


What Geoscience Experts And Novices Look At, And What They See, When Viewing Data Visualizations

Kim A. Kastens, Lamont-Doherty Earth Observatory of Columbia University, USA;
Education Development Center Oceans of Data Institute, USA
Thomas F. Shipley, Temple University, USA
Alexander P. Boone, University of California, Santa Barbara, USA
Frances Straccia, Education Development Center Oceans of Data Institute, USA

ABSTRACT

This study examines how geoscience experts and novices make meaning from an iconic type of data visualization: shaded relief images of bathymetry and topography. Participants examined, described, and interpreted a global image, two high-resolution seafloor images, and 2 high-resolution continental images, while having their gaze direction eye-tracked and their utterances and gestures videoed. In addition, experts were asked about how they would coach an undergraduate intern on how to interpret this data. Not unexpectedly, all experts were more skillful than any of the novices at describing and explaining what they were seeing. However, the novices showed a wide range of performance. Along the continuum from weakest novice to strongest expert, proficiency developed in the following order: making qualitative observations of salient features, making simple interpretations, making quantitative observations. The eye-tracking analysis examined how the experts and novices invested 20 seconds of unguided exploration, after the image came into view but before the researcher began to ask questions. On the cartographic elements of the images, experts and novices allocated their exploration time differently: experts invested proportionately more fixations on the latitude and longitude axes, while students paid more attention to the color bar. In contrast, within the parts of the image showing the actual geomorphological data, experts and novices on average allocated their attention similarly, attending preferentially to the geologically significant landforms. Combining their spoken responses with their eye-tracking behavior, we conclude that the experts and novices are looking in the same places but “seeing” different things.

Keywords: Geoscience Education Research; Data Skills; Eye-Tracking; Perceptual Learning; Expert-Novice

 Data are the foundation of science. Every insight and every fact in every science textbook is grounded in data. Making meaning from data is a central activity in the life of a scientist. But making inferences from data is far from straightforward. A geoscientist looks at a data-driven visualization (Clark & Wiebe, 2000) and “sees” Earth processes. A student may look at the same data set and see dots, wiggles, blotches of color, or perhaps “pretty rainbows” (Phipps & Rowe, 2010, p. 316). An event or events (*sensu* Shipley, 2008) happened in the Earth System, and left a trace, a physical manifestation in the arrangement of molecules, that was detected by human senses or instrumental sensors. The geoscientist looks at this trace and infers the event. How did the scientist do that? And how can the student learn to do the same?

In recent decades, the geosciences community has amassed vast treasure troves of data documenting how the atmosphere, hydrosphere, cryosphere (ice), and geosphere (solid earth) vary over space and change over time. Government agencies and university-based consortia have assembled data archives covering almost any aspect of the Earth that can be mapped or measured. Pooling data has become a cultural norm in the geosciences community, in part because many geoscience research questions can only be answered with data of multiple types, spanning multiple field areas or multiple time scales, crossing national boundaries, and drawing on multiple subspecialties (Manduca & Kastens, 2012). This approach of gathering, compiling and interpreting global Earth data has led to some of geoscience’s most important advances, such as providing evidence for plate tectonics and global warming, and some

of its most practical products, such as earthquake hazard maps and the tsunami warning system. Some scholars of science predict that all of science is moving in this direction, towards a “Fourth Paradigm” in which data exploration of large, professionally-collected and managed data volumes becomes a primary mode of scientific discovery, alongside the three older paradigms of empirical, theoretical and computational science (Grey, 2009; Hey *et al.*, 2009).

The data in such archives are available to anyone with a high-bandwidth Internet connection. Although the archives were built and are maintained for scientists’ use, students and teachers have also been an intended audience from the earliest days of web-served data (Manduca & Mogk, 2002). Geoscience curriculum developers have responded enthusiastically to this opportunity. There are now hundreds of online lessons plans in support of inquiries in which students work with professionally-collected data obtained from web-served data archives (Science Education Resource Center, 2014). Curriculum developers’ enthusiasm for such inquiries is grounded in one of the fundamental challenges of geoscience education: The Earth is approximately 16-orders of magnitude larger than the classroom, and thus most important Earth phenomena will not fit inside a classroom. Access to archival Earth and environmental data allows students to engage in true scientific inquiry, formulating questions and seeking answers about important Earth processes, even those too big, too far away, too slow, too long-ago, or too dangerous to visit on a field trip.

Although geoscience curriculum development has responded robustly to the availability of Earth data archives, education research lags behind. Swenson and Kastens (2011) categorized research studies on students’ interpretation of data according to the relationship between the data collector and the data interpreter. By far the majority of studies fall in the category of “first hand data,” in which students collect a data set and then interpret the same data. Hug & McNeill (2005, 2008) pioneered a mode of study in which students interpret a data set that they did not collect, but which is nonetheless student-collected data of a type that they could have collected. There is also a growing literature on citizen-science, the situation where students collect data that is subsequently interpreted by professional scientists (Trumbull, *et al.*, 2000; Bonney, *et al.*, 2009). This study deals with the fourth and final category, “professionally collected data,” in which data are collected by scientists, technologists and information specialists, and then interpreted by students. Prior work in this genre includes Edelson, *et al.*, (1996), Hegarty, Canham, and Fabrikant (2010), and Phipps and Rowe (2010).

Interpretation of professionally-collected data differs in important ways from interpretation of student-collected data (Kastens, 2011; Kastens, Krumhansl & Baker, 2015). When people collect data themselves, they gain a deeper understanding of the process by which the data were generated and possible limits on validity (Hug & McNeill, 2008). In the process of making “first inscriptions,” they can develop an embodied, holistic sense of the circumstances or environment from which the data were extracted, and then draw on this understanding in making causal inferences (Roth, 1996), whereas when working from data collected by others, a sense of the data-acquisition process has to come indirectly from metadata and understanding of the instrumentation used. Whereas a student-collected data set typically has a few dozen or perhaps a few hundred data points, professionally collected data sets are measured in megabytes. Students typically display their own data in simple tables or graphs, whereas professional visualization tools enable sophisticated spatial visualization techniques. Student collected data are limited to that which can be collected in a lab or a short field trip, whereas professionally-collected data can be global in scope and span years to millennia in time. If the student-collected data were collected in a controlled experiment, the contrast is even more extreme. In a controlled experiment, the event that caused the observed trace is known--because it was caused by the experimenter. In an observational inquiry based on geoscience data, the causative event is unknown and must be inferred. In summary, it is not safe to extrapolate from research on students’ interpretation of self-collected data to students’ interpretation of professionally-collected geoscience data.

Generating a high-quality interpretation from professionally-collected geoscience data presents many challenges. First, the data interpreter needs some knowledge about the instrumentation used to collect the data, and the relationship between the parameter measured by the sensor (e.g. time for a sound wave to travel from a ship to the seafloor and back) and the Earth attribute of interest (e.g. shape of the seafloor). Next, she needs some knowledge of potential causative processes that may be or have been active in the region under study. In addition, she needs some kind of theory/model about how each of the potential causative processes can leave its trace on the Earth in the form of one or more observable attributes, distributed in a distinctive and diagnostic pattern in time and space. Perhaps she has a mental library of exemplars of the temporal/spatial patterns that are typically caused by common Earth processes, but then how does she decide whether the to-be-interpreted data set is a sufficiently exact match to the exemplar? And

what does she do with data that matches no exemplars? She needs to be aware that the causal processes may or may not be still active and measurable (Kastens & Manduca, 2012), and that traces may have been partially removed, as for example by erosion or subduction. Moreover, traces of multiple causal processes or multiple events may be superimposed on the same section of the Earth (Manduca & Kastens, 2012).

This study approaches the problem of data interpretation by analyzing the gaze behaviors and utterances of geoscience experts and novices as they view a series of five data visualizations. Expert/novice comparison can be an effective starting point for designing educational interventions to help novices move towards more expert-like practices (Singer, et al, 2012; Shipley, Tikoff, Ormand & Manduca, 2013). Participants viewed data visualizations on a computer screen and answered questions posed by a researcher, while an eye-tracking device recorded the point on the screen at which participants were looking. Questions, answers, and participant's gestures were video-recorded.

The data visualizations used in this project are colored shaded relief images of bathymetry and topography (figure 1). The shape of the Earth's surface is one of the most fundamental data sets available to geoscientists, and one of the richest in explanatory power. Solid Earth scientists use topography/bathymetry to identify tectonic plate boundaries, and hydrologists use topography to predict and monitor surface and ground water flow. Physical oceanographers and paleo-oceanographers use bathymetry to identify the gateways and boundaries that steer ocean currents. For atmospheric scientists, topography is a critical boundary condition for explaining the location of tornados and the onset of monsoons.

As data representations go, shaded relief bathymetry/topography is considered highly "iconic" (Myers & Liben, 2008), in other words, the representation resembles its referent. The images resemble a landscape seen from an airplane at low sun angle, and the colors resemble vegetated continents and water-covered oceans. The high iconicity should make this data representation relatively easy for novices to interpret, when compared to representations of Earth attributes that cannot be sensed by the human senses such as strength of the gravity field or magnetic field. And yet an earlier study showed that many 8th, 9th, and 12th grade students currently or previously enrolled in a year-long Earth Science course, were not able to correctly interpret a shaded relief map of global bathymetry/topography (Swenson & Kastens, 2011). The map was projected on a screen in the front of the students' classrooms, and participants were asked to write answers to three open-ended questions: "What do you think this is?"; "How do you think this was made?" and "What do you think this is useful for?" For the first question, all students identified the image as a map, but only 30% mentioned height/depth, shape of Earth's surface or topography. A plurality (44%) of the students mentioned only basic geographic features such as continents or latitude/longitude, without mentioning any Earth science data type. Twenty-one percent mentioned Earth attributes that are not in the image, including tides, currents, ice, weather patterns, winds, clouds, air temperature, water temperature, salinity, terrain types (glacier, tundra, swamp), sediment type, rock type, vegetation, and animal migration patterns. In a subsequent study, Swenson (2010) found similar results among non-science-major undergraduates, using a different, less iconic representation of global bathymetry/topography.

The study presented here was designed to address the following research questions:

- *How do novices and experts invest their attention when exploring a data visualization?* Participants were given 20 seconds to visually explore each visualization, after which they knew they would be asked questions about their observations and interpretations. In eye-tracking studies, the allocation of gaze effort among different areas of the viewed image is taken as evidence of what the viewer considers to be interesting or important.
- *How do experts and novices differ in how they describe what they observe and how they interpret its meaning as they view a data visualization?* We designed a series of questions that would probe for both observations (e.g. "... imagine you are talking ... to a person who can't see the image. Just use words to describe ...") and interpretations (e.g. "How do you think the marked part of the Earth got to be this way, the way you just described...")
- *On the continuum of novices between best and worst data interpreters, can we identify attributes that distinguish the better from the worse?* Previous eye-tracking studies in education contexts have identified gaze allocation behaviors that are clearly counter-productive, such as looking at the text but not the figures in a science textbook (Cromley, Snyder-Hogan, Luciw-Dubas, 2010). If we could identify

- such attributes in data interpreters, it might be possible to design learning experiences to encourage more fruitful approaches to data exploration.
- *Does spatial ability predict the quality of novices' interpretation of topo/bathy data?* As a probe of spatial ability, we used the Hidden Figures task, which has some resemblance to the data interpreter's task of looking for meaningful patterns in the complex visual field of a data visualization.
 - *What are the [self-reported] strategies used by geoscience experts to instruct novices on how to interpret this type of data?* We anticipated that this line of questioning would provide a preliminary look at current instructional practices and might yield promising teaching strategies that could be explored in further research or incorporated into instructional materials.

The discussion section of the paper builds from our observations to (a) hypothesize about the reasons for the observed differences and similarities observed between experts and novices, (b) suggest elements of a learning progression for data interpretation mastery, and (c) explicate implications for instruction.

METHOD

1. Participants

The “novice” sample comprised 45 students drawn from the psychology participant pool at Temple University. An additional 15 students had incomplete or flawed data (incomplete eye-tracking, incomplete video, color-blind, investigator error) and were excluded from the analysis. After giving informed consent and before beginning the interview, novices filled out a demographic form that asked for their sex, handedness, age, race/ethnicity, education level, major, and whether they had ever taken a course in Earth science. The group included 30 females and 15 males. Age ranged from 18-30, with a mean of 20.5 years. Sex and age distributions are representative of the participant pool. Most (26/45 or 58%) reported having taken no Earth Science coursework at all. Seven reported having taken one course containing Earth Science content in college or community college, and an additional 12 recalled studying Earth Science pre-college. No students were majoring in Earth & Environmental Sciences. The most common reported major was psychology (18/45 or 40%), with the rest spread thinly across 18 other majors.

The “expert” sample included geoscientists drawn from the scientific staff of the Lamont-Doherty Earth Observatory of Columbia University. Participants were required to have at least 10 years of professional experience doing research in solid Earth geosciences. Expertise researchers cite ten years as the time investment required to develop expertise in a specialized domain (Simon & Chase, 1973; Feltovich, Prietula & Ericsson, 2006). Participants were recruited by email solicitation and personal contact by author KK, who sought a balance in male/female representation and in oceanic/continental expertise. A total of 14 experts provided useable data. For 2 of these, the eye-tracker recording failed, but since data on experts is scarce, we kept these participants in the analysis pool for the video portion of the study.

Approval for work with human subjects was granted by the Institutional Review Boards of Columbia University and Temple University. All participants granted informed consent in writing after the experiment had been explained. Novices received participation credits required for a psychology course in which they were enrolled. Experts were not compensated.

For anticipated correlational analyses a power analysis indicates that there is a 20% chance of missing effects ranging from $r=.3$ to $r=.66$. The experiment was primarily designed to characterize novice and expert's approaches to professionally collected data and not detect subtle differences among groups.

2. Images

Participants saw five shaded-relief color-coded data visualizations (Figure 1) of terrestrial geomorphology made from a digital elevation model, the Global Multi-resolution Topography Synthesis (Ryan, *et al*, 2006). The “global” image showed the Earth's surface from 77° S to 78°N. The other four images were higher resolution zoomed-in views of areas of geological interest. Data visualizations (images) were made with the *GeoMapApp* system (<http://www.geomapapp.org>) using the default color scheme, shaded relief settings, and projection. A color bar (map key) was added to all five images and a distance scale bar was added to the four high-resolution images.

The global image was provided to help orient the participants to what they were seeing and to build on the earlier work of Swenson and Kastens (2011), which used an almost identical image. The four high-resolution images were selected as classic examples of the geomorphological expression of important Earth processes. We used two images from oceanic and two from continent crust, reasoning that this would provide landforms that were less familiar (oceanic) and more familiar (continental) to the novices. The oceanic images showed a section of the Mid-Atlantic Ridge straddling the Kane Fracture Zone and a seamount-studded portion of the southwest Pacific. The continental images showed a portion of the Valley & Ridge province in Pennsylvania and a portion of the Columbia River with its incised tributary system in Washington/Oregon.

3. Eye-Tracking

Novices viewed the images on a *Tobii T60* eye-tracker at Temple University, and experts viewed the images on a *Tobii model T60XL* eye-tracker at Lamont-Doherty Earth Observatory. These systems have a data rate of 60Hz, i.e. 60 gaze points per second are collected for each eye. Calibration, sequencing of images, and data acquisition were controlled using *Tobii Studio*. Participants sat 60cm away from the monitor.

For all participants, the global image was presented first, followed by the four high-resolution images presented in random order. Preliminary analysis of the quality of the novices' responses to questions about the high-res images showed no systematic relationship to image position in the sequence, and so image sequence was not further considered in the analysis.

As each image came into view, the participant was given 20 seconds of uninterrupted time to scrutinize the image, after which the researcher began to ask questions. The choice of 20s was based on pilot testing, in which we found that novices began to look restless and ready for something else to happen after about 20s of silence. During the Q&A for each image, the researcher controlled the advancing of the images manually, allowing as much time as needed for questions, responses, and follow-ups.

Analysis of the eye-tracking data focused on the 20s interval of uninterrupted exploration time, when the participants were gathering information in preparation for the questions they knew would be coming. We chose this analysis strategy so that we would not have to differentiate between eye-tracking behavior while listening versus speaking, nor account for variation in the length of time that each participant spent discussing each image.

For analysis of the eye-tracking data, each image was divided into "areas of interest" (AOI's), using *Tobii Studio* (Figure 1). AOI's were chosen to separate areas that were either geologically or visually distinct, and were the same for both experts and novices. Discrete AOI's were then clustered into "category AOI's," which encompassed larger areas that were geologically coherent. For example, on the global image, North America, South America, Greenland, Africa-east, and Africa-west were each discrete AOI's, and they were combined with other land areas to form the "Continent" category AOI.

The two basic data types considered were number of fixations spent within each AOI and each category AOI of each image during the 20 s exploration period, and the time spent within each AOI and each category AOI during that same 20 s exploration. For most analyses, both data types were normalized by the percent of the image area that was occupied by that AOI. For these normalized data, the units of data are number of fixations invested per 1% of the image, and msec invested per 1% of the image. Had we not normalized by area, the observed difference between AOI's would be swamped by larger AOI's capturing more attention than smaller AOI's.

4. Think aloud Research Protocol

After the initial 20s period of unguided exploration, the investigator began asking questions. The questions and answers were videotaped, with the camera looking over the participant's shoulder and focused on the screen so as to record the participants' gestures and cursor movements. The audio portion of the videotapes was commercially transcribed, and then annotated by a project team member to indicate gestures or cursor motions. Analysis was primarily based on the transcripts, with occasional reference to the videos to resolve ambiguities or to combine eye-tracking with utterances.

The interview questions were intended to probe both what the respondent was seeing (which we will refer to as “observations”) and how they were making sense of what they were seeing in terms of Earth processes (“interpretations”). The novice and expert protocols were slightly different. An abbreviated version of each protocol is included as Table 1. Both protocols began with a description of the project and its goals, followed by informed consent, and calibration of the eye-tracker. Novices filled out a demographic form following the consent form, and experts filled out a questionnaire about their field of expertise at the conclusion of the interview.

As novices viewed the global image, the researcher asked “What do you think this is?” along with follow-up questions until participant committed to an interpretation, then asked further questions to elicit what in the image had led the participant to that interpretation. With the global map still on the screen, the researcher explained what the map was showing, pointed out the color bar, and then said that the next four images will be “...like this global map, but zoomed-in...” For each of the four hi-res images, the same series of questions were asked. As the participant looked at the whole image, the researcher asked “What do you think this image is showing?”, and “What processes do you think might have shaped this part of the Earth’s surface?” with follow-on questions seeking further elaboration.

Next a white circle or oval was presented on the screen marking a feature of interest, and the participant’s attention was directed to that area. The marked features of interest were as follows: a section of mid-ocean ridge axis, a transform fault, a pair of conical seamounts, a pair of starfish-shaped seamounts, a section of ridge and valley Appalachians, a water gap (river cutting across ridges), and an incised river gorge. The novice was asked to “describe what you see in and around the area that was just marked,” with follow-up questions that were adaptive based on whether the response described geomorphology, a process, a non-geomorphological thing (i.e. the Great Wall of China) or merely described the image itself with no mention of any referent (i.e. I see two brownish parts surrounding a blue stripe.) After all four hi-res images were completed, there were four wrap-up questions probing what sources of information the students drew on as they interpreted the images, how they thought the images were made, and what they thought the images could be useful for.

Our original intent had been to ask the same questions of the experts, adding one question about pedagogy. Initial trials, however, showed that the experts were much more talkative, and the interviews were taking too long and exceeding the storage capacity of the eye-tracker. We therefore omitted the follow-up questions on the global image, the wrap-up questions, and one of the follow-up questions on the high res images. The pedagogical question was asked for each of the high resolution images, and was worded as “Now please pretend that you are coaching an undergraduate summer intern on how to look at images like this, what to pay attention to, how to interpret the data. What would you say?”

Author AB was the interviewer for the novices, and author KK for the experts. AB did not know any of the novices; author KK did know the experts.

5. Spatial Visualization: Hidden Figures

After completing the eye-tracking task, the novices took a test of spatial ability, 16 items from Hidden Figures Test A from the ETS Kit of Factor-Referenced Cognitive Kits (Ekstrom, et al., 1976). For each item, the participant is shown drawings of five simple shapes, along with one complicated drawing. The task is to pick out which of the five simple shapes is contained within the more complicated drawing (Witkin, 1950). This task is considered a test of field dependence/field independence (Witkin, Dyk, Faterson, Goodenough & Karp, 1962; Kirton, 1978) or the ability to attend to relevant information amid visual distractions. Hidden figures was chosen because it bears a face similarity to the task of spotting a significant pattern or shape amid a complex data visualization. The test was scored by assigning one point for a correct selection, zero points for an item left blank, and -0.25 points for a wrong answer to penalize for guessing.

6. Construction of Geoscore from the Spoken Responses

Our goal in analyzing the participants’ utterances and gestures was to quantify how well they observed and interpreted the data visualizations, to investigate whether stronger and weaker data interpreters differed in their approach to visually exploring a new image. The same coding scheme was used for experts and novices.

The coding scheme, summarized in Table 2, aggregates across all of an individual's responses to all questions asked about a given image, searching across all the responses for evidence of ten attributes of understanding per image. The choice and definitions of attributes was developed iteratively. We began with an *a priori* knowledge (author KK) of what Earth features/processes each image was selected to showcase, and then developed a comprehensive catalog of all features/processes mentioned by any expert for each image. From this we selected features/processes that were mentioned by multiple experts, and that pertained to widely known Earth processes (erosion, faulting, plate tectonics) taught in elementary and middle school science. Finally, as the data analysis proceeded, we eliminated a few features/processes that were mentioned by no novices, substituting features/processes that had greater recognizability. The experts' responses also included a variety of more subtle observations and more sophisticated interpretations, but these were not included in our coding scheme.

For the four high-resolution images the coded attributes are as follows: The first four attributes are *quantitative observations*: at any point in discussing the image did the participant describe the trend of a feature, a depth/elevation, vertical relief, or a horizontal length/distance? The next four attributes are *salient geomorphological features*, which were defined individually for each image. For example, on the mid-Atlantic Ridge image, the four coded salient features are the ridge itself, the rift valley, the large offset of the ridge axis (Kane Fracture Zone), and the presence of other off-axis lineaments (fracture zones). Credit was given for lay-language equivalents to the technical terms and for gestures. For example, "this part here [pointing to MAR north of Kane] looks like it should line up with this part here [pointing to MAR south of Kane] but it doesn't" would earn credit for having noticed the "offset" salient feature. The final two attributes were *interpretation* about process or cause. As with the salient features, these were defined individually for each image. For example, on the seamount image, the two coded interpretations are that the conical bumps are caused by volcanism and that the seafloor between the seamounts is relatively flat because there has been sedimentation. The interpretation did not need to be scientifically correct; it merely needed to articulate an Earth process that the respondent suggested could be plausibly responsible for something observed in the image. For example, "this [pointing to MAR north of Kane] moved over here [pointing to MAR south of Kane]" is not a scientifically accurate process interpretation, because in transform fault motion the ridge offset remains constant, but would get credit as an interpretation because it describes a process that plausibly could be causally related to the observed phenomenon of two non-aligned, similar-looking linear features.

For the global image, there were again 10 coded attributes. Almost no one, even among the experts, described quantitative attributes such as the height, depth, or size of features on the global image, so we substituted the following for the first four attributes: Latitude/longitude, elevation, bathymetry, and the meaning of color.

For analysis, we constructed two composite scores. The four-image *GeoScore* sums all ten attributes across all four high-resolution images, of a total possible score of 40. The five-image *GeoScore* sums all ten attributes across all five images, including the Global image, out of a total possible score of 50. For comparison between experts and novices, the four-image *GeoScore* is used, because the expert interviews passed quickly over the Global image to conserve eye-tracking recording time. The five-image *GeoScore* provides some additional nuance to analyses within the novice group.

For each image, author FS extracted from each transcript the text/gesture snippets that she judged to be definitely or possibly relevant to each of the 10 attributes, and placed them in a spreadsheet. She did a first coding for the presence or absence of each attribute for each participant for each image, noting those that seemed potentially problematic. Author KK did a second coding based on the snippets selected by FS, and the two coders reconciled differences through discussion, clarification of the description of the attribute, and in a few cases by development of a replacement attribute. When all coding was complete, an independent third coder re-coded 10% of the responses (350 responses), spanning all images, and including experts, strong novices and weak novices. The third coder had experience in Earth Science teaching and in qualitative education research, and was otherwise uninvolved with the project. Inter-rater consistency between the independent coder and the consensus FS/KK codes was 95.4%.

7. Coding of How to Teach Question

The experts were asked one pedagogical question: "Now please pretend that you are coaching an undergraduate summer intern on how to look at images like this, what to pay attention to, how to interpret the data. What would you

say?” The question was repeated for each of the four high-resolution images. The wording of the question was intended to evoke a situation of one-to-one mentoring, with the respondent and a bright and eager-to-learn apprentice looking at the data visualization together. This question addressed the geoscientist-respondents in their role as education practitioners. Asking practitioners to reflect on their professional practice is an established methodology in curriculum development where the learning goals involve expert-level proficiency (e.g., Norton, 1992; Malyn-Smith & Ippolito, 2015). As always, self-reflections should be viewed with care, as they can be incomplete or distorted (Feldon, 2007).

The coding schema used for this question is shown in Table 3. The construct embodied in the coding scheme is a continuum from specificity to generalizability. At the one extreme (category 1) are guidance utterances that are specific to the image being viewed. At the other extreme (category 4) are guidances that could be applicable to any type of data visualization. In between, are guidances that would be applicable to any visualization of bathymetry/topography data (category 2), and guidances that would be applicable to any type of map (category 3). The four categories and the specific-generalizable dimensionality of the coding scheme were developed *a priori*. The subcategories (e.g. 2a elevation) were developed bottom-up based on the participants’ responses.

The granularity of our coding of this question was the “guidance utterance” or “guidance”: a coherent phrase that offered a specific piece of advice or guidance to the imagined student. Such utterances ranged from less than a sentence to several sentences or sentence fragments. We tallied guidances in two ways: (a) by the number of guidances in each coding category for each image for each expert, and (b) by the presence or absence of at least one guidance in each coding category for each image for each expert.

Responses from 14 experts were coded; this includes two experts for whom we had good video but unusable eye-tracking data. Every response was coded twice, by authors KK and FS. Differences were reconciled by discussion.

RESULTS

1. Spatial Ability: Hidden Figures

Novices’ scores on the Hidden Figure test ranged from -3.75 to 11.0, with a mean of 2.49, out of a possible range from -4.0 to +16.0. These are comparable to scores observed in other samples from this same participant pool (unpublished data).

2. Total GeoScore

The mean (SD)total 4-image *GeoScore* among the novices was 8.9 (5.2) out of 40 possible, as contrasted with a mean of 31.9(4.3) among the experts. There was no overlap between the two groups: the lowest scoring expert was 3 points stronger than the highest scoring novice (Figure 2). Recall that the coded attributes were chosen to be ones that a novice had a chance of answering correctly. Had we chosen more arcane attributes that were mentioned only by experts the expert/novice separation would have been larger.

Among the novices, *GeoScores* varied by more than an order of magnitude, ranging from 0 to 20 (out of 40 possible) for the 4-image score and from 2 to 25 (out of 50 possible) for the 5-image score. Strong *GeoScore* performance is associated with prior Earth Science course-taking. When asked on the demographic form “Have you ever taken a course in Earth science? If so, please explain,” 26 participants said no and 19 said yes. The students who reported taking no Earth Science had a mean 4-image *GeoScore* of only 6.7 (4.1), whereas the students who reported taking Earth Science had a mean 4-image *GeoScore* of 12.0 (5.0), a statistically significant difference (two-tailed *t*-test, $p=0.0005$). A similar pattern was seen with the 5-image *GeoScore*, although not quite as strong [$M=9.9$ (5.5) for no ES course versus $M=15.9$ (6.3) for ES course takers, $p=0.002$].

We examined several other aspects of our dataset for factors that might be potentially contributory towards the wide range of individual differences in *GeoScore*, and found no candidates. Novices’ score on the Hidden Figures test showed a very weak positive correlation with 4- and 5-image *GeoScore* ($r^2<0.02$), and a negative correlation with age and year of college. There was no significant difference between the *GeoScores* of male and female students (*t*-test, $p=0.4$ for 4-image *GeoScore* and 0.7 for 5-image).

3. Components of GeoScore

Table 4 and Figure 3 show the breakdown of *GeoScore* into its component parts: description of *salient* geomorphological features, attempted *interpretation* of the processes or mechanism responsible for the salient observed features, and use of precise and *quantitative* metrics (elevation/depth, relief, distance/length, and trend) in describing the data.

As is to be expected, the experts outperformed the novices on all components of the *GeoScore*. However, the gap between experts and novices differs markedly on the three components. On the salient features component, the novices' mean score was 39.9% of the experts' mean score, whereas on the quantitative component, the novices' mean score was only 12.3% of the experts' mean (Table 4). The Interpretation component falls in between, with novices' mean equalling 30.1% of the experts' mean.

Among the novices, there is a trend in the prominence of each *GeoScore* component within the individual's score. The weakest-performing novices (ranked by total *GeoScore*, bottom of Figure 3), get almost all of their points from describing the salient features that are visually available in the data visualization in front of them. The intermediate-performing novices pick up more and more points from the interpretation components, and then the high-performing novices begin to pick up points from the quantitative components. The experts' point distribution across the three components closely follows the distribution of available points (Figure 3, top bar).

4. Eye-Tracking:

Eye-tracking areas of interest (AOI) data can be presented as either number of fixations spent within each AOI or as the time spent within each AOI. For our dataset, these two measures are very strongly correlated (Figure 4), with $R^2=0.96$ for the experts and $R^2=0.98$ for the novices. Figure 4 shows the relationship between time and number of fixations per AOI for experts and novices with separate symbols and with separate regression lines. Note that the experts and novices behave almost identically with respect to the relationships between time and number of fixations per AOI, as indicated by the near-superposition of the expert and novice regression lines. This need not have been the case: to the extent that fixation duration reflects information processing, it might have been the case that experts or novices tended towards longer or shorter fixation duration for parts of the image that they were more or less interested in, in which case the two regression lines of Figure 4 would have diverged.

Given near perfect correlation between time and number of fixations spent per AOI, and the near identical behavior of experts and novices with respect to fixation duration, for the rest of the paper we will present data only in terms of number of fixations per AOI. Recall from "Methods" that we normalized the AOI data by dividing the number of fixations in an AOI by the area of that AOI as a percentage of the total image area, so these data are in units of number of fixations per 1% of the image area. The average number of fixations recorded per image across the 20s viewing interval was 56.1 for novices and 56.9 for experts, so that if attention had been evenly distributed across the image, each 1% of the image would have received 56/100 fixations, or 0.56 fixations. In fact, attention was quite unevenly distributed across the image. Many AOI's received almost no attention, signified by the large number of data points near the origin of Figure 4, while some AOI's received an order of magnitude more attention, signified by data points around 5 or 6 on the horizontal axis of Figure 4.

Next we examine which AOI's and types of AOI's commanded the most and least attention, and compare the distribution of allocation of attention of experts versus novices. For this analysis, we use the category AOI's, which lump together individual AOI's into areas that have similar visual appearance and similar histories. Figures 5 and 6 are scatterplots on which the mean number of fixations by experts in a category AOI is plotted against the mean number of fixations in that same AOI by novices. In each case, the number of fixations is normalized by dividing by the area of the AOI. On such a plot, points that lie above the diagonal 1:1 line represent AOI's to which novices attended more than did experts, whereas points below the 1:1 line received more attention from experts than from novices.

Consider first the AOI's for the cartographic aspects of the image (Figure 5). For the color bars (aka "map key" or "map legend") on each of the four high-resolution images, the novices invested far more looking effort than did the

experts, signified by the marked location of these symbols substantially above the 1:1 line. Conversely, experts invested more attention towards the latitude and longitude axes than did the novices, depicted by the position of those symbols below the 1:1 line. Results for distance scale were mixed.

Now consider the category AOI's that comprise the actual data, which we will call Geomorphology AOI's. The expert and novice mean number of fixations per Geomorphology AOI are plotted in Figure 6. A linear regression line has been fitted through the points showing that the attention that experts and novices invest in the geomorphological AOI's is strongly correlated ($r^2 = 0.89$); in other words, experts and novices are behaving similarly in this regard. As shown by the labeled data points falling farthest off the 1:1 line, experts pay a bit more attention to Ridge-transform intersections and to the oceanic sections of the Global image, and the novices pay a bit more attention to Seamounts and the continental portion of the Global image. The same data are shown numerically in Table 5. In this table, the AOI's are sorted by normalized mean number of fixations within each sample group, so that the AOI to which experts devoted the most attention is at the top of the expert column, and similarly the AOI to which the novices devoted the most attention is at the top of the novice column. The order in the two columns is nearly identical. Both novices and experts devoted the most time to the ridge-transform intersections on the Mid-Atlantic Ridge images, followed by the Seamounts on the Seamount image, the canyon on the River image, the Valley & Ridge province of the VR image. Both devoted the least to the off-axis portions of the Mid-Atlantic Ridge image (Flank and Fossil FZ AOI's). The only break in the expert/novice agreement of order lies in the Global image: compared to the experts, the novices devoted more attention to continents and less attention to oceans.

5. Combining GeoScore with Eye-Tracking

We have shown that there is a wide range of *GeoScore* within the novice sample, and a handful of ways in which novices and experts differ in their eye-tracking behavior, notably in tendency to focus on various cartographic devices on the map and their relative attention to oceanic and continental portions of the Global image. We now ask whether high-*GeoScore* novices act more like experts in their eye-tracking behavior than do low-*GeoScore* novices, focusing our attention in the category AOI's that showed an expert/novice difference in Figures 5 and 6.

Figure 7 (upper) shows individual participants' *GeoScores* plotted versus the attention that they paid to the continental portions of the Global image. Because individuals had different numbers of fixations in any given image, attention is expressed as the ratio between the number of fixations an individual invested in the continental AOI's and that individual's total number of fixations on the image. One sees substantial individual differences within each group and substantial overlap between the groups in the degree of attention they pay to the continents. On average, the experts devoted 46.2% of their fixations to the continents, while the novices devoted 69.2%. The experts' fraction of attention devoted to the continents is quite close to the percentage of the image falling in continental AOI's, which is 48.3%. The novices, in contrast, overinvested in continents.

Within the Continent category AOI, novices had a favorite continent. Figure 7 (lower) shows participants' attention to the single AOI that covered North America. The North America AOI covers 7.3% of the image, and the experts allocated approximately that fraction of their fixations (mean 6.9%). The novices, in contrast, allocated almost twice that percentage (13.3%). Although there is a lot of scatter in the data, there is a notable tendency for the high-*GeoScore* novices to behave in a more-expert-like way by allocating less attention to North America than did their low-*GeoScore* peers (see trendline figure 7, $R^2=0.13$). No such trend was seen for Continents as a whole.

Figure 8 shows the allocation of individuals' attention to color bars (upper graph) and the latitude and longitude axes (lower graph). Color bar is quite a small AOI, only 2.3-3.3% of the images. Almost everyone gave it more than that share of their fixations, as is appropriate given its importance in map reading. Among the novices, there is large variation in attention to the color bar, with the high end of the range being two individuals who spent more than 20% of their fixations on the color bars. Most of the novices paid little attention to the latitude and longitude axes, with the low end of the range being two individuals who had not a single fixation in and latitude or longitude AOI on any of the five images. There is no trend of number of fixations versus *GeoScore* on either Color bar or Lat/Long (Figure 8).

6. Experts' Teaching Strategies (Pedagogical Question)

Across all four high-res images, the 14 experts offered a total of 338 instances of guidance or advice to the imagined summer intern. Figure 9 shows the relative abundance of the various types of guidance, according to the coding scheme of Table 3. The data are displayed in two ways. The upper graph allows for multiple instances of a code within an individual's response to an image, and shows what percentage of all of the 338 recorded guidances fall into that coding category. The lower graph notes the presence or absence of at least one occurrence of that coding category in each response. There were 54 opportunities for each code to manifest (4 images x 14 experts – 2 opportunities missed through investigator error), and the graph shows the percentage of guidances observed in that category relative to maximum of 54 opportunities.

Both data displays show that our experts provided more guidances of the more specific types (categories 1 and 2) and fewer guidances of the more generalized types (categories 3 and 4). Of all of the guidances offered, nearly 50% were in category 1 (image-specific details), and almost every expert made a category 1 utterance for almost every image (96% of opportunities, or all but two). In category 1, the expert indicates by word or gesture a specific feature on the image and may state what it is called (“these are the abyssal hills”; “I guess this is a propagating rift of some sort”) but does not explain his/her reasoning.

Two types of guidance useful across topographic/bathymetric data were next most common: 2a (elevation) and 2b (rules of thumb about geomorphology). In category 2a, the expert directs the learner to attend to elevation or relief using the provided scale. Category 2b differs from Category 1 in that the expert does not just indicate and name a feature, but also articulates some observable aspect of the feature that is diagnostic to the geomorphological eye, some aspect that could be used to identify another instance of the same kind of feature, such as linear or dendritic pattern. Among the guidances generalizable across all maps (category 3) or all data (category 4), the most commonly offered was to use the available information (scale, latitude) to get a handle on sizes and distances (category 3a). A number of additional useful strategies were offered in small numbers: figure out where you are in the world from the lat/long (3b), use the key or legend to figure out what type of representation you are looking at (4a), and access additional information sources (regional map, other data types, scientific papers about the field area) to provide context (4b).

DISCUSSION

1. Expert/Novice Similarities

The big surprise in this data is that the experts and novices behaved so similarly in how they allocated their attention during their exploration of the topography/bathymetry data. Although there were notable expert/novice differences in their attention to the cartographic devices (lat/long axes, color bar, and distance scale), their attention to the major geomorphologic provinces is remarkably similar (Table 5, Figure 6). Kellman & Massey's (2013, their table 4.1) articulate characteristics of information extraction by experts and novices; in their schema, the least demanding aspect of information extraction is “selectivity.” Selectivity refers to the tendency of novices to pay attention to both irrelevant and relevant information, while experts engage in selective pickup of relevant information. For the specific task of observing geomorphology data visualizations, novice and experts seem to be behaving similarly at the level of “selectivity.”

Topping the list of which category AOI's commanded the attention of both experts and novices (Table 5) are the most geologically significant regions of each of the high-resolution images: the ridge crest and transform areas of the MAR image, the seamounts on the seamount image, the dissected canyon area of the river image, and the area of eroded tight folds of the Valley & Ridge image. Within 20s of seeing the images for the first time, the novices, like the experts, have identified where the geologic action has been happening in each high resolution image and have focused their attention towards those parts of each image.

Evolutionary psychology provides a possibly valuable lens through which to consider the novices' relatively expert-like behavior with respect to attention-allocation across geomorphological provinces. This school of thought makes a distinction between cognitive processes that were critical or adaptive for survival and reproduction in the lives of our evolutionary ancestors, and those that were not. The former may be more deep-seated and possibly hard-wired in

modern humans, while the latter have to be learned through painstaking education and experience (Cosmides & Tooby, 1994; Evans, 2003; Geary, 2012a). Hunter-gatherers would have benefitted from awareness of landforms in their environment. Landforms control watercourses, determine easy and difficult walking routes, shape microenvironments in which specific plants and animals thrive, and are spatially associated with natural hazards such as landslides and floods. We can posit a type of “folk geomorphology,” alongside the “folk physics,” “folk biology,” and “folk psychology” (Solomon & Zaitchik, 2012; Geary, 2012b), which would cause humans to attend to distinctive or unusual landforms, which are often those shaped by active or strong geological processes.

The developer of GeoMapApp, William F. Haxby, explicitly designed the data visualizations schema of GeoMapApp and its predecessor software to mimic the appearance of a natural landscape as seen from a high vantage point. He combined a naturalistic color palette and with shaded relief of a simulated low sun-angle sun, rather than using the topographic contours that were the standard approach to showing topographic data when he began his work. His stated intent in this design was to tap into human’s intuitive perceptual and interpretive ability rather than relying solely on analytical approaches to data interpretation (Haxby, et al, 1983; Haxby, personal communication). Our results suggest that he succeeded.

2. Expert/Novice Differences

In this dataset, the expert/novice differences in ability to articulate observations and interpretations about what is being seen (GeoScore) greatly outstrips the expert/novice difference in gaze allocation behavior during visual exploration of the image (eye-tracking). There is zero overlap in the two groups’ GeoScores (Figure 2). For the eye-tracking, even in the metrics where we did find expert/novice differences, there is substantial overlap between the two groups (Figures 8 and 9).

The areas where we detected expert/novice differences in eye-tracking make sense. The finding that novices attend preferentially to continents in the global image aligns with Swenson & Kastens’ (2011) finding that high school students shown this same image and asked to write about “What do you think this is?” wrote mostly about the continents. Swenson & Kastens interpreted this as due to students looking at what was familiar to them. Our additional finding that within the continents they focused preferentially on North America suggests that they were drawn to the locality with which they had a personal connection. This interpretation is compatible with international comparison studies in which high school and college students, when asked to sketch a map of the world, tended to draw their home country and continent larger in size and more central in location relative to other features (Saarinen, 1973).

Another possible interpretation, compatible with continent-preference but perhaps not with America-preference, would be that novices are seeing the continents as figure and the oceans as ground. Gestalt psychology posits that humans visual processing system tends to classify views into “figure,” more deserving of attention, and “ground,” ignorable. Relatively small size and convex shape are cues for figure, while appearing to be behind or in the background is a cue for ground (Rubin, 2001). On the shaded relief image viewed by our participants, continents are convex, smaller in area than oceans, and appear to project outwards towards the viewer.

In their responses to the coaching-an-intern question, experts referred to the latitude/longitude grids on the high-res images as a source of two different types of information: about position in the world, and about horizontal map scale. The novices’ paucity of fixations in the latitude and longitude AOI’s suggests that many or most of them lack the geographic knowledge to make use of this information, by converting latitude to distance, or by using latitude/longitude to position the image relative to global landmarks.

On average, novices attended to the color bar more than did experts. The novice mean is pulled upwards by some individuals who invested impressively much attention to the color bar (more than 15% of their fixations on an AOI that occupies 2.3% of the image and is far from the center of the image). Looking at the gaze plots or replaying the eye-tracking sequence for these individuals shows that the gaze kept returning to and dwelling on the color bar, as though the student had been taught that the map key is important and kept going back there seeking guidance. Experts, in contrast, tended to glance at the color bar once per image and then move on.

In addition to difference in overall GeoScore, we also saw expert/novice differences in the components of GeoScore. Of the GeoScore components, experts and novices were most similar in their ability to articulate the qualitative aspects of the salient features in each image, allowing the novices to use lay language rough equivalents of the experts' technical descriptions. The strongest expert/novice GeoScore contrast lay in their use of quantitative elements (relief, altitude/depth, distance/length, trend) in describing the image.

3. Perceptual Learning

When the experts were asked how they would coach an undergraduate in how to interpret this type of image, the largest number of guidances offered took the form of drawing attention to and naming specific features or areas on the map. On the face of it, this seems like a tremendously inefficient pedagogical approach, as taken to scale it would imply that each learner will need to be shown each individual Earth feature or type of Earth feature in each type of data that he or she will encounter. It is tempting to say that these experts are just bad teachers, and they need a healthy dose of professional development to help them develop ways of giving higher-added-value guidance, of offering strategies that will be transferable across to other data visualizations and other data types.

On second thought, however, there may be something more going on here. The approach the experts are describing is the process by which we all learned our first language, one of the greatest feats of learning that each of us has ever accomplished. An adult seeking to help a child learn to speak points out features in the environment and states what they are called in the target language: "See the doggie." Learning language is not just about learning to mimic adults' sounds; it is learning how to partition observed reality into different phenomena that we shall agree to call by different words. And yet, the adult does not state the criteria for determining what label to use: "See, it is moving so it is an animal, and it has four legs and fur so it is a mammal, and it is attached to a human by a leash, so it is probably a dog." The adult just calls attention to the phenomenon and states its name. Likewise, our experts often did not state their criteria; they just called attention to the phenomenon and stated its name.

This kind of learning has been studied by cognitive scientists, beginning with Gibson and Gibson (1955), under the name of "perceptual learning," defined as "experience-induced changes in the way perceivers pick up information" (Kellman & Massey, 2013, p119). The idea is that perceptual mechanisms do not just deliver low-level information such as color and shape to be processed by higher cognitive functions. Rather, with experience, perception alters and provides more complex, differentiated, and abstract descriptions of reality. Thus, with experience, the toddler learns to perceive a dog, without any conscious effort of classification or criteria-matching. Perception is faster and more energy efficient than cognition, and thus it is postulated to have been evolutionarily advantageous for our ancestors to accomplish as much as possible of the huge task of making sense of one's surroundings through perceptual rather than cognitive mechanisms. It is a characteristic of perceptual learning that when people have reached fluency, they "see" things as falling into the correct classification. One of our experts expressed this feeling: "From a geology background you can just take one look at this and look at the shape and the morphology and go 'this is obviously a mid-ocean ridge.'" Experts perceive their classification as a percept, not as the outcome of inference or analysis, and in some cases they may not even be able to articulate the reasons for their classification (Biederman & Shiffrar, 1987).

So it may be that our experts are telegraphing to us that they are drawing on their own perceptual learning as they interpret the images, and they are trying to trigger perceptual learning in the imagined undergraduate by pointing to important features and stating the name: "See the doggie" becomes "See the transform fault." If so, it may be that perceptual learning plays an underappreciated role in mastery of data interpretation, as least as applied to data presented in the form of visualizations. If this is true, then the research literature on perceptual cognition provides some insights about how to scaffold and foster perceptual learning around data visualizations. Perceptual learning is accelerated if the learner is provided with multiple trial instances of a target phenomenon, and the instances have a high degree of similarity (providing the commonality that justifies placing them in the same category) and yet at the same time have differences that span the full range of variability within the category (Kellman & Massey, 2013). Perceptual learning is also accelerated when feedback is provided about whether the learner's determination is right or wrong, when the diagnostic features are pointed out to the learner (Biederman & Shiffrar, 1987), and when contrasting cases are juxtaposed (Le, Silver, Shemwell, Capps & Voyer, 2015). For certain high-stakes learning tasks in medical imagery, perceptual learning software has been developed to provide the necessary multiple trials, with feedback, across instances that present both similarities and differences (Kellman & Massey, 2013).

4. Toward a Learning Progression for Making Meaning from Data

Based on our own experience as data interpreters and our reading of the literature, we envision that three types of knowledge/information must be accessed and integrated in order to make meaning from a dataset one has not previously been exposed to. These are:

- a) *Data-derived information*: Information that is contained within the data itself. In the current study, this is information that is visually available to the data analyst as the meaning is being constructed. In the broader context of science and science education, it would include information throughout a dataset that one is manipulating and analyzing.
- b) *Knowledge about the referent*: Knowledge about and understandings of structures, processes, behaviors and functions of the system represented by the data (the referent system). “Meaning making” in this study refers to constructing and articulating understandings about the referent system, and thus some foundation of knowledge about the system is needed upon which to construct the new understandings.
- c) *Representational competence*: Also known as “meta-representational competence” (Hegarty, 2014; diSessa, 2004). Proficiency with data representations in general and the available type of data representation in particular. This competency includes the ability to use never-before-seen external representations productively.

The mechanism by which the brain combines, integrates, and processes these three types of information to generate an insight about the referent system remains deeply mysterious. This mechanism has not been addressed by the current work. What we can address is availability of the three types of input knowledge among our novices as compared to our experts.

The attributes of our respondents’ protocols that we coded as “salient geomorphological features” are manifestations of the first type of knowledge needed for data interpretation: information that is contained within the data itself. The attributes that we coded as “Interpretation” draw heavily upon the second type of knowledge: knowledge about the processes and behaviors of the Earth System, such as erosion, faulting, and volcanism.

We assert that the attributes coded as “quantitative” are manifestations of the third type of knowledge: the aspect of representational competence that allows one to use novel external representations productively. The experts in our study had not seen these exact high-resolution data visualizations previously. And yet they knew that a good thing to do when trying to describe or interpret a novel data representation is to characterize the highs and lows and typical values, the range of values, the size or extent of features, and any prominent trends or patterns observed. The quantitative attribute we coded as “elevation” is a manifestation of the respondent characterizing the highs, lows, and typical values in the data (e.g. the elevation of the coastal plain in the Valley and Ridge image, or the ridge crest in the MAR image). The attribute we coded as “relief” is a manifestation of the respondent characterizing a range of values (e.g. the difference in depth between the base and peaks of the seamounts). The attribute we coded as “distance/length” appears when the respondent is attending to the size of features in the horizontal dimension (e.g. the width of the MAR rift valley). The attribute we coded as “trend” appears when the respondent has noted a prominent directionality in the data (e.g. the orientation of the fracture zone ridges in the MAR image). Note that this is a different kind of understanding than that assessed by data-skills questions such as “where is the shallowest point in the area of seafloor shown in this image, and what is the water depth at that point?” We are probing understanding of what are the factors of importance in analyzing something. If the data skills question in the previous sentence is analogous to a vocabulary question in literacy education, knowledge of the value of attending to these quantitative attributes of data would be analogous to knowing that in analyzing a novel one should pay attention to character, plot development, setting in place and time, symbolism, etc.

Among the novices, data-derived information seems to be the best developed or the most accessible. From our eye-tracking finding that experts and novices allocate their data exploration time similarly, we infer that the novices are taking appropriate actions to gather this first type of information. From our finding that novices score highest on the salient component of GeoScore, relative to either the experts or relative to the maximum score available (Table 4), we infer that they are being relatively successful in pulling forth this information from the data. For some of the lowest-performing novices (bottom of Figure 3), this is the only one of the three data types that we see manifested. Travelling

up the continuum of overall data competence (upward on Figure 3), scores on this component increase; the strongest two novices on this component score as high as the weakest expert.

The least well-developed or least accessed type of information/knowledge among the novices seems to be the third type, representational competence, which we believe is probed by the quantitative component of the *GeoScore*. The weakest 13 novices, according to overall *GeoScore*, got zero points (out of 16 available) on this component (bottom of Figure 3), and the mean novice score on this component was far worse than the others, whether judged relative to the experts or relative to the maximum points available (Table 4). The remaining type of knowledge, knowledge of the Earth system, which we believe is probed by the Interpretation component of the *GeoScore*, occupies an intermediate position in the proficiency profile of our novices.

Our data show a snapshot at one time, rather than a longitudinal progression of individuals over time. Nonetheless, the distribution of the three components of the *GeoScore* within our novice group suggests that there may be an ordering or progression in which, under existing conditions of instruction and life experiences, learners build access to the three types of knowledge that they will need to interpret data of this type. First they learn to spot and qualitatively describe salient features in the data. Second, they learn to make broad, qualitative interpretations of the processes that formed distinctive landforms. Finally, they learn to attend to quantitative aspects of the landscape, including elevation, relief, size and trend. Once they have acquired the habits of mind of attending to quantitative aspects of the data—and the skill set to do so—they would be in a position to develop a more sophisticated set of interpretations grounded in quantitative models of tectonic and erosional processes. This last step is outside the part of the learning trajectory exhibited in the current study.

This proposed progression follows the evolution of the humanity's knowledge across intellectual history. The propensity to attend to salient features in the landscape probably goes far back into humanity's evolutionary past, as it would have been of high value to our hunter-gatherer ancestors. Making qualitative process-oriented interpretations of the origin of landforms was the research agenda of the field of geomorphology from its founding in the nineteenth century through the mid-twentieth century. With the advent of quantitative geomorphology in the latter half of the 20th century, the scientific community began to extract insights from the quantitative as well as qualitative aspects of topographic data.

When asked how they would coach an undergraduate summer intern on how to look at and interpret the data, the experts mostly provided knowledge of types one and two, pointing out salient features in the data and stating the interpretation. They were not consistent in drawing attention to quantitative attributes of the data—although they themselves attended to such attributes.

5. Implications for Instruction

Our students' interview responses confirm, as many researchers and instructors have found before us, that the degree of knowledge of college undergraduates about their own planet is highly variable and the low end is extremely low (Figure 2). This may be in part because so many students in the United States, especially college-bound students, do not study Earth Science in high school (Wilson, 2014). But it may also be that much of modern science teaching focuses on building an understanding of concepts and processes, and that students have less opportunity to construct an understanding of what the traces of these processes look like in the real world, especially when the traces of multiple processes are superimposed on each other in the same dataset (Gould, Sunbury & Dussault, 2014).

Establishing the connection between the conceptual model of an Earth process and the observable trace left behind by that process can proceed in two different instructional contexts, summarized by two questions: (1) How do scientists know that? and (2) What is going on in this data? The former is a historical approach, in which the data and reasoning that enabled scientists to construct and bolster a specific scientific claim is laid out. The latter approach assumes that students are already familiar with the candidate processes that may have shaped the area represented by the data, and the challenge for the student is to recognize the trace of the process in the data. Many of our novices were broadly familiar with the concepts of erosion, rivers, faults, volcanoes, but they could not necessarily recognize the traces of these phenomena in data.

Based on our findings and prior work, we offer the following suggestions for how to guide students along the pathway towards making meaning from data:

- *Keep in mind that you know a lot more than students do:* Obviously, you have a stronger conceptual framework (Bransford, *et al.*, 2000): you know more than they do about plate tectonics, and erosion, and volcanism, and other Earth processes. Less obviously, you also know more about relationships between Earth processes and the traces they leave behind; you have an eye educated to spot salient patterns and trends; you have the habits of mind of quantifying observations and seeking causes for spatial patterns; and you have a toolkit of representational strategies.
- *Use perceptual learning effectively.* Point out distinctive features in data, and identify what they are, using professional terminology. “See the doggie” was the most common type of guidance offered by our experts to the imaginary undergraduate learner, and it is how we all learned our first language, and it taps into a powerful human cognitive ability. To be effective, perceptual learning requires exposure to multiple instances of each type of feature to be learned and named. The multiple instances should have both commonalities and differences: the shared aspects that make each instance part of the category being learned (for example, “moraine”), and the range of differences that make this a category rather than an identity (for example, the Ronkonoma Moraine).
- *Reinforce perceptual learning with feedback:* Kellman’s work (Kellman & Massey, 2013) shows that perceptual learning can be accelerated by providing rapid feedback on correctness of the learners’ identification. It would be possible to adapt the Perceptual/Adaptive Learning Module (PALM) approach from medical education to visually-identified geoscience features, such as rocks and minerals, landforms, structures, and features on remote sensing imagery.
- *Articulate the criteria:* In pure perceptual learning, the criteria are not stated explicitly. The parent doesn’t say to the child, “The way that I know it’s a dog is because it has 4 legs and fur and is attached to a human by a leash.....” Likewise, our experts typically did not say to the imaginary undergraduate learner, “The reason I think these are volcanoes is because they are conical in shape, and stick up by themselves against an otherwise relatively smooth background....” However some of our experts did articulate their criteria as they identified features for the imagined undergraduate, and research on other difficult perceptual learning tasks has shown rapid learning when previously unarticulated criteria are explicitly presented to the learner (Biederman & Shiffrar, 1987).
- *Have students articulate the criteria for the category:* In perceptual learning, the learner unconsciously does what scholars of analogic reasoning call “extracting the schemata” (Gentner & Colhoun, 2010; Jee, *et al.*, 2010). During the course of examining multiple non-identical instances of a phenomenon, a human can extract the features that are in common across the instances and ignore the features that are not. This process requires and constructs quite a deep understand of the phenomena, and is therefore a powerful learning process (Gentner, 2010). Rather than having the instructor articulate the criteria for a specific kind of feature in a type of data, it could be more powerful to set up a learning opportunity in which students examine and categorize phenomena in data. Discovering Plate Boundaries (Sawyer, *et al.*, 2005) uses this approach, in which student groups decide upon and articulate their own classification scheme for plate boundaries based on maps of earthquakes, volcanoes, crustal age and topography.
- *Encourage use of the full set of available cartographic or graphic clues:* The novices in our study rarely (relative to the experts) looked at the lat/long axes or the distance scale. Getting a handle on the scale of the map and thus the size of the features shown was the most frequently offered generalizable guidance given by our experts in coaching the imagined undergraduate. Similar patterns recur at different scales in the physical world, and thus students should know that scale is an important discriminant: scale distinguishes a conical pimple from a molehole from a seamount. Novices varied widely in their attention to the color bar (depth scale) provided on each image, and so even with college students it would be appropriate to reiterate the fundamental map skill of reading the map legend.
- *Foster the habit of mind of characterizing data quantitatively:* The biggest difference between our novices and experts in their spoken responses was that experts spontaneously gave us quantitative descriptions of highs, lows, characteristic values, and sizes (vertical relief and horizontal length/distance), as well as trends. Novices mostly did not. Note that it’s probably not enough to pose questions like: “Where is the shallowest point, and how deep is the water depth at that point?” Although we didn’t ask such questions, we anticipate that the expert/novice difference on such questions would have been narrower than it was on the construct

that we did probe. Our experts exhibited a habit of mind of attending to the values, the range, the trends in data, and our novices mostly did not.

- *Talk metacognitively about generalizable practices in making meaning from data:* Nearly half of the data interpretation guidances that our experts offered to the imagined undergraduate were not generalizable to the next data visualization that the student will encounter, or the one after that, or the one after that. It should be possible to design guiding questions as part of instruction that will trigger metacognitive thinking to extract more generalizable strategies from class discussion. A useful line of discussion could go along the lines of: “OK, so we have just finished interpreting some data visualizations that show bathymetry data. Are there approaches that we used today that we’ll be able to use again next time we have bathymetry data? What about the next time we have data that is a map, but it’s another kind of data? What will we know to look for next time? What will we know to do next time?”
- *Encourage attention to data quality and possible errors introduced during data acquisition or processing:* Most of our experts but few of our novices mentioned the two different levels of data quality in the seamount image (higher resolution is from a ship-mounted sonar; blurrier is from satellite gravimetry data). Hug and McNeill (2008) found that attention to issues of data quality was one of the largest difference between students’ work with data that they had and had not collected themselves. As geoscience education makes more use of large, professionally collected datasets obtained from the internet--as opposed to smaller datasets that students collected themselves--the burden is on instructors and instructional materials to bring students up to speed on how the data were collected and potential error sources.
- *Encourage using contextual information:* As one of our experts said “Context is everything.” Outside of the context of this experiment, professional geoscientists rarely encounter a data visualization in isolation and out of context. Encourage students to seek out and draw on relevant information to provide context, such as a regional map or atlas, scientific literature, or other data types.

ACKNOWLEDGEMENTS

We appreciate the time and insights provided by the expert and novice participants in our study, as well as graphic design work by Linda Pistolesi and Silvia La Vita. This work was supported by the National Science Foundation’s program for Fostering Interdisciplinary Research in Education (FIRE) through NSF Award numbers 1138616 (Columbia University), 1138619 (Temple University), and 1331505 (Education Development Center, Inc.). This is Lamont-Doherty Earth Observatory contribution number 8006.

AUTHOR BIOGRAPHIES

Kim A. Kastens is a Special Research Scientist at the Lamont-Doherty Earth Observatory of Columbia University. Her training and early career were in geology and oceanography. The second half of her career has focused on geoscience education research and public understanding of the Earth and environment. She has a long-standing interest in how teachers teach and students learn from geoscience data, especially in support of problem-solving and decision-making. In pursuit of this interest, she co-founded the Oceans of Data Institute at the Education Development Center, where some of the work report herein was conducted. E-mail: kastens@ldeo.columbia.edu (Contact author)

Thomas F. Shipley is on the faculty of the Department of Psychology and the Spatial Intelligence & Learning Center at Temple University. His early research focused on perception of objects and events. Currently, he applies formal methods from his previous research to understand the perceptual and cognitive processes underlying navigation and visualization. He collaborates frequently with geoscientists and geoscience educators. His research in STEM education works towards two synergistic goals: improving fundamental understanding of the cognitive processes that enable spatial reasoning, and improving teaching and learning by application of insights from cognitive science. E-mail: tshipley@temple.edu

Alexander P. Boone is a Ph.D. candidate in the Department of Psychological and Brain Sciences at the University of California, Santa Barbara. His work has focused on individual differences in both small and large scale spatial cognition. Prior to joining UCSB, he was a lab manager for the Spatial Intelligence & Learning Center at Temple University, where he contributed to the study reported here. E-mail: boone@psych.ucsb.edu

Frances Straccia has a BA in biology from Denison University and an MS in geology from the University of Michigan. She did her thesis on freshwater limestones of the Snake River Plain of Idaho. Her career was in the oil business, first as a wildcat oil finder then as an executive editor for Shell Oil Company's international in-house technology magazine. Several stints overseas broadened her geologic knowledge base and exposed her to different educational approaches. While working on this project, she was a Research Assistant at the Education Development Center. E-mail: franc.g.straccia@gmail.com

REFERENCES

- Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4), 640-645.
- 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 *CAISE Inquiry Group Report*. Washington, D.C.
- Bransford, J., et al. (2000). *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Academy Press.
- Clark, A. C., & Wiebe, E. N. (2000). Scientific visualization for secondary and post-secondary schools. *Journal of Technology Studies*, 26(1), 24-32.
- Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Cognitive activities in complex science text and diagrams. *Contemporary Educational Psychology*, 35(1), 59-74.
- Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. *Cognition*, 50(1), 41-77.
- diSessa, A. A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction*, 22(3), 293-331.
- Edelson, D. C., Gordin, D. N., & Pea, R. D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. *Journal of the Learning Sciences*, 8(3-4), 391-450.
- Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor referenced cognitive tests*. Princeton, NJ: Educational Testing Service.
- Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10), 454-459.
- Feldon, D. F. (2007). The implications of research on expertise for curriculum and pedagogy. *Educational Psychology Review*, 19, 91-110.
- Feltovich, P. J., Prietula, M. J., & Ericsson, K. A. (2006). *Studies of expertise from psychological perspectives*. In Ericsson, K. A.; Charness, N.; Feltovich, P.J.; Hoffman, R. R. (2006). *The Cambridge handbook of expertise and expert performance*. (pp. 41-67). New York, NY, US: Cambridge University Press.
- Geary, D. (2012a). Evolutionary educational psychology. In Jarris, K. R., Graham, S., and Urdan, T. (Eds.), *APA Educational Psychology Handbook: Vol 1. Theories, Constructs, and Critical Issues*, American Psychological Association.
- Geary, D. (2012b) The evolved mind and scientific discovery. In Shrager, J., and Carver, S. M., (Eds.), *From child to scientist: Mechanisms of learning and development* (pp. 87-115). Washington, D.C.: American Psychological Association.
- Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34(5), 752-775.
- Gentner, D., & Colhoun, J. (2010). Analogical processes in human thinking and learning. In B. Glatzeder, V. Goel, & A. von Müller (Vol. Eds.), *On Thinking: Vol. 2. Towards a Theory of Thinking* (pp 35-48). Berlin: Springer-Verlag.
- Gould, R., Sunbury, S., & Dussault, M. (2014). In praise of messy data. *The Science Teacher*, 81(8), 32-36.
- Gray, J. (2009). Jim Gray on eScience: A transformed scientific method. In T. Hey, S. Tansley & K. Tolle (Eds.), *The Fourth Paradigm: Data-intensive Scientific Investigation* (pp. xvii - xxxi): Microsoft Research.
- Haxby, W. F., Karner, G. D., LaBrecque, J. L., & Weissel, J. K. (1983). Digital images of combined oceanic and continental data sets and their use in tectonic studies. *Eos, Transactions American Geophysical Union*, 64(52), 995-1004.
- Hegarty, M. (2012, August). *Metarepresentational competence as an aspect of spatial intelligence*. Annual meeting of the Cognitive Science Society, Sapporo, Japan, online at: <http://mindmodeling.org/cogsci2012/papers/0222/paper0222.pdf> (last accessed 3 Feb 2015).
- Hegarty, M., Canham, M. S., & Fabrikant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(1), 37-53.
- Hey, T., Tansley, S., & Tolle, K. (Eds.). (2010). *The fourth paradigm: data-intensive scientific discovery*: Microsoft Research.
- Hug, B., & McNeill, K. L. (2008). Use of First-hand and Second-hand Data in Science: Does data type influence classroom conversations? *International Journal of Science Education*, 30(13), 1725-1751.
- Jee, B. D., Uttal, D. H., Gentner, D., Manduca, C., Shipley, T. F., Tikoff, B., ... & Sageman, B. (2010). Commentary: Analogical thinking in geoscience education. *Journal of Geoscience Education*, 58(1), 2-13.
- Kastens, K. A. (2015). Philosophy of Earth science. In Gunstone, R. (ed.), *Encyclopedia of Science Education* (pp. 354-357). Dordrecht, Heidelberg, New York, London: Springer.
- Kastens, K. A. (2011). *Learning to learn from data*. Retrieved from <http://serc.carleton.edu/earthandmind/posts/datalearningpro.html>.

- Kastens, K. A., Krumhansl, R., & Baker, I. (2015). Thinking Big: Transitioning your students from working with small student-collected data sets towards "big data". *The Science Teacher*, 82(5), 25-31.
- Kastens, K. A. and C. A. Manduca (2012). Mapping the domain of Time in Geosciences. *Earth & Mind II: Synthesis of Research on Thinking and Learning in the Geosciences*, Geological Society of America Special Publication. K. A. Kastens and C. Manduca. Boulder, Geological Society of America: 13-19.
- Kellman, P. J., & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. *The Psychology of Learning and Motivation*, 58, 117-165.
- Kirton, M. (1978). Field dependence and adaption-innovation theories. *Perceptual and Motor Skills*, 47(3f), 1239-1245.
- Le, T. K., Silver, E. J., Shemwell, J. T., Capps, D. K., & Voyer, C. E. (2015). *How learning during scientific observation can influence students' reasoning with evidence*. Paper presented at the National Association for Research in Science Teaching, Chicago, IL.
- Manduca, C. A., & Kastens, K. A. (2012). Geoscience and geoscientists: Uniquely equipped to study Earth. *Geological Society of America Special Papers*, 486, 1-12.
- Manduca, C., & Mogk, D. W. (2002). Using Data in Undergraduate Science Classrooms. Carleton College: National Science Digital Library. Online at: <http://d320goqmya1dw8.cloudfront.net/files/usingdata/UsingData.pdf> (last accessed Nov. 24, 2015)
- MayIn-Smith, J., & Ippolito, J. (2015). *Defining emerging occupations: Social technology enabled professional, Big data enabled professional*. Paper presented at the Society for Information Technology & Teacher Education International Conference 2015, Chesapeake, VA.
- Myers, L. J., & Liben, L. S. (2008). The role of intentionality and iconicity in children's developing comprehension and production of cartographic symbols. *Child Development*, 79(3), 668 – 684.
- Phipps, M., & Rowe, S. (2010). Seeing satellite data. *Public Understanding of Science*, 19(3), 311-321.
- Ryan, W. B., Carbotte, S. M., Coplan, J. O., O'Hara, S., Melkonian, A., Arko, R., ... & Zemsky, R. (2009). Global multi-resolution topography synthesis. *Geochemistry, Geophysics, Geosystems*, 10(3).
- Roth, W. M. (1996). Where IS the Context in Contextual Word Problem?: Mathematical Practices and Products in Grade 8 Students' Answers to Story Problems. *Cognition and Instruction*, 14(4), 487-527.
- Rubin, E. (2001). Figure and Ground. In S. Yantis (Ed.), *Visual Perception*. Philadelphia: Psychology Press.
- Saarinen, T. F. (1973). Student views of the world. In R. M. Downs & D. Stea (Eds.), *Image and environment: Cognitive mapping and spatial behavior* (pp. 148-161). Chicago: Aldine.
- Sawyer, D. S., Henning, A. T., Shipp, S., & Dunbar, R. W. (2005). A data rich exercise for discovering plate boundary processes. *Journal of Geoscience Education*, 53(1), 65-74.
- Science Education Resource Center. (2014). Using data in the classroom: A site for educators and resource developers, on-line resource. Retrieved from. <http://serc.carleton.edu/usingdata/index.html> last accessed Feb 27, 2015.
- Shiple, T. F. (2008). An invitation to an event. In T. F. Shipley & J. M. Zachs (Eds.), *Understanding events: From perception to action* (pp. 3-30). Oxford: Oxford University Press.
- Shiple, T. F., Tikoff, B., Ormand, C., & Manduca, C. (2013). Structural geology practice and learning, from the perspective of cognitive science. *Journal of Structural Geology*, 54, 72-84.
- Simon, H. A., & Chase, W. G. (1973). Skill Chess: Experiments with chess-playing tasks and computer simulation of skilled performances throw light on some human perceptual and memory processes. *American Scientist*, 61(4), 394-403.
- Singer, S. A., Nielsen, N. R., & Schweingruber, H. A. (2012). *Discipline-based education research*. Washington, D.C.: National Academy Press.
- Solomon, G. E. A., & Zaitchik, D. (2012). Folkbiology. *Wiley interdisciplinary reviews*, 3(1), 105-115.
- Swenson, S., & Kastens, K. (2011). Student interpretation of a global elevation map: What it is, how it was made, and what it is useful for. In Feig, A., & Stokes, A., (Eds.), *Qualitative Inquiry in Geoscience Education Research*, *Geological Society of America Special Papers*, 474, 189-211.
- Trumbull, D. J., Bonney, R., Bascom, D., & Cabral, A. (2000). Thinking scientifically during participation in a citizen-science project. *Science Education*, 84(2), 265-275.
- Wilson, A. H., (2014). *Status of the Geoscience workforce*. Alexandria, VA: American Geosciences Institute.
- Witkin, H. A. (1950). Individual differences in ease of perception of embedded figures. *Journal of Personality*, 19, 1-15.
- Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R., & Karp, S. A. (1962). *Psychological differentiation: Studies of Development*. New York: Wiley.

APPENDIX

Table 1. Interview protocols

Novices	Image	Experts
<ul style="list-style-type: none"> • Explanation & informed consent • Eye-tracking calibration 	None	<ul style="list-style-type: none"> • Explanation & informed consent • Eye-tracking Calibration
(2.1) What do you think this is? <i>[follow up questions]</i>	Global map	(2.1) What do you think this is?
(2.2) OK, good, so you have told me that you think this is a <i>[map/image/picture]</i> of <i>[interpretation]</i> . What clues in the image led you to think that it shows <i>[interpretation]</i> ?		(2.2) OK, good, so you have told me that you think this is a <i>[map/image/picture]</i> of <i>[interpretation]</i> . What clues in the image led you to think that it shows <i>[interpretation]</i> ?
(2.3) Was there anything else in the image that led you to that interpretation?		(2.3) Was there anything else in the image that led you to that interpretation?
(2.4) Could you please point to an example of where you think the image is showing <i>[interp.]</i>		
<ul style="list-style-type: none"> • Explanation of the global map 		
(3.1) What do you think this image is showing? <i>[follow-up questions]</i>	One of four hi-resolution images, in random order.	(3.1) What do you think this image is showing? <i>[follow-up questions]</i>
(3.2) What processes do you think might have shaped this part of the Earth’s surface? <i>[follow-up questions]</i>		(3.2) What processes do you think might have shaped this part of the Earth’s surface? <i>[follow-up questions]</i>
(3.3N) Can you give me any more detail about what you think is going on in this image?		(3.3E) Now please pretend that you are coaching an undergraduate summer intern on how to look at images like this, what to pay attention to, how to interpret the data. What would you say?
(3.4N) What do you think you would see if you could see a larger area of the Earth than we are seeing here, if you could see outside the frame of this image? <i>[follow-up questions]</i>		(3.4E) Where do you think this place is?
(3.5N) Please describe to me what you see in and around the area that was just marked. While you are answering this question, please pretend that I’m in another room and can’t see the image; just use words to describe the marked part of the image as best you can.	Hi-res image with a significant feature marked by a white circle or oval.	(3.5E) Please describe what you see in and around the area that was just marked. Pretend that you are talking on the phone to a colleague in another country who can’t see the image and can’t see you; just use words to describe the marked part of the image as best you can.
<i>Repeat (3.1) to (3.5) for the other hi-res images.</i>	<i>3 more hi-res images</i>	<i>Repeat (3.1) to (3.5) for the other hi-res images.</i>
(4.1) How do you think these images that we’ve been looking at were made?	“Thank you”	
(4.2) What do you think these images are useful for?		
(4.3) Think back over all the images. As you were trying to come up with the answers to my questions, what sources of knowledge did you draw on?		
(4.4) Now I’m going to read you a list of sources of knowledge that some people use as they try to interpret these types of images. For each choice, tell me whether that was one of the sources of knowledge you used....		
		And finally, could I ask to you fill out this form about your background and education. <ul style="list-style-type: none"> • Fill out demographic form

Table 2. Coding Schema for GeoScore

Coded element		High-resolution Image			
		Mid-ocean Ridge	Seamount	Valley & Ridge	River
Quantitative observation	<i>Elevation</i>	Mention the elevation of a specific feature or region of the map in meters.			
	<i>Relief</i>	Mention the vertical difference between the elevation of two mapped features or regions in meters.			
	<i>Distance</i>	Mention the horizontal length or width of a specific landform OR the horizontal distance between two geomorphologic features, in meters or kilometers.			
	<i>Trend/azimuth</i>	Use the terms “north,” “south,” “east” and/or “west” to describe a direction of elongation (as of a ridge) or alignment (as of seamount chain) or motion (as of river). “Right,” “left,” “top” and “bottom” are not acceptable substitutions.			
Salient features	Unique for each image. Lay language acceptable.	Ridge	Seamounts	Ridge	River
		Rift	Flat seafloor	Valley	Plateau
		Offset of ridge	More detail in Mark 1 than 2	Zigzag pattern	Tributaries
	Fracture zone	Multibeam sonar tracks	River	Smooth-textured region of plateau	
Interpretation	Unique for each image. Lay language acceptable.	Divergent motion	Volcanism	Compression/collision	Direction of river flow
		Strike-slip motion	Sedimentation	Erosion	Erosion

Table 3. Coding for experts' responses to the pedagogical question*

Code	Examples
(1) Image-specific details	
(1a) Draw attention to a feature, perhaps name it, but provide no criteria for recognition	“we see the very strong river erosion up here” [V&R] “And then these are the abyssal hills.” [MAR]
(2) Strategies useful across bathymetry/ topography visualizations	
(2a) Elevation: Realize you are looking at depth/elevation/ shape of Earth's surface	“... the elevation that we're looking at, that goes from sea level basically up to about 900 meters give or take and you get an idea of the vertical extent” [V&R]
(2b) Rules of thumb about geomorphology: Things useful to look for (e.g. lineations, symmetry); connections between landforms and process.	“look at the pattern, so principally we see a pattern that looks dendritic with an overall downward direction of fluvial flow.” [River] “you're using your eye to look for linear or near linear features” [MAR]
(2c) About the data: pointers about the data, not about the Earth; includes differing resolutions of ship versus satellite bathymetry	“I guess the first thing they would need to know is how to interpret a color-coded false illumination elevation model...” [V&R]
(3) Strategies useful across maps in general	
(3a) Scale: Look at lat, long, distance scale to figure out how big things are	“Establish the scale of map, and there is a scale bar provided, but in addition one knows this is apparently a Mercator projection that you can trust the dimensions are such that one minute north south is like a mile.” [River]
(3b) Position: Use lat/long to figure out where you are in the world	“... take a look at where on Earth we actually are. So actually now that I read this 116°W and 47°N that would put us, I guess that's the US, right? Anyway check the latitude and longitude and figure out where in the world we are.” [River]
(3c) Other generalizable map strategy (envision that you are there; break up the map into homogeneous units)	“So, they can imagine if they were walking through this ravine, it would actually be like a ravine that shoots deep down.” [River] “Don't trust that every map that you see of the sea floor reveals everything you hope to find.” [SM]
Code	Examples
(4) Strategies useful across data visualizations in general	
(4a) Key or legend: figure out what the representation system is	“First to understand what the image is: that it is depth to the seafloor” [SM]
(4b) Context: Refer to additional information sources (field experience, regional map, literature, other data sets)	“When you put that together with field experience, they tell the story that makes sense.” [V&R] “I would encourage the student to get an atlas.” [SM]

* Experts (n=14) were asked: “Now please pretend that you are coaching an undergraduate summer intern on how to look at images like this, what to pay attention to, how to interpret the data. What would you say?”

Table 4. Expert vs Novice Performance on 4-image GeoScore

	Total GeoScore		Salient Features		Interpretation		Quantitative	
	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert
Mean(SD) Correct	8.9 (5.2)	31.8 (4.3)	5.6 (2.7)	13.9 (1.3)	2.0 (1.4)	6.6 (0.7)	1.4 (2.0)	11.2 (3.5)
Percent Correct	22.3%	79.5%	34.7%	87.1%	25.0%	83.0%	8.6%	70.1%
Percent of Expert Score	28.1%		39.9%		30.1%		12.3%	

Notes: Maximum possible score is 40 for Total GeoScore, 16 for Salient, 8 for Interpretation, and 16 for Quantitative. N = 45 for novices and 14 for experts.

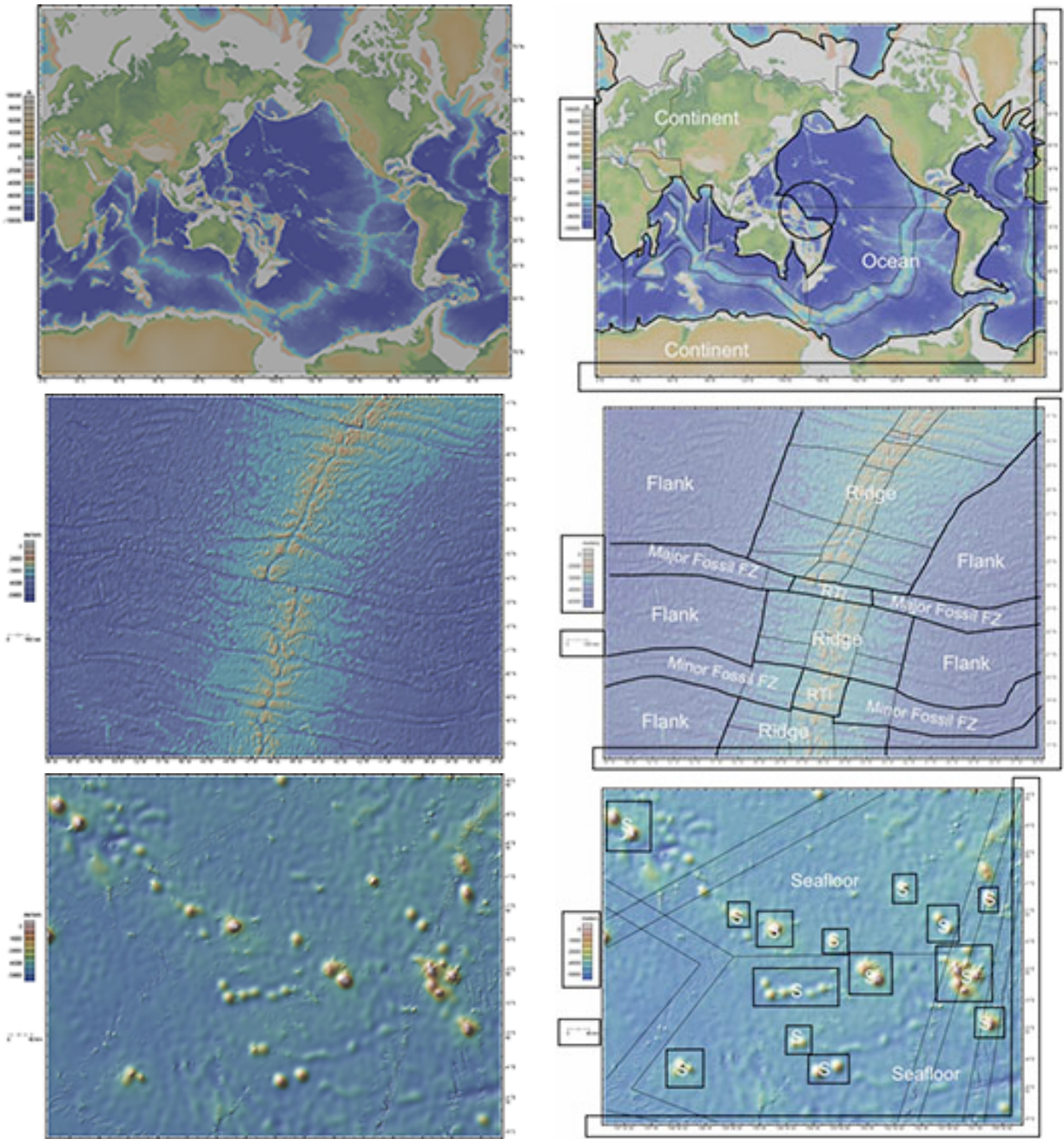
Table 5. Mean number of Fixations per Geomorphologic AOI, sorted for experts and for novices

Experts		Novices	
Category AOI	Mean #	Category AOI	Mean #
MAR: Ridge-Trans Intersec.	3.24	MAR: Ridge-Trans Intersec.	2.34
SM: Seamount	1.67	SM: Seamount	1.94
River: Canyon	1.22	River: Canyon	1.09
MAR: Ridge	1.03	MAR: Ridge	0.98
VR: Valley& Ridge	1.03	<i>Global: Continent</i>	0.88
<i>Global: Ocean</i>	0.75	VR: Valley&Ridge	0.80
<i>Global: Continent</i>	0.55	SM: Seafloor	0.58
SM: Seafloor	0.54	River: Non-canyon	0.53
River: Non-canyon	0.48	VR: Plateau	0.52
VR: Plateau	0.44	VR: Lowlands	0.44
VR: Lowlands	0.42	<i>Global: Ocean</i>	0.41
MAR: Major Fossil FZ	0.32	MAR: Major Fossil FZ	0.41
MAR: Flank	0.14	MAR: Flank	0.19
MAR: Minor Fossil FZ	0.10	MAR: Minor Fossil FZ	0.11

Notes: Abbreviations for images: MAR: Mid-Atlantic Ridge; SM: Seamount; VR: Valley & Ridge. Values are normalized by dividing mean number of fixations in an AOI by the area of the AOI in percent of image. Thus units are # of fixations per 1% of image.

Figure 1.

(Left) The data visualizations as seen by the participant. The image size as seen by the participants was approximately 27cm across in the horizontal. All participants saw the global image first, followed by the four high-resolution images in randomized order. The high resolution images depict a portion of the Mid-Atlantic Ridge, an area of seafloor in the southwestern Pacific studded with seamounts, the Valley & Ridge province of the Appalachians, and a portion of the Columbia River and its tributaries. (right) The same images subdivided into areas of interest (AOI's). Individual AOI's are separated by fine lines, and heavier lines separate the category AOI's. The geomorphological category AOI's for each image are labeled. FZ = fracture zone; RTI = Ridge-transform-intersection; S = Seamount.



(Figure 1 continued on next page)

(Figure 1 continued)

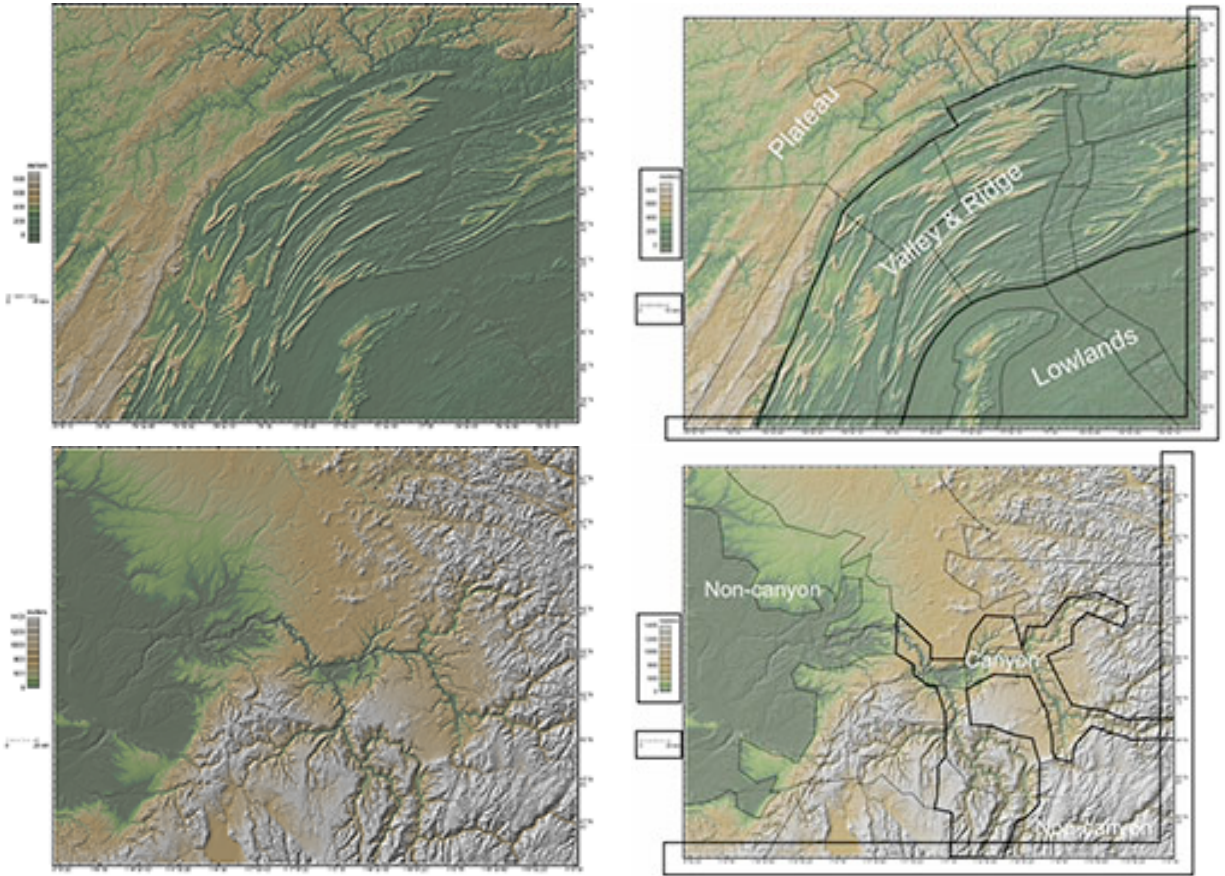


Figure 2. Bar graphs show the total 4-image GeoScore for each participant, ordered from high to low, out of a total possible score of 40. Red bars are novices and blue bars are experts. Among the experts, the two lighter bars are individuals for whom we have video but no eye-tracking data. There is a very wide range of scores in our sample, and no overlap between novices and experts

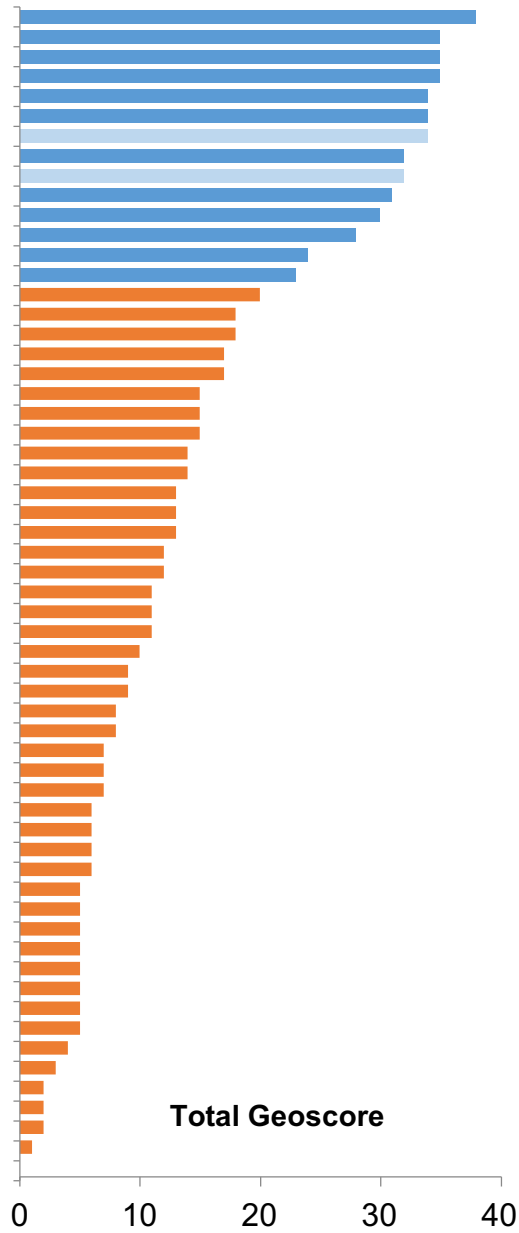


Figure 3. The 4 image GeoScore is broken down into its constituent parts. Low-performing novices get most of their points from describing the salient features that they observe in the data. High-performing novices, and especially experts, pick up additional points from interpreting processes and from articulating quantitative attributes of the features viewed. Arrows show domain of experts and novices respectively, and the top bar would be a perfect score.

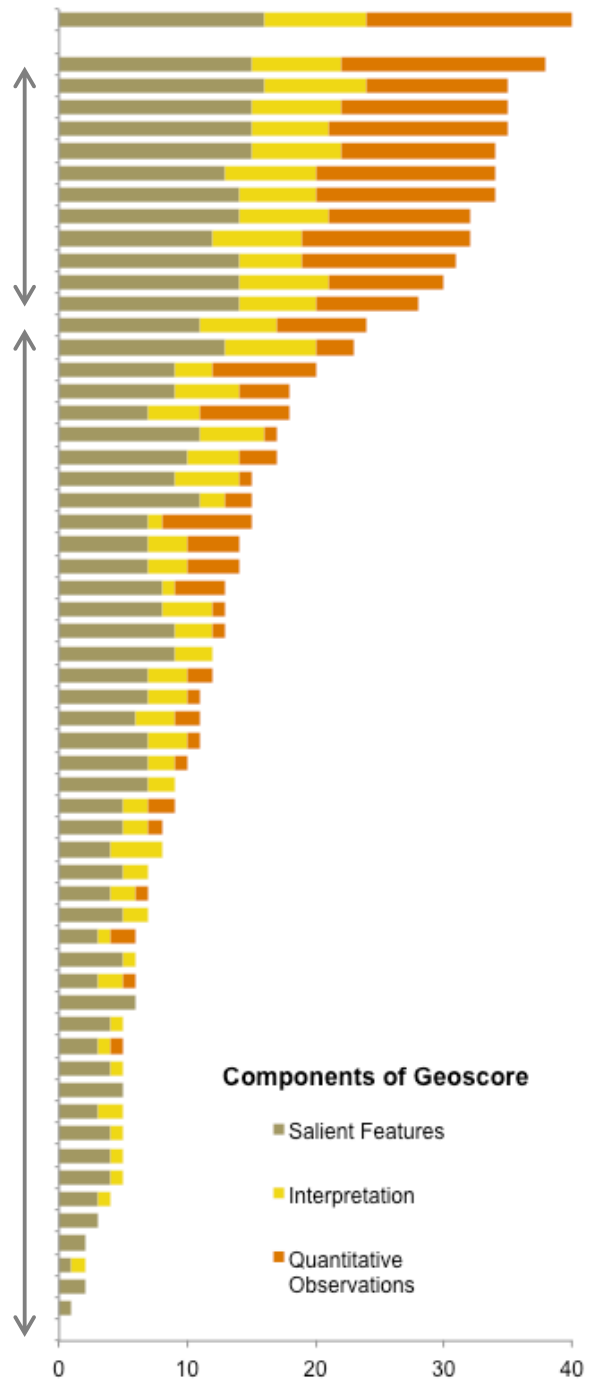


Figure 4. For each AOI across all five images, the scatterplot shows the mean amount of time participants spent in that AOI versus the mean number of fixations participants gave to that AOI, both normalized by the area of the AOI. Time and number of fixations in AOI are strongly correlated. Experts (red open circles) and novices (blue filled circles) behaved similarly.

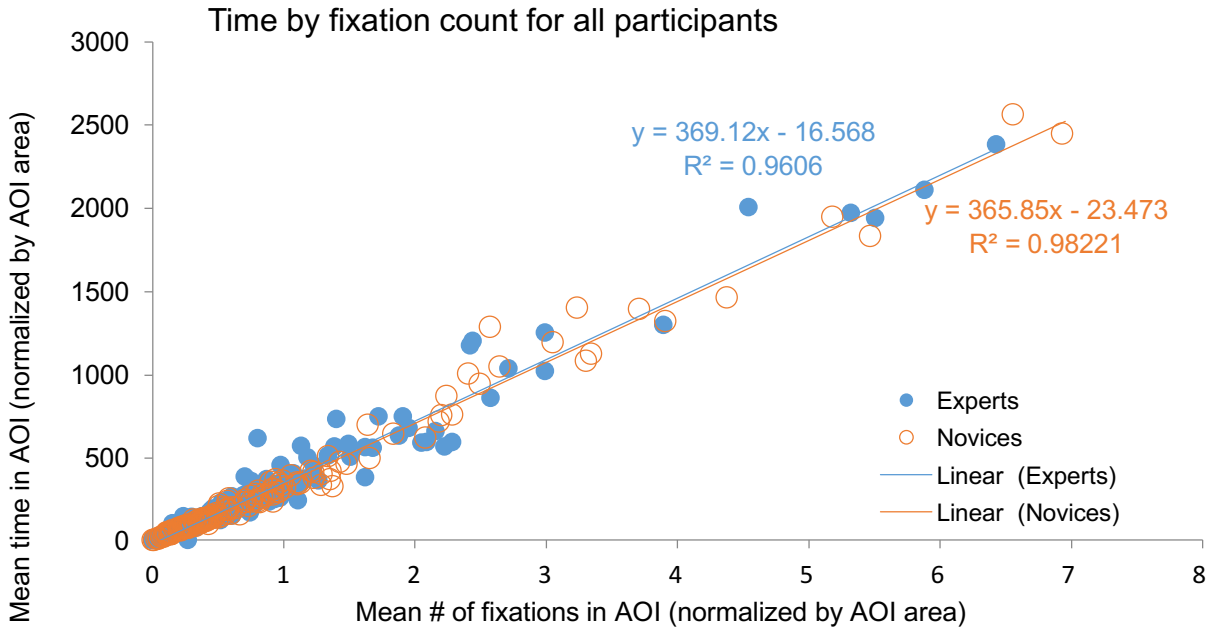


Figure 5. Each symbol represents one category AOI on one image, and plots the mean number of fixations experts have invested in that AOI (normalized by area of the AOI) versus mean number of fixations novices have invested in that same AOI. Symbols above the 1:1 diagonal line represent AOI's where the novices have invested more fixations than the experts; symbols below the diagonal line show where experts invested more fixations than the novices. As highlighted by the marked symbols, novices attended strongly to the color bar (map key) on each of the four high-res images. Experts attended to the lat/long axes more so than the novices.

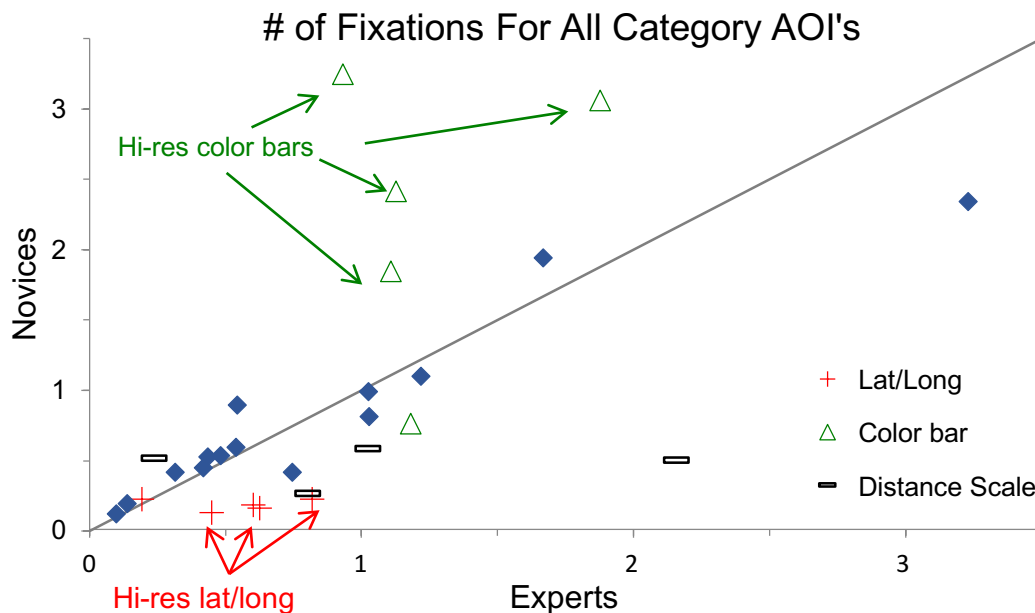


Figure 6. Looking only at the category AOI's that encompass the data, the geomorphology AOI's, we find that there is a strong correlation between the mean number of fixations (normalized by AOI area) that experts and novices invest in each AOI. In other words, experts and novices invest their exploration time similarly. Points that lie furthest from the regression line are labeled and discussed in the text.

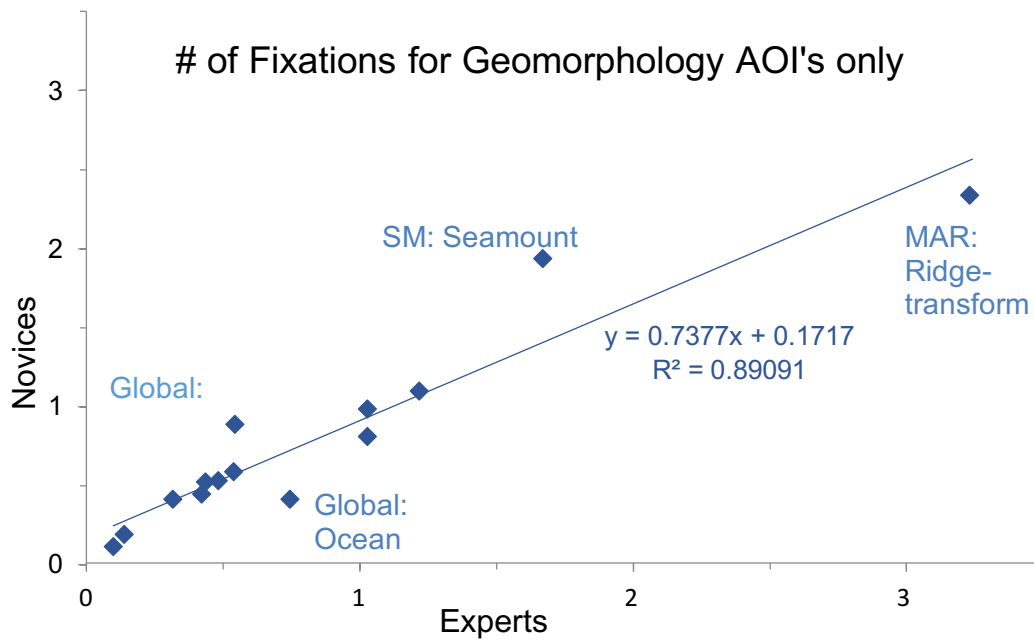


Figure 7. On these scatterplots, each symbol represents one participant. The horizontal axis shows the participant's GeoScore, while the vertical shows what percentage of his/her Global image fixations were spent in continents (upper graph) or in North America (lower). Relative to experts, novices spend more time on continents, and especially in North America. There is a slight tendency for high-GeoScore novices to be more expert-like in their lower degree of attention to North America. The green arrow on the vertical axis marks the percentage of the image area occupied by continents (upper) and North America (lower). Experts' allocation of viewing time is proportional to the occupied area of the image, while novices overinvest in continents and North America.

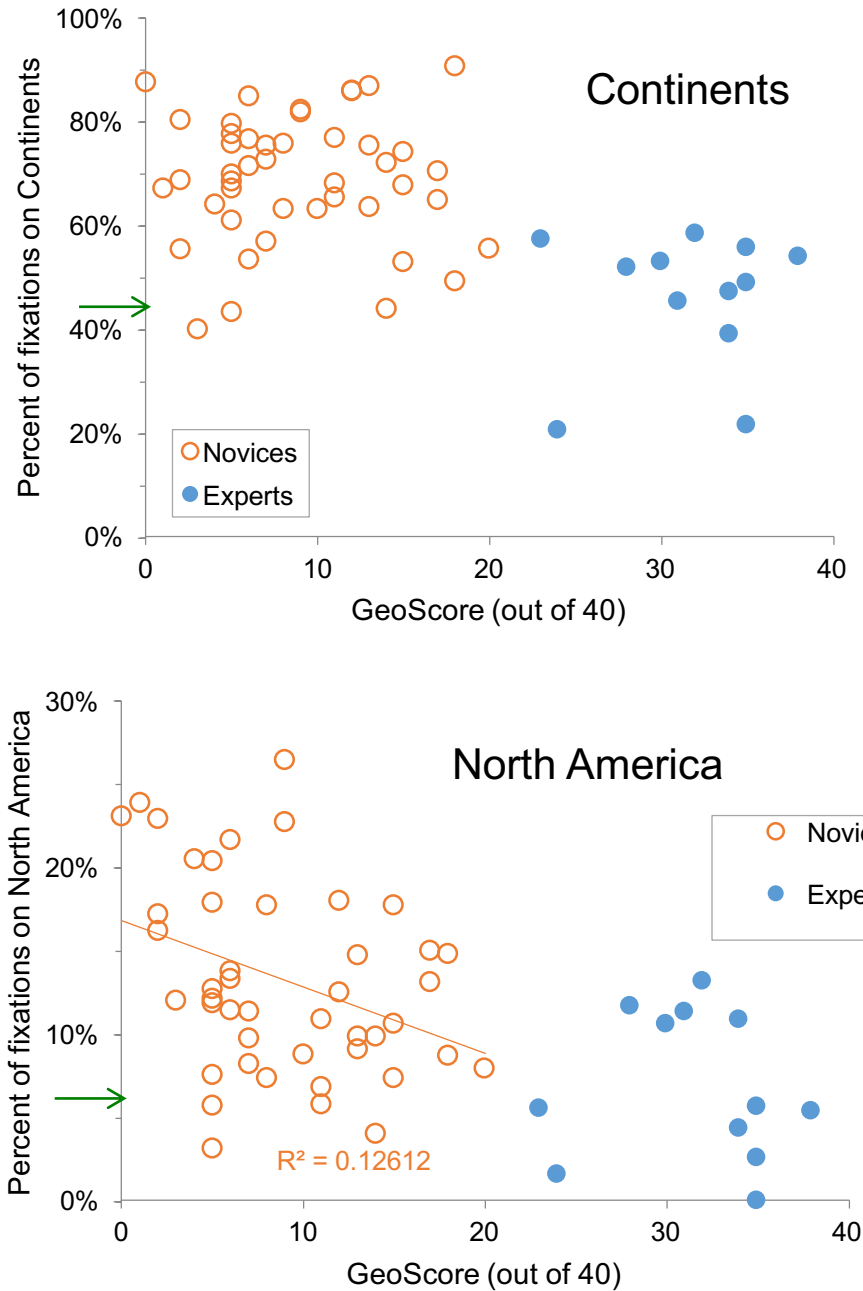


Figure 8. On these scatterplots, each symbol represents one participant. The horizontal value shows that participant's GeoScore, while the vertical axis shows what percentage of his/her fixations were spent in the color bars (upper graph) or in the latitude and longitude AOI's (lower graph). On average, novices spend more time than experts on color bars, and less time on the lat/long axes, but there is substantial overlap. There is not a tendency for high-GeoScore novices to be more expert-like in their exploration behavior on these two metrics. The green arrow on the vertical axis marks the approximate percentage of the image area occupied by the color bar (upper) and latitude + longitude bars (lower).

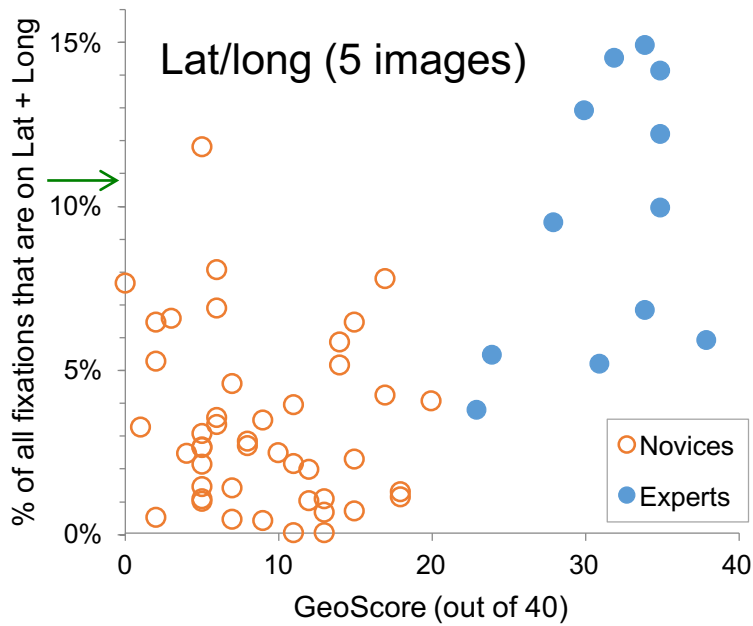
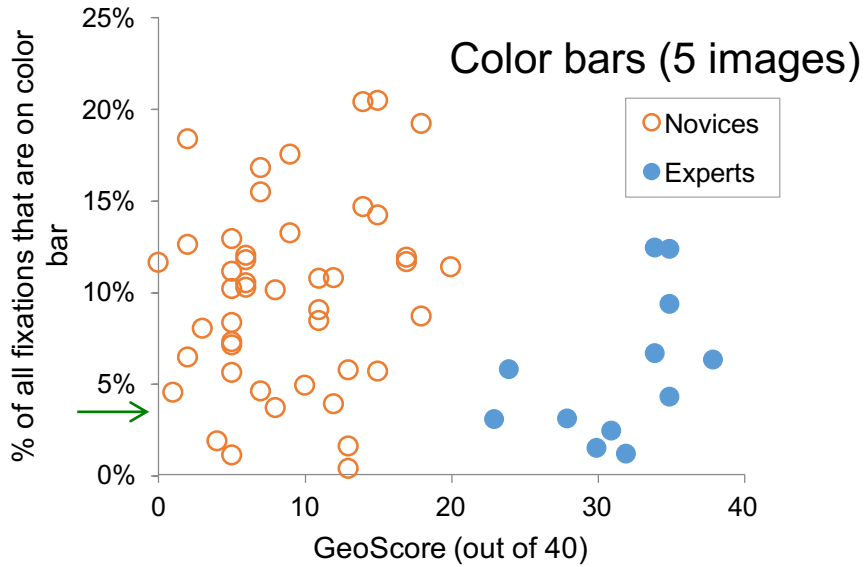


Figure 9. Experts' responses to the question about how they would coach an undergraduate on how to interpret this type of data, with coding categories as in Table 3. (Upper) Percentage of guidances in each category out the total of 338 recorded guidances. (Lower) Percent of opportunities in which a guidance in that category was recorded, where an opportunity is one expert viewing one image. The most common type of guidance is to point out a feature that is specific to that image (coding category 1).

