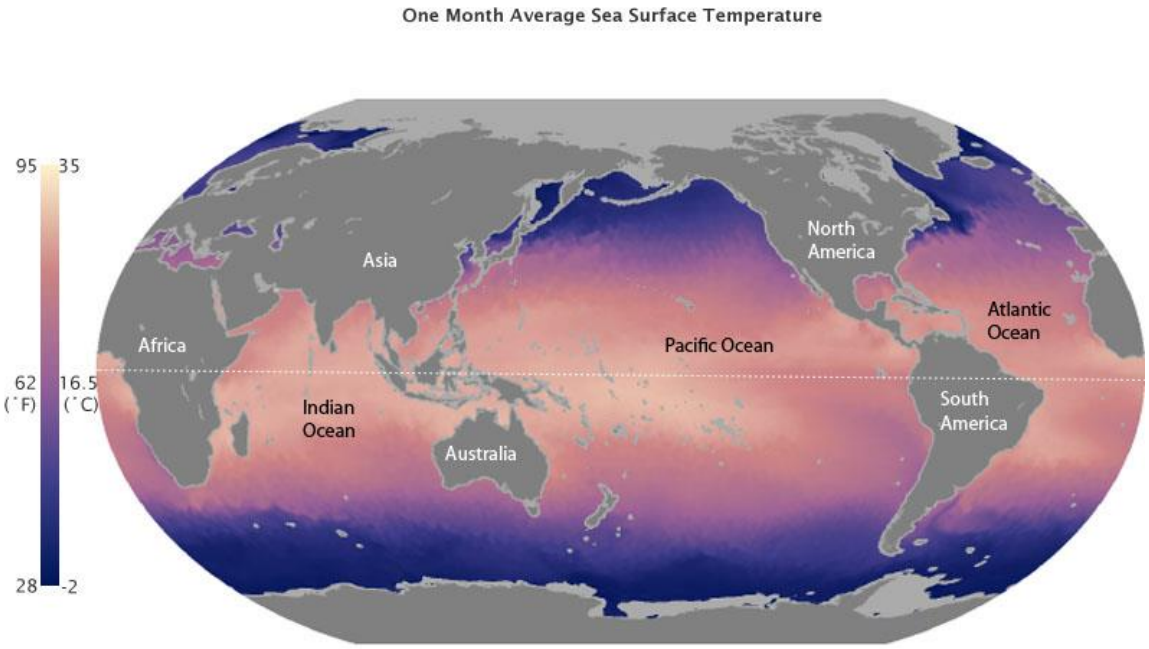


Figure A4. Fully-Scaffolded (FS) SST visualization, including geographic (GS), color (CS), and title and key (TS) scaffolding.

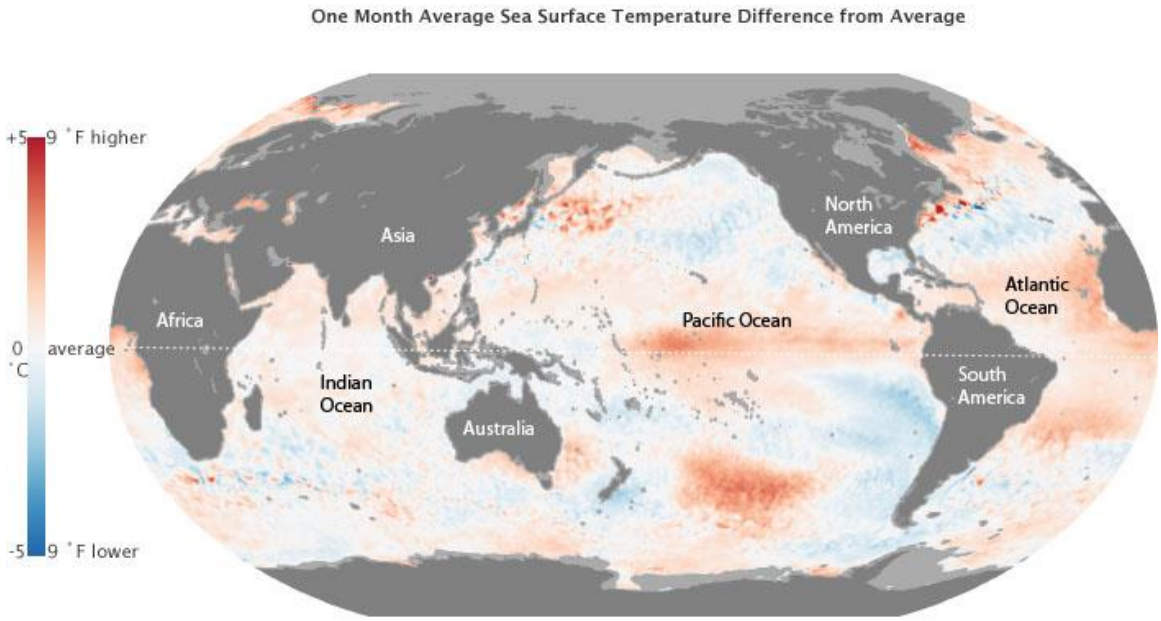


Figure A5. Fully-Scaffolded (FS) SST anomaly visualization.

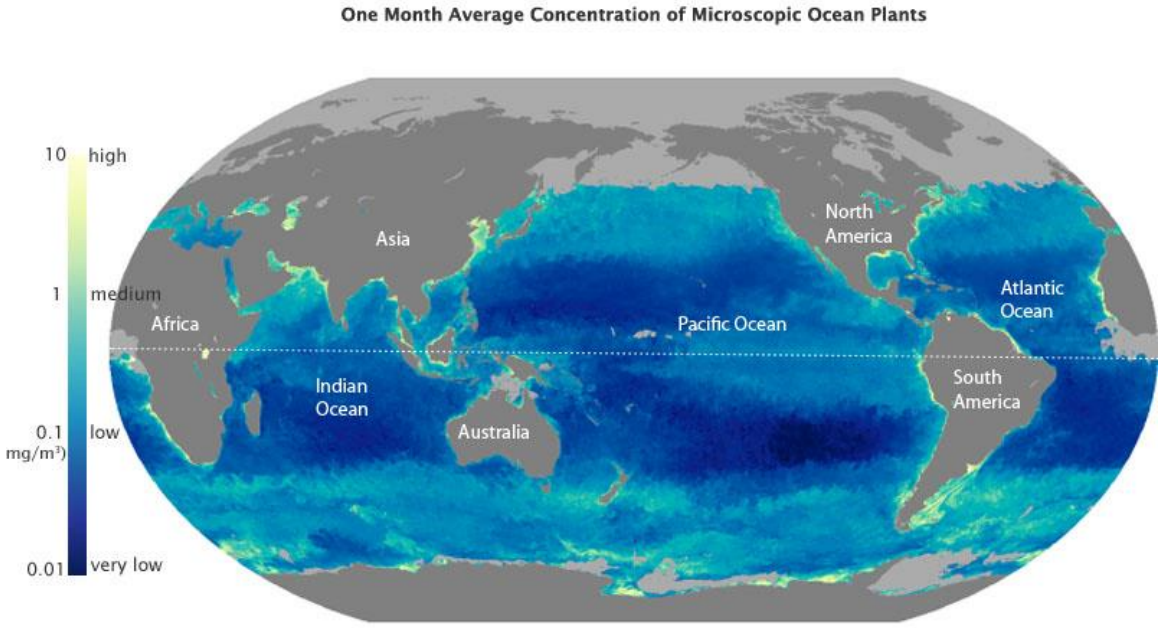


Figure A6. Fully-Scaffolded (FS) chlorophyll visualization.

Appendix 2 - Interview Protocols

Clinical Interviews

Questions to ask one time throughout entire interview, generally on the first stimulus
Where is the equator on the image? (Medium difficulty)
Probe: How do you know?
[Ask this question again for an unscaffolded (US, CS, or TS) image if the first stimulus has the geographic scaffolding (GS, FS)]

What does the grey depict in the image? (Medium difficulty)
Probe: Is there a difference in dark versus light grey?
Probe: How do you know?

Geography
Which coast of which continent has the highest value? (Medium difficulty)
Probe: How do you know?

Colors/values
What do the colors tell you about this image? (Easiest)
Probe: What does that mean?/tell me more
Probe: How do you know?

Where are the most extreme values in the image? (Easiest)
Probe: Is that both high and low extremes?
Probe: How do you know?

What is the measurement unit in this image? (Medium difficulty)
Probe: How do you know?

Main idea
What is the main idea of this visualization?
Probe: What does that mean?/tell me more
Probe: How do you know?

What in the image confirms or goes against that as the main idea? (Medium difficulty)?
Probe: How you know that?

What else would you tell someone this image is about? (Medium difficulty)
Probe: How do you know that?

Time of year/Season
What time period does this image represent? (Medium difficulty)
Probe: How do you know that?

Do you think this is a typical example of that time period? Is this image showing what you would expect to see for this time period? (Medium difficulty)
Probe: How do you know?

(If they know it's less than a year) What season of the year do you think is represented? (Most difficult)
Probe: How do you know?

Equatorial Pacific/El Niño
Describe the conditions in the equatorial Pacific. (Most difficult)
Probe: What could that indicate?
Probe: How do you know?

Is that normal or different from normal in this image? (Most difficult)
Probe: How do you know?

Linear versus Logarithmic Scale
What is the interval between (point to areas and give verbal idea)* (Most difficult)
How does that interval compare to the interval between (point to two more areas) – is it roughly similar or widely different?
Probe: How do you know?

*Intervals:
SST CS/FS scaffolded versions – Interval designed to be 10 degrees Fahrenheit Difference between Peachy red around Hawaii and magenta in Mediterranean Sea compared to Difference between magenta in Mediterranean Sea and dark purple north of Europe

SST unscaffolded (US/GS/TS) – Interval designed to be 10 degrees Fahrenheit Difference between yellow south of Australia and light blue south of Australia compared to Difference between yellow south of Australia and red north of Australia

SST Anomaly Scaffolded (CS/FS) – designed to be 5 degrees Celsius different Difference between Medium blue East of New Zealand and white to the west of New Zealand compared to Dark red east of North America and medium red directly north of it

SST Anomaly unscaffolded – designed to be 5 degrees Celsius different
Difference between yellow under equatorial Pacific and dark blue west of South America compared to light blue edges in mid-North Atlantic medium red west of Greenland

Chlorophyll scaffolded – designed to be factor of 10 different
Difference between dark blue west of South America and medium blue equatorial eastern Pacific ocean basin compared to difference between pale yellow along the Gulf Coast and pale green west coast United States

Chlorophyll unscaffolded – designed to be factor of 10 different
Red East of Asia and orange just adjacent to the east compared to purple west of S. America and darker blue just west of it/surrounding

Eye-tracking Interview

Instructions to participants: “Similar to last time, I’m going to be asking you questions as you look at images. This time, try to answer only the question I ask, and don’t think ahead to other questions I might ask that you remember from last time or from other images as we go along.

First, we are going to calibrate your eyes to the machine. Look at the dot and follow it with your eyes. Try not to move your body and head around.

If you do need to take a break, let me know.

[Calibration] Now, please take a look at this image for 10 seconds before I ask the first question.”

Questions:
What is this image depicting?
Probe: How do you know?

Is there anything in the image specifically that tells you that’s what it’s about?

What do the colors represent?
Probe: How do you know?
What about the grey?
Probe: How do you know?

What is the measurement unit?
Probe: How do you know?

Which coast has the highest values?
Probe: How do you know?

What time period is represented?
Probe: How do you know?

Appendix 3 – Areas of Interest

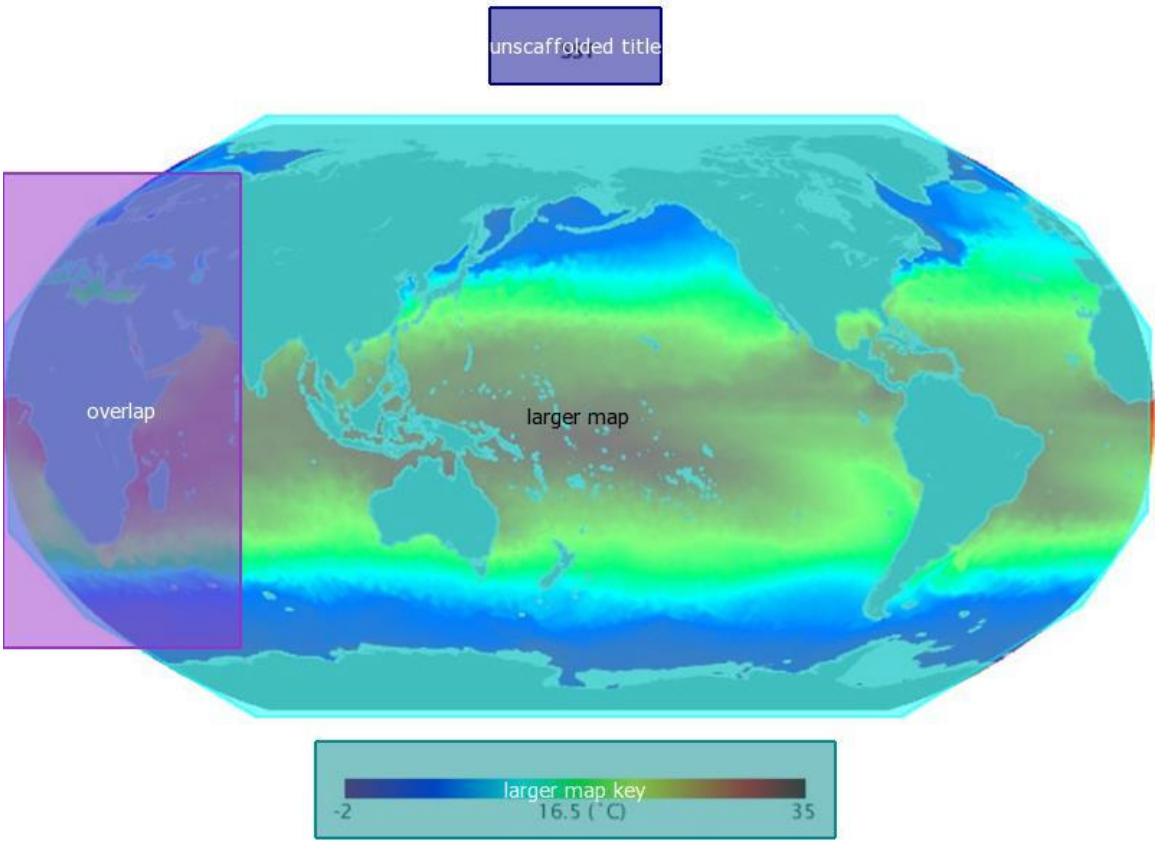


Figure A7. “Larger Map” Areas of Interest. These are the areas of interest used for Cases US, CS, and GS.

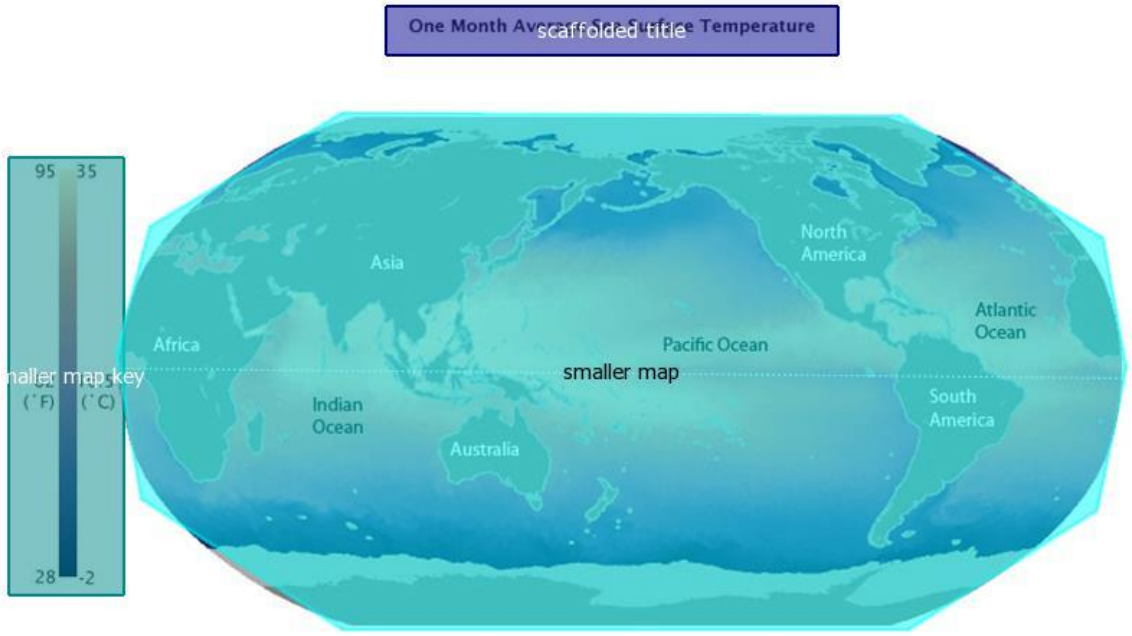


Figure A8. “Smaller Map” Areas of Interest. These are the areas of interest used for Cases TS, FS

Appendix 4 – Interview Scoring Rubrics

The following four tables (Table A1 – A4) present the rubrics used to score the clinical and eye-tracking interviews.

Table A1

SST or One Month Average Sea Surface Temperature Scoring Rubric

Question	Irrelevant/didn't understand Question	Completely Incorrect	Partially Correct	Completely Correct	Sophisticated (must meet conditions of correct as well)
Main Idea			Temperature without reference to ocean	Sea surface temperature	Global OR variation/distribution of
Evidence in visualization for main idea			"the pattern", only oceans colored	Distribution of temperatures, cold at poles and warm at equator	Sunlight, references to specifics about the pattern, upwelling
Colors			Some correct, some incorrect	Pink/peach/white/yellow highest, dark blue/purples lowest	Mention actual values
				Red/yellow high/warm/hot, green middle/cool, blue/purple low/cold	
Time span	Some point in geologic history ("After Pangaea broke up")	Day, snapshot, year, multiple years	Average, A few weeks, a couple months, season	One month/30 days	one January (all clinical interviews, and US in eye-tracking), one April (eye-tracking CS and GS), one July (eye-tracking TS and FS)

Table A1 (Continued)

Question	Irrelevant/didn't understand Question	Completely Incorrect	Partially Correct	Completely Correct	Sophisticated (must meet conditions of correct as well)
Season		East/West Hemisphere; Northern Hemisphere spring, summer, fall; Southern Hemisphere spring, fall, winter	Winter without reference to the hemisphere	Northern Hemisphere winter, Southern Hemisphere summer	one January (all clinical interviews, and US in eye-tracking), one April (eye-tracking CS and GS), one July (eye-tracking TS and FS)
Measurement unit		K, Kelvin, anything else	C (and F, if appropriate)	Celsius (and/or Fahrenheit)	Degrees
Equatorial Pacific conditions		Talk about north/south distribution or Indian or Atlantic ocean basin	Hot/warm	Warmer parts west than east	upwelling
Is Equatorial Pacific Normal?		no		Yes	

Table A2

SST Anomaly or One Month Average Sea Surface Temperature Difference from Average Scoring Rubric

Question	Irrelevant/didn't understand Question	Completely Incorrect	Partially Correct	Completely Correct	Sophisticated
Main Idea		Temperature, no reference to anomaly/difference/ normal	Temperature anomaly/ difference without reference to ocean	Sea surface temperature anomaly, read title	Global OR variation/ distribution of
Evidence in visualization for main idea			Pattern, only the oceans colored		It's not just SST, eddies/currents, El Niño
Colors		No reference to anomaly/difference/ highER-lowER	Some correct, some incorrect	White=average/ zero, reds higher/ warmer than normal, Blues lower/ colder than normal	Mention actual values
				Green=average/ zero, red/yellow/ orange higher/ warmer than normal, blue/ purple lower/ colder than normal	

Table A2 (Continued)

Question	Irrelevant/didn't understand Question	Completely Incorrect	Partially Correct	Completely Correct	Sophisticated
Time span		Day, snapshot, year, multiple years	A few weeks, a couple months, season	One month/30 days	one January (all clinical interviews, and US in eye-tracking), one April (eye-tracking CS and GS), one July (eye-tracking TS and FS)
Season		East/West Hemisphere; Northern Hemisphere spring, summer, fall; Southern Hemisphere spring, fall, winter	Winter without reference to the hemisphere	Northern Hemisphere winter, Southern Hemisphere summer	one January (all clinical interviews, and US in eye-tracking), one April (eye-tracking CS and GS), one July (eye-tracking TS and FS)
Unit		K	C (and F, if appropriate)	Celsius (and/or Fahrenheit)	Degrees, Reference to difference from average

Table A2 (Continued)

Question	Irrelevant/didn't understand Question	Completely Incorrect	Partially Correct	Completely Correct	Sophisticated
Equatorial Pacific conditions		Talk about north/south distribution or Indian or Atlantic ocean basin, hot/warm without reference to average	Hotter/warmer than average, no mention of specific w/E pattern differences	Warmer than average in east, about average in west	El Niño
Is Equatorial Pacific normal?		Normal, different from normal because it's the only area in the image that is that way.		Different from normal	El Niño definition, anomaly definition

Table A3

Chlorophyll-a or One Month Average Ocean Productivity or One Month Average Microscopic Ocean Plant Concentration Scoring Rubric

Question	Irrelevant/didn't understand Question	Completely Incorrect	Partially Correct	Completely Correct	Sophisticated
Main idea			No reference to ocean, concentration or amount of something (unknown)	Ocean Chlorophyll, ocean plant life,	Global OR variation/distribution of, variety of chlorophyll (the 'a' part)
Main idea evidence in map			Pattern, only oceans colored	Highs near coastlines, lows open ocean	upwelling
Colors		green = average, anything as zero	Some correct, some incorrect;	Dark blue =low, white/yellow =high Red high, blue/purple low	Mention actual values
Time span	Some point in geologic history ("After Pangaea broke up")	Day, snapshot, year, multiple years	A few weeks, a couple months, season	One month/30 days	one January (all clinical interviews, and US in eye-tracking), one April (eye-tracking CS and GS), one July (eye-tracking TS and FS)

Table A3 (Continued)

Question	Irrelevant/didn't understand Question	Completely Incorrect	Partially Correct	Completely Correct	Sophisticated
Season		East/West Hemisphere; Northern Hemisphere spring, summer, fall; Southern Hemisphere spring, fall, winter	Winter without reference to the hemisphere	Northern Hemisphere winter, Southern Hemisphere summer	one January (all clinical interviews, and US in eye-tracking), one April (eye-tracking CS and GS), one July (eye-tracking TS and FS)
Unit		Meters cubed or milligrams alone, meters or meters squared	Abbreviations mg/m3	Milligrams per meter cubed	Concentration, density (but it's technically not uniform distribution) – measurement theory
Equatorial Pacific conditions		Talk about north/south distribution or Indian or Atlantic ocean basin	Low or just reference to different, no mention of specific equatorial pattern differences	Lower in (east?), higher In west (?)	upwelling
Is Equatorial Pacific normal?		No		Yes	

Table A4

Topic-independent Questions Scoring Rubric

Question	Irrelevant/didn't understand Q	Completely Incorrect	Partially Correct	Completely Correct	Sophisticated
Grey		Both dark and light are land	Land only, incorrect reasoning for no data (shallow water, etc.), describe only the dark or light	Dark= land AND light= no data, or all=no data	Why it's no data (satellite coverage, ice to some extent, clouds)
Equator location		Gesture too high or low	Don't really gesture across the whole image	Reference to middle of image, correct gesture across entire span of globe	Latitude/longitude reference, specific countries it passes through

Appendix 5 – Qualitative Code Book

The codes I used fell into three themes: *critique*, where participants offered commentary, positive or negative, on the visualizations or their elements; *confusion*, where participants came to incorrect conclusions about meaning or simply indicated they were struggling to make meaning; and *meaning-making*, which were all the types of evidence offered by the participants for their conclusions about meaning, whether those conclusions were correct or not. Codes falling under the *critique* theme mostly concerned the color scale, though some were more general suggestions that the font was too small, or the title was confusing specifically because of jargon or abbreviations it contained. These were the least-frequently occurring codes.

Confusion codes included those for the alternate conceptions of geography or conflation of different color meanings. They also were expressions of difficulty or uncertainty, instances where clarifying questions were asked, or indications that the participant did not know the answer and was unwilling or unable to venture even a guess. The conflation of the terms *typical, average, normal,* and *usual* also were coded under this theme, as well as instances when the participant had expressed that they had used a particular visualization element but subsequently ignored it when trying to make meaning.

The most-frequently coded theme was for *meaning-making*. This included two main sub-themes, *sources of information* and *what they are doing*. The latter covered several codes that expressed how the participants were arriving at their conclusions, such as the ways they divided the color scale into different numerical values, hypotheses they made, and evidence that they were arriving at new insight.

The *sources of information* sub-theme codes were the ones most relied on for the current analysis of meaning-making. These also encompassed two meta-codes for *use of visualization elements,* for instances when participants referred to the title, the key, patterns in the data, or other elements of the stimuli directly,

and *knowledge or experience*, for instances of use of information from outside the experimental conditions directly. Instances where the users compared one stimulus visualization to another from the experiment, either within the clinical or eye-tracking interviews, or when they referenced the clinical interview in the eye-tracking interview were also *use of visualization elements.*

Sources of information from outside the experiment broke down into several groups of codes: *school*, *scientific culture, other culture*, *model representation and construction knowledge.* This meta-code also included over 20 other specific-detail codes that did not fit in one of the groups above, such as *El Niño*, to mark instances where the phenomenon was referred to by name, use of jargon, and discussion of the seasons, as well as general references to prior knowledge or experience that were not specifically attributed to school. The full code book is available upon request.

Appendix 6 – Lists of Jargon Words Used

Table A5

Most Frequently Used Jargon by Participant Subgroup

Novice		Expert	
Jargon	Number of Interviews	Jargon	Number of Interviews
concentration	3	upwelling	15
dense/density	2	anomaly	7
photosynthesis	2	distribution	7
45th parallel	1	gyres	7
currents	1	eddies	6
equinox	1	tongue	6
exponential	1	composite	6
intensity	1	warm pool	5
magnitude	1	austral	5
model	1	margin	5
solstice	1	phytoplankton	5

Note. Words listed in decreasing order of frequency of use.

Table A6

Additional Jargon Used by Experts

Jargon	Number of Interviews	Jargon	Number of Interviews
advected	1	circumpolar	2
aerosols	1	climatological mean	2
algorithm	2	continental shelf	3
altimeter	1	current	3
backscatter	1	cycles/seasonality	2
band	3	density	1
basin	1	dissolved	1
bathymetry	2	drift	1
biomass	2	dynamics	1
bloom	2	electromagnetic radiation	1
boreal	2	extension (Kuroshio)	2
boundary	4	flow	2
channels	1	foam line	1
circulation	3	forcing	1

Table A6 (Continued)

Jargon	Number of Interviews	Jargon	Number of Interviews
front	1	passive measurement	1
gradient	2	photic	1
graininess	1	photosynthesis	1
hydrography	1	pixel	1
in situ	1	platform	1
inertia	1	plumes	2
infer	2	primary productivity[a]	4
insolation	1	projection	1
instability wave	2	regression	2
integration	2	remote sensing	1
interfaces	1	saturated	1
internal wave	1	scatterometer	1
interpolate	1	sediment	2
jets	2	sensor	2
kriging	1	signal	3
latency	1	signature	2
logarithm	2	single track	1
(Aleutian) low	1	smooth	1
magnitude/on the order of	4	solar irradiance	1
meanderings	3	spectrum	1
mesoscale	1	subtropical	2
mixing	1	synoptic	2
model	3	synthetic aperture radar	1
(micro)molar	2	temperate	1
nutrients	4	transient	1
ocean color	1	turbulence	1
organic matter	1	waning	1
parameter	1	warm core	2

Note. Words listed alphabetically.

[a]Used outside of visualizations that had this in the title.

Appendix 7 – Eye-Tracking on a 3-D Digital Exhibit[1]

At the Hatfield Marine Science Center Visitor Center, Oregon State University researchers are engaged in a long-term research program about how visitors make meaning from scientific visualizations of satellite-based Earth system science data (Phipps & Rowe, 2010; Rowe et al., in preparation, 2010). As part of this process, in this dissertation I am testing visualizations with various levels of visual scaffolding, from an unscaffolded visualization that is straight out of an ocean sciences' journal publication to a highly-scaffolded version intended to be clear to people with minimal oceanography background. I have used eye-tracking in a laboratory setting to investigate differences between everyday scientists (novices) and university oceanography professors (experts) in attention to various elements of the visualizations in order to inform my design. Here, I explore the use of eye-tracking with the digital 3-D exhibit piece in its gallery setting.

The Visitor Center is part of Oregon State University, located at its Hatfield Marine Science Center campus in Newport, Oregon, offering a free admission interactive science center experience. It houses hands-on, video, and computer-based science exhibits based around current marine science and technology research by on-site and other university and partner researchers, plus live animal exhibits including touch pools. The facility also offers public talks and lectures around the research.

The exhibit under study here is a digital spherical display system, which projects imagery from a computer via specialized software and a lens in front of the projector which bends the visualization for an internal projection on a globe-shaped screen. The system is the Global Imagination Magic Planet™ three-foot diameter globe. It is located in a small exhibit room containing interactives and information about the remote-sensing systems scientists use to collect much of the data presented in other exhibits throughout the center.

A touch-screen kiosk on one side of the globe allows users to choose among five stories about satellite data. The stories feature a single or series of static images or animations of worldwide ocean data on the globe, always accompanied by text, which in four out of five cases is narrated, and sometimes accompanied by other supporting images on the flat kiosk screen. The static images on the globe rotate while the animations do not. All the stories feature marine themes: Recipe for a Hurricane, Tsunamis, Life in the Ocean (featuring global wind and chlorophyll data on the globe), Coral Bleaching, and El Niño. The Recipe for a Hurricane, El Niño, and Life in the Ocean stories feature visualizations of data similar to that examined in these experiments.

The Free-Choice Learning Lab at the university has been studying how visitors use visualizations in this exhibit in various iterations of both globe content and supporting kiosk material. I have also examined the use of the types of visualizations that might be displayed on the globe in studies in the center but separate from the exhibit. The visualizations designed for this dissertation were based on findings from those and related research work at other science centers with spherical display systems or visualizations of scientific data that accompany current research stories.

Annual visitation to the center is approximately 150,000 visitors, primarily in intergenerational groups of three people on average. Demographically, the center's visitors resemble other science centers in the United States, with the exception of tending to have a higher level of education that in typical of science center visitors, 52% with a four-year or advanced degree versus 46% nationally for similar centers and 28% for the entire population of the United States. Female visitors also outnumber males, 65% to 35%. The exhibit room is located near the end of the typical visitor path through the center, and approximately two-thirds of the center's visitors pass by the exhibit room without entering (L. Good, personal communication, April 28, 2013).

Physical Setup

The Free-Choice Learning Lab purchased an SMI-RED™ eye-tracker integrated with a 22” monitor as described in the body of the dissertation. Our primary interests were flexibility of the system to serve multiple purposes, including the laboratory-based experiments and the *in situ* tracking in the Visitor Center. The SMI system offered powerful stimulus presentation and analysis software as well, plus 120 Hz recording capability. It allows for stand-alone use for presenting large-format posters and videos on a wall, offering us the capability to track visitors as they view versions of posters, videos, and other exhibit text undergoing development and testing. The lab did consider a head-mounted system, but analysis with those systems requires significantly more processing time, and given our inexperience with the methodology, I ultimately opted for a balance of ease of use and flexibility to start.

In the exhibit room, the SMI-RED™ stand-alone eye-tracker was placed on a tripod. After inputting the screen size and selection of participants who would be standing, iViewX™ software recommended position for the eye-tracker height and distance from the screen so that the tracker was 60 -- 80 cm from the visitor’s eyes. The Magic Planet™ sphere was then 169 cm in front of the visitor, behind the eye-tracker, which was set at a height of 151 cm from the floor to the bottom of the stand or foot of the tripod and an initial angle of eight degrees. See Figure 2 in the body of the dissertation for a photograph of the setup.

Due to the constraints of the eye-tracker, participants were only able to view about 15% - 25% of the screen, instead of the whole globe as would normally be available to exhibit users if they chose to walk around. This was considered an acceptable tradeoff as previous work by our lab (Rowe et al., in preparation, 2010) has suggested the availability of 360º use is not necessarily an affordance for visitors, who tend to remain on the side of the exhibit with the kiosk that controls the globe unless explicitly invited via the kiosk to walk around. In this case, since I were interested in meaning-making from visualizations

removed from context, and the tracking system limited the participants' ability to view both the globe and the kiosk at once, participants also were not able to access the touch screen exhibit kiosk; it was both out of the range of readability and physically out of reach. Thus, they were not offered any supporting audiovisual material that normally accompanies the exhibit.

Experimental Conditions

Because I am interested in understanding how visitors view visualizations of data with supporting material presented in titles and color keys, I prepared two versions of visualizations for each of two ocean topics, sea surface temperature anomaly ("SST anomaly"; data used to determine El Niño, La Niña, or neutral conditions), and chlorophyll-a (a proxy for primary producers in the ocean food web). Each participant was presented with one version from each topic; one version that was essentially straight from a scientific publication (*Scientific Level*) and one version (*General Level*) that was altered in colors, title, geographic labels, color key placement, and measurement units intended to be more understandable to a general audience (Phipps and Rowe, 2010). These visualizations were the same as those used in the laboratory experiments for the unscaffolded and fully-scaffolded cases (See Appendix 1). The order of topic and version was alternated, for a total of four combinations (A, B, C, and D) of visualization presentations. See Table A7.

Table A7

In Situ Eye-tracking Stimulus Presentation Order by Level and Topic

Presentation Order	First Image	Second Image
A	Scientific, Chlorophyll	General, SST Anomaly
B	Scientific, SST Anomaly	General, Chlorophyll
C	General, Chlorophyll	Scientific, SST Anomaly
D	General, SST Anomaly	Scientific, Chlorophyll

While normally the title would be displayed on the adjoining kiosk, due to the field of view restrictions presented by the stand-alone eye-tracker, the titles and color keys were displayed on part of the visualization on the sphere that the participant could view without moving around the screen.

Procedure

After obtaining informed consent, visitors' eyes were centered for the eye-tracker as directed by iViewX™ software by moving the subject forward or back or left to right, including if necessary, adjusting the angle of the eye-tracker. Subjects were then instructed to stand as still as possible for calibration and the experiment. Next, calibration proceeded with five calibration points shown on the spherical display; visitors were instructed to look at the center cross first, then I verbally directed the visitor to look at the subsequent points in order while I verified the subjects' gaze on the laptop screen through iViewX™. I first collected data through the iViewX™ software instead of the ExperimentCenter™ stimulus presentation software because the StoryTeller™ proprietary software that displays images on the sphere could not be installed on the computer with ExperimentCenter™. This meant that I had to manually synchronize the start and end of the eye-tracking recording with the stimulus presentation by using the touch screen at the same time as starting recording in iViewX™. In later trials, I used ExperimentCenter™ to time the 10 seconds of spontaneous looking and control data recording but still manually advanced the stimulus image on the touch-screen kiosk.

For the experiment, subjects were instructed to look at the image as long as it was necessary to feel as if they understood it as well as they could. The researcher presented the first image on the spherical screen at the same time as starting the eye-tracker data collection. Upon verbal indication by the subject that they were through looking, the researcher stopped the eye recording. The researcher then asked five questions about the image topic, meaning of the

colors, area of the image with the highest values, time span of data presented, and season of the year presented. During the questions, the subject remained still and the image remained viewable. The researcher recorded answers in field notes.

Once these five questions were answered, the researcher gave the same instructions to view the image as long as necessary for understanding, then displayed the second image while simultaneously starting the eye-tracking through iViewX™. Once the subject verbally indicated completion, the recording was stopped, and the subject was asked the same five questions, though they were allowed to relax as the recording portion of the experiment was finished at this point.

Finally, the subjects were asked about their comfort with interpreting such visualizations as were presented on a Likert-type scale of not at all comfortable, comfortable, or very comfortable, and what school, professional and informal science background they had. These questions are similar to ones asked in the laboratory setting.

Findings

Data were imported into SMI's BeGaze™ software for analysis. Eye-tracking data collected in early trials through iViewX™ was overlaid onto photos taken of the sphere and imported into BeGaze™ after cropping to match the calibration image of 1920 x 1920 pixels. However, despite uploading a calibration image and matching points as described in the manual, the stimulus visualizations did not immediately line up in a meaningful way with the tracking data. Instead, the data seemed confined to the upper-left portion of the stimulus visualization.

In the later trials where ExperimentCenter™ controlled the timing of spontaneous looking sections and data collection, the size of the calibration area which was set to match the visitor's view of the sphere in iViewX™ was actually

overridden by ExperimentCenter™, which thought the calibration and stimulus screen size was that of the laptop running the eye-tracker software. Warping the stimulus images still produced suspicious images as described below, though there were several that were more likely to be correct.

This qualitative view of the data is an important first step to interpreting the quantitative data that the tracker records. For one thing, it allows us to visually inspect the data and determine whether it “makes sense” on a first-pass. For another, it allows us to match fixations to particular locations on the visualization, as I are interested in not only how long (duration) visitors look at the visualization, but also where they look, specifically the title, key, “hot spots” in the map, and geographic labels.

At this point, it is difficult to tell whether or not the eye-tracking data collected in this manner is usable. Of 26 trials, only about half (16) displayed scan paths similar to those found from laboratory experiments, that is, with fixations and paths tracked between those points. See Figure A9.

The path in dark blue represents a “typical” scan path, with fixations and paths drawn between multiple points over 20 s of viewing. The path in lighter blue shows only two paths in 20 s, with fixation sizes too small to make up for the remaining time (larger diameter of spot indicates longer dwell time). In the gaze replay video, however, the gaze is shown moving all around the visualization, casting suspicion on all the path data displayed. This limits the usefulness of any analysis of where participants are looking at the visualizations.

In the later trials, a few of the participants’ data did appear more likely to be a valid match with the stimulus visualization. However, I cannot be certain of the precise alignment due to the unknown algorithm for warping applied by BeGaze™ to the eye-tracking data as compared to the manually cropping algorithm for the visualization by the image manipulation software Microsoft Paint™. In addition, some of the later trials still had long periods where the

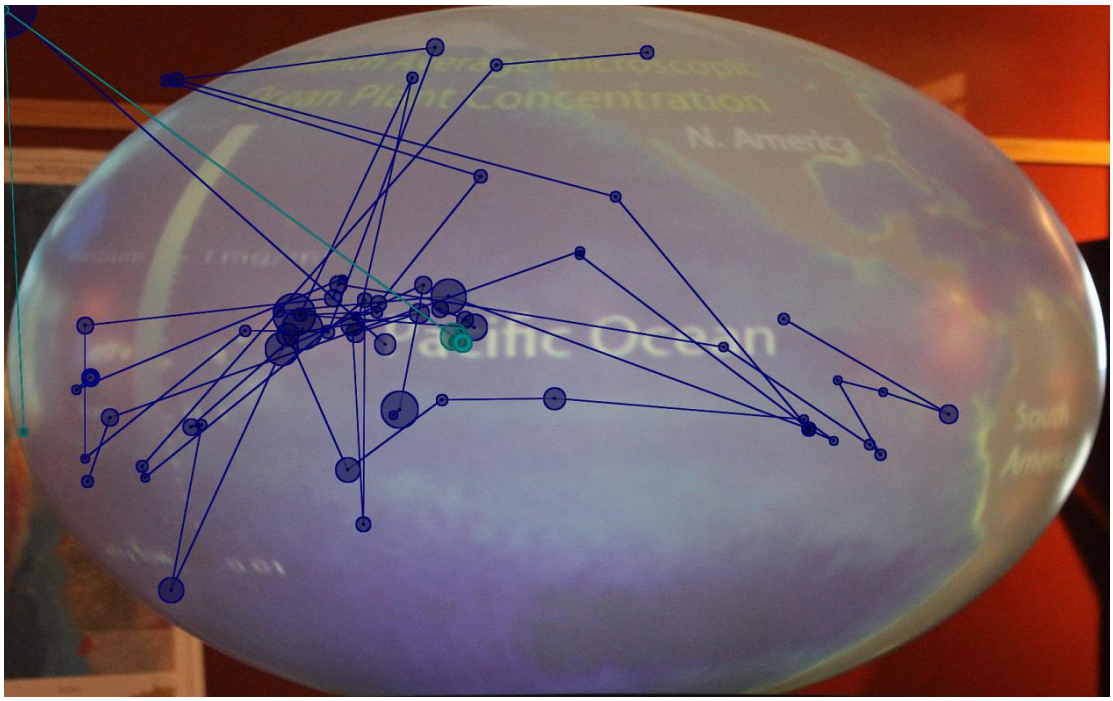


Figure A9. Comparison of typical and suspicious scan paths. The typical scan path is shown in dark blue, while the scan path that seems atypical is shown in lighter blue.

subject appeared to look at the upper-left corner of the image in a single fixation. Discussions with SMI led me to believe this could be a result of overhead lighting behind the participant that tricked the eye-tracker into using it as a source instead of the participants' pupils. Finally, in all cases, participant movement was likely to have caused a significant loss of data, as the tracking ratios for the participants ranged from 7.5% to 82%, meaning that at best, participants were not captured by the system 18% of the time, and at worst, they were not captured nearly 93% of the time.

Lessons Learned

The stand-alone eye-tracker should allow us to collect data in a more naturalistic setting than previous laboratory-based experiments, i.e. an exhibit in a science center. It allowed us to present data at the normal size of the exhibit, under typical less-than-ideal lighting conditions, and in three dimensions in the

noisy, high-distraction context, rather than on a smaller 22”, well-lit flat screen computer monitor in a quiet, isolated room. However, it limited the examination of the full exhibit, including kiosk and 360-degree affordance of the screen that visitors could normally walk entirely around.

Data preparation, collection, and analysis *in situ* is highly time-consuming compared to the laboratory setting, which allows software-driven calibration verification, synchronization of recording and stimulus presentation, and automated overlay of data to stimulus imagery. Another researcher would be handy to help coordinate setup and ensure precision, but this requires more ongoing human resources. The *in situ* data are also harder to interpret due to imperfect matching of the stimulus visualizations to the data paths. It is unclear how many of these problems are due to the digital or spherical nature of this exhibit, but given the results of other projects in this paper, it seems likely the eye-tracking system itself and user inexperience accounts for not insignificant parts of the problem. Subsequent experiments must solve the problem of carefully assuring the calibration image size and the stimulus presentation size assumed by the various pieces of software are set identically. Finally, the problem of the small tracking ratios indicated greater participant movement than is desirable. The system can compensate for movement out of the head box area by retaining and reapplying calibration data when the eyes come back into view, but there is a lag time. Future experiments should consider the use of a chair to help participants remain still.

A glasses-based system for eye-tracking would allow even further naturalistic examination and provide an integrated “participant video” to match with what the participant looked at, but the expense may still not be justified if a research group has existing access to a stand-alone system. However, the object of interest and stand-alone eye-tracking system would ideally have capabilities to perform calibration, present stimuli using presentation software, and capture synchronized participant video using an integrated web camera or similar.