JOURNAL OF GEOSCIENCE EDUCATION
2018, VOL. 66, NO. 1, 55–76
https://doi.org/10.1080/10899995.2018.1411724

Routledge
Taylor & Francis Group

How students reason about visualizations from large professionally collected data sets: A study of students approaching the threshold of data proficiency

Ilyse Resnick[a], Kim A. Kastens[b], and Thomas F. Shipley[c]

[a]Department of Psychology, Penn State University Lehigh Valley, Center Valley, Pennsylvania 18034; [b]Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964; [c]Department of Psychology, Temple University, Philadelphia, Pennsylvania 19122

ABSTRACT

This study identifies a population of students who have an intermediate amount of relevant content knowledge and skill for working with data, and characterizes their approach to interpreting a challenging data-based visualization. Thirty-three undergraduate students enrolled in an introductory environmental science course reasoned about salinity data as shown in map and vertical profiles from the Mediterranean while thinking aloud and being eye-tracked. Students reasoned about 2D and 3D interpretations in the context of two hypothesis arrays (a suite of potential interpretations about a set of data). Findings suggest the students have some effective strategies in reading data: They look at cartographic elements, correctly identify the image as a salinity map, and draw inferences from the data. Common looking strategies include scanning along the salinity gradient, comparing areas of interest, and aligning the color bar with the map. Individual differences emerge in the interpretation of the data, with no interpretations being fully aligned with the scientifically normative explanation. Post hoc analyses identify reasoning tasks and spontaneous behaviors related to a construct we refer to as "data expertise," which is intended to capture the degree of conceptual sophistication and resourcefulness in reasoning about data. A data expertise scale was developed, with scores ranging from zero (weak) to six (strong) that were normally distributed. Our findings suggest that appropriately coordinating data with a model, comparing and contrasting across data representations from different times or places, and extracting 3D structure from 2D representations are associated with data expertise.

ARTICLE HISTORY
Received 18 August 2016
Revised 09 June 2017
Accepted 05 September 2017
Published online 23 February 2018

KEYWORDS
Reasoning; data expertise; novice; STEM

Introduction

Benchmarks for Science Literacy (American Association for the Advancement of Science, 2008), *A Science Framework for K–12 Science Education* (National Research Council [NRC], 2012), *Next Generation Science Standards* (NGSS Lead States, 2013), and many educators (Manduca & Mogk, 2002) emphasize students' direct engagement with data to develop scientific habits of mind and practices. Working with real data can improve students' reasoning about uncertainty and their quantitative skills (Creilson et al., 2008). Although scientists often reason about small-scale data sets from experiments and observations, reasoning about multiple and large-scale data sets has become increasingly important in science (National Research Council, 2010; Wolkovich, Regetz, & O'Connor, 2012). Having access to large-scale data sets allows scientific breakthroughs that could not be achieved from data collected by a single researcher, and may allow students and early career researchers to ask and answer bigger questions than would be possible if they were limited to only data they had collected themselves (Kastens, 2012; Linik, 2015; National Research Council, 2010).

To effectively incorporate large, professionally collected data into student activities, we need to better understand how such data is read, analyzed, and interpreted by students at different levels of mastery. However, most science education research on students' understanding of data involves small, student-collected data sets, and findings from such studies may not carry over to large, professionally collected datasets (Kastens, Krumhansl, & Baker, 2015). When students collect their own data, they develop an embodied understanding of the setting, methods, and potential data issues such as missing data and measurement error (Hug & McNeill, 2008), which they may lack with other-collected data sets. Interpretation of professionally collected data sets may also involve domain-specific knowledge of the referent systems and of specialized representational strategies.

CONTACT Ilyse Resnick imr9@psu.edu Department of Psychology, Penn State University Lehigh Valley, 2809 Saucon Valley Road, Center Valley, PA 18034, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ujge.

Reasoning about large, professionally collected data can require the coordination of multiple sources of data (which can be from different times and locations), as well as an assortment of data-based visualizations and conceptual process models (which provide candidate explanations for the observed phenomena). Being able to connect different scientific ideas to form a coherent model is referred to as *knowledge integration* (e.g., Clark & Linn, 2003/2009; Linn, 2000). Scientific knowledge integration can be challenging for students, especially in geoscience, which draws from such a range of disciplines (Kastens & Manduca, 2012). An added difficulty is that naïve conceptions are often robust to change (Chi, 2005). Openness to conceptual change depends on the learner's level of engagement, depth and organization of background knowledge, motivation, disposition, willingness to engage with complex messages, and perception of the new content as understandable, coherent, plausible, and compelling (Dole & Sinatra, 1998; Lombardi, Sinatra, & Nussbaum, 2013). To change one's conceptual model when faced with data that disagree with the model is a sophisticated habit of mind that students often lack (Chi, 2005).

As described above, reasoning about data involves a host of different skills and knowledge. Consequently, developing a construct for how and how well people reason about data is challenging, because such reasoning is multifaceted (Mandinach & Gummer, 2012). Defining such an important construct is crucial, however, for developing appropriate teaching materials as well as valid and reliable assessments (Pellegrino, Wilson, Koenig, & Beatty, 2014). There has been a recent focus on defining *data literacy* versus *data fluency* (e.g., Greenberg & Walsh, 2012; Mandinach & Gummer, 2012; Manduca & Mogk, 2002). These terms imply empirically distinct stages (i.e., achieving literacy and fluency); however, conceptualizing one's ability to reason about data as a continuum is aligned with the learning progression framework presented by the National Research Council (2007, 2014). The authors referred to such a continuum as a *data expertise continuum*, and suggested that as persons move along the data expertise continuum, they develop increasing degrees of conceptual sophistication and resourcefulness in reasoning about data. A person with some experience working with data may be able to reason about simple and complete data sets for well-structured problems. As this person gains more experience, he or she would eventually be able to reason about much more complex, novel, and ill-structured problems based on incomplete data sets. We note that this approach is not necessarily mutually exclusive from approaches to learning progressions that include multiple threads connecting different content domains (e.g., Wilson, 2005, 2009; for reviews of approaches to learning progressions, see Salinas, 2009). Although data expertise may be composed of a single continuum, it is also likely intertwined with understanding specific scientific phenomena and concepts.

Current study

In this observational study, we document student behaviors when reading a data-based visualization, such as looking strategies and scientific reasoning. We are interested in students who have some experience with data; they are no longer novices but not yet experts. The abilities, behaviors, and strategies of this population of student are less researched compared to complete novices or even experts. However, this period likely plays a key role in transitioning from novice to expert, and may represent a barrier or falling off point for many students. In examining the knowledge and abilities of intermediate data interpreters, we can also better understand the kind of curriculum this population requires to support their learning. In particular, the current study addresses the following research questions:

1. What is the portfolio of students' looking and behavioral strategies when interpreting complex data-based visualizations, and how common are each of these strategies? This includes what information they attend to, how they coordinate multiple sources of information, and how they connect their claims and interpretations with observations through scientific or logical reasoning.
2. Within this portfolio, which looking and behavioral strategies are associated with greater and lesser levels of data expertise?
3. Do students who have higher levels of data expertise perform better or differently than students with lower levels of data expertise when completing interpretations that are more challenging? For example, here we examine a summative interpretative task that requires integrating 2D and 3D information from multiple complex data visualizations along with process models.

Criteria for selecting the task and stimuli

Because a main aim of this study was to document reasoning processes, behaviors, and strategies of students, we used a task that went beyond simply decoding and describing data, and required interpretation at the outer limits of their data interpretation skill set. The selected task (described in "Methods") is modified from a classroom activity that coauthor KK had used in teaching undergraduate geoscience majors and nonmajors, which

was known to be challenging but not completely out of reach for this population.

This task was also chosen because it exercises habits of mind that are particularly characteristic of geosciences (Manduca & Kastens, 2012). Geoscience often involves spatial thinking (Kastens & Ishikawa, 2006); the chosen task requires grappling with Earth phenomena in three spatial dimensions plus time. Geoscience is primarily an observational rather than an experimental science; the chosen task requires abductive reasoning, reasoning in which the results are observed and the causative process (es) must be inferred (Magnini, 2004; Oh, 2010). The chosen task importantly includes color-coded data-based visualizations of Earth processes. Visualizations are frequently used for seeking patterns and trends in geoscience data (e.g., Merwade & Ruddell, 2012; Ware, 2004), and are known to present difficulties for students (e.g., Hegarty, Canham, & Fabrikant, 2010; Kastens, Shipley, Boone, & Straccia, 2016; Swenson & Kastens, 2011). The final criterion in selecting the task was that it should require students to reason from models of Earth processes as they interpret the data, coordinating between model and data. In science, data are used to inform the development of models, and then models are used to shape the interpretation of data. The ability and habit of mind of comparing the behavior of models with the behavior of the Earth as captured in data, with the goal of evaluating and improving models, is central to modern geosciences—and yet this process can often be invisible to students (Kastens, 2015).

Scaffolding students' reasoning

Because we wanted to observe students as they interpreted complex data in terms of process models, we needed to ensure that they had access to models from which to reason. We used a series of hypothesis arrays (Kastens, 2009; Kastens et al., 2015) to set up increasingly complex problems. A hypothesis array provides a range of candidate hypotheses or explanatory models, which students can use to organize their data exploration, much as alternative working hypotheses can help direct an expert's exploration of new data (Cleland, 2001). For example, reasoning about the cause of an observed concentration gradient could be scaffolded by a hypothesis array comprising (a) solute is being added at the high concentration end, (b) solute is being removed at the low concentration end, (c) solvent is being added at the low concentration end, or (d) solvent is being removed at the high concentration end. Hypothesis arrays can be verbal, diagrammatic, or mathematical. They resemble multiple-choice questions, but they are distinct in that the alternative response options are designed to guide students' thinking rather than assess their understanding. Hypothesis arrays may scaffold students' attempts to organize their observations by laying out the pertinent dimensions along which observations could profitably be considered (Kastens, Agrawal, & Liben, 2009; Mayer, Mautone, & Prothero, 2002). This structure may allow for a visual alignment so that similarities and contrasts are particularly salient, which is an underlying mechanism of reasoning by analogy (Gentner, 1983). Hypothesis arrays allow the student to have some immediate success during the interview session, scaffolding them to be able to continue to reason around a challenging data set. A major benefit of this approach in the research context is that complex reasoning can develop within one interview session rather than taking all semester.

Methods

Study population and setting

Participants were enrolled in an introductory environmental science course during the Fall 2012 semester at a private women's liberal arts college. In this course, the students studied oceanographic processes in a local estuary through both laboratory and field experiences. Thus, all students had some experience engaging with oceanographic data: They were no longer complete novices but not yet experts. Their estuary unit included the following concepts: (a) Estuaries are regions where salty water from the ocean mixes with fresher water from inland, forming a gradient; (b) water density varies depending on salinity and temperature; (c) density contrast can drive movements of water masses; and (d) dense salty water tends to flow underneath lower density fresher water.

All of the students who attended class on the day the study was launched participated in the in-class portion of the study (95 students). The whole class was invited to participate in the laboratory portion of the study, which took place outside of class, with 33 students agreeing to participate. Because the laboratory portion included reasoning about data-based visualizations that included color, and eye movements were recorded, to participate in the laboratory portion, participants were required to not have epilepsy or color blindness, and to have normal or corrected-to-normal vision. Participants who participated in the laboratory portion of the study were given a $15 bookstore gift card.

Materials and equipment

Data-based visualizations

We had students reason about the two-dimensional distribution of salinity in a map of the Mediterranean Sea

and adjacent areas, as well as the three-dimensional flow of salty water, from three vertical profiles in combination with the map. Students had learned about relevant processes in class, but in a sufficiently different context that they would have to reason from processes and not simply remember what they had been taught.

Salinity map. The salinity map (Figure 1A) of the Mediterranean Sea and surrounding area showed a strong east–west salinity gradient, from highest salinity in the eastern Mediterranean Sea (39 parts per thousand [ppt]) to intermediate salinity in the western Mediterranean Sea (37 ppt) and lowest salinity in the adjacent North Atlantic (35 ppt). Salinity was shown using color, and a color bar was underneath the map. Latitude and longitude were shown on the western and southern edges of the map. The image was created using the data viewer of the International Research Institute for Climate and Society (Blumenthal et al., 2014).

Data profiles. The data profiles consisted of three north–south vertical data profiles of the water column, successively farther west of Gibraltar (Figure 1D). Each profile contains a high salinity lens depicted on the image as a red-centered bull's-eye. The peak salinity value at the center of the bull's-eye diminishes westward. Note that the high salinity lens is at about 1000m water depth, not at the seafloor or at the sea surface.

Hypothesis arrays

Students were presented with two kinds of hypothesis arrays (verbal and diagrammatic), which provided scaffolding for reasoning about the salinity map and data profiles. The verbal hypothesis array outlined four possible interpretations for the two-dimensional surface salinity gradient across the Mediterranean Sea shown in the salinity map (Figure 1B). The interpretations were written descriptions of two relevant dimensions: if salt or water is added or removed, and the location of that process (e.g., eastern vs. western Mediterranean). The diagrammatic hypothesis array outlined four possible interpretations for the three-dimensional movement of salinity between the Mediterranean Sea and Atlantic Ocean using diagrams (Figure 1C). The interpretations were drawings that represent combinations of two relevant dimensions: location of salinity in the water column and the direction salinity moves.

Scientifically normative explanations of the hypothesis arrays. The scientifically normative model within the array of verbal hypotheses is that an excess of evaporation over precipitation plus runoff in the dry climate of the eastern Mediterranean causes a net removal of fresh water in the east (option (b) in Figure 1B). As fresh water evaporates, salt ions are concentrated in the water remaining in the basin, and the salinity increases. The scientifically normative model within the diagrammatic hypothesis array is that water flowing out of the Mediterranean is saltier than the surrounding Atlantic water, which tends to makes it denser and inclined to sink, and yet it is warmer than the Atlantic water, which prevents it from being dense enough to sink all the way to the seafloor (option (c) in Figure 1C).

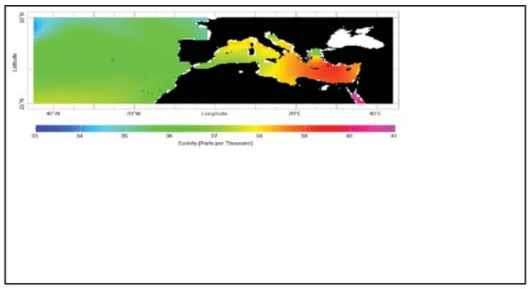
Video, audio, and eye tracking

Interviews were audio- and video-recorded. The video camera was positioned to record the full screen, including the mouse cursor, and could capture where participants were pointing. Participants' eye movements were also analyzed using a Tobii 2.2 eye tracker, paired with a Dell Precision laptop. Participants sat 60 cm away from the monitor. The eye tracker has a data rate of 60 Hz (i.e., 60 gaze points per second are collected for each eye). Visual fixations were quantified as duration of individual periods of looking within each AOI using the Tobii software.

Spatial reasoning assessments

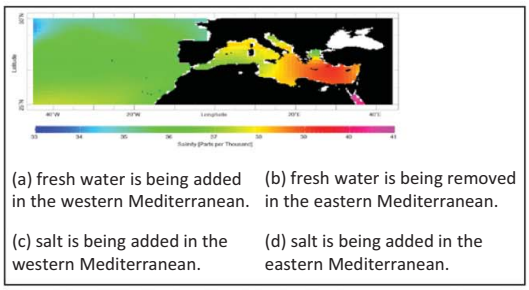
All students were assessed on two separate measures of spatial reasoning: the water-level test (Inhedler & Piaget, 1964) and the geologic model block test (Ormand, Manudca, et al., 2014, 2014b). The water-level test assesses perceptual frames of reference (Uttal et al., 2013). It is untimed and is composed of six drawings of straight-sided bottles oriented 30°, 45°, and 60° to the right and left of upright. Participants indicate where the water would be if the bottle were about half full, with the correct response always being a horizontal line. The geologic model block test assesses a type of spatial visualization referred to as *penetrative thinking* (Kali & Orion, 1996; Ormand, Manudca, et al., 2014; Ormand, Shipley, et al., 2014) or *volumetric thinking* (Shipley & Tikoff, 2016). It is composed of block diagrams showing different geologically possible formations (e.g., a fold) and a vertical slice through the block. Participants identify which of four candidate cross sections is correct for that vertical slice. Although items and cross sections have a geologic foundation, the task is possible to complete without any geologic knowledge by integrating the geometric information presented on the different faces. The current study used a shortened version of the geologic block test; students had 5 minutes to complete eight items that ranged in difficulty.

A. Participants saw a salinity map of the Mediterranean Sea



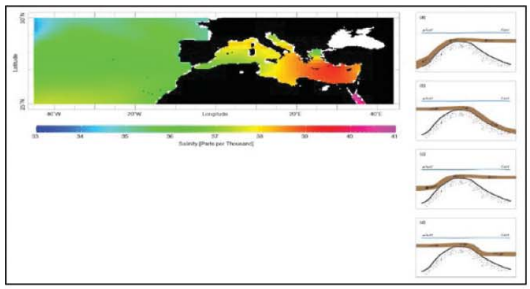
- 20 seconds of free exploration
- Describe image.
- What is image of?
- Where is image?
- What process(es) led to pattern of salinity?

B. Then HALF of the participants were presented with a verbal hypothesis array



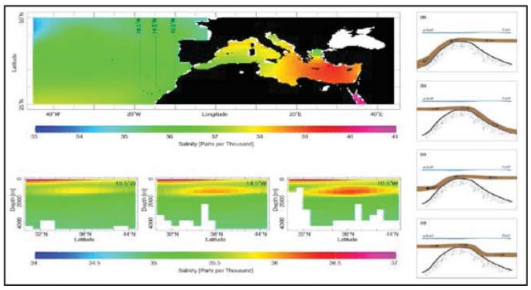
- Options presented simultaneously
- Participant read options aloud.
- Which is correct? Why?

C. All participants were presented with a diagrammatic hypothesis array



- Options were presented one at a time: A, B, C, then D
- Experimenter described each option
- Which is correct? Why?

D. All participants were presented with data and the diagrammatic hypothesis array



- Diagrammatic hypothesis array options remained on screen
- Data profiles were presented one at a time: 10.5°W, 14.5°W, then 18.5°W
- Participant described data profiles.

Summative task:
- Which is correct? Why?

Figure 1. Interview protocol.

Procedure

In-class portion

An investigator attended a lecture session during the second week of the semester to describe the goals and procedures of the study, obtain informed consent, and administer the in-class portion. The in-class portion consisted of two pencil-and-paper spatial reasoning measures: water-level test (Inhedler & Piaget, 1964) and geologic model block test (Ormand, Manudca, et al., 2014, 2014b). Spatial assessments were given to assess the relation between data expertise and spatial reasoning (part of Research Question 2), as well as assess if the volunteers were representative of the classroom as a whole (i.e., did volunteers and nonvolunteers have similar levels of spatial reasoning).

Laboratory portion

Students electing to participate in the out-of-class portion of the study were interviewed and eye tracked individually. They were released from 30 minutes of the

laboratory portion of the class, and interviewed in a quiet room. The interview protocol was designed to move students from relatively simple interpretations (e.g., identification of what the data represented) to more complex reasoning (e.g., inferring 3D structure and processes). Participants made interpretations while thinking aloud in response to a series of guiding questions. The guiding questions, which included the two hypothesis arrays, scaffolded participants' observations, allowing for increasingly complex reasoning about a challenging data set. (See Figure 1 for a summary of the interview.

Participants began by looking at the salinity map (Figure 1A) for 20 seconds so that we could capture looking patterns during undirected exploration of a novel data display. After the initial 20 seconds, participants were asked a series of questions to examine participant conceptions about what the image represented: They were asked to identify what the image showed and where it was located, and to describe the image to a person in another building who could not see the image. After the participant's description, the investigator identified the image as a map of salinity data in the Mediterranean Sea and surrounding area, and explained key features required to read the map (e.g., less salty areas are shown in blue and green); this allowed all participants to be able to make relevant interpretations going forward.

Next, students were asked to hypothesize about the Earth process(es) responsible for the observed map pattern. This task required the participants to make interpretations from the salinity map plus their knowledge of Earth processes (Figure 1A), without any interpretive aids. Half of the participants were then provided with the verbal hypothesis array and asked to pick the correct interpretation (Figure 1B).

The next section of the study examined students' ability to detect and interpret the tongue of salty water that emerges from the Straits of Gibraltar and flows in the middle of the water column westward into the Atlantic. This task is a three-dimensional reasoning task, requiring students to coordinate information from the surface salinity map, hypothesis array options, and vertical data profiles. Students were initially provided with the salinity map and diagrammatic hypothesis array (Figure 1C). They were then asked to identify the correct interpretation by selecting one of the four diagrams. However, the data provided are insufficient to distinguish between the hypothesis array options, because the hypotheses pertain to subsurface processes and the map data are from the sea surface only. Students were asked what type of data they would collect and where they would collect them, as a prelude to reasoning about the data profiles. The data profiles were then provided, giving the participants enough information to complete this task.

Finally, the students completed a summative task. At this point, the students had progressed through the interview, which provided scaffolding with the verbal and diagrammatic hypothesis arrays and provided all required information and data to determine the location and direction of the salt tongue. On the summative task, students were asked for a second time to select one of the four diagrams that best depicted the salt tongue, and to explain why they preferred the selected option.

Analysis of data

Reliability of open-ended data

One cognitive scientist and one geoscientist double-coded the open-ended portions of the interview. The cognitive scientist was an expert in analyzing quantitative and qualitative measures of human behavior, and the geoscientist experienced at scoring student performance on geoscience tasks. Intercoder reliability was initially 88%. The two coders discussed the remaining 12% until 100% intercoder reliability was reached. This indicates a high degree of reliability in our coding schemes designed to capture different elements of students' reasoning about data.

Analysis of eye movement patterns

To analyze eye movement patterns, the image was divided into "areas of interest" (AOIs; Figure 2). The map was divided into three AOIs based on contrasts of the salinity data: the Atlantic, Western Mediterranean, and Eastern Mediterranean. The map color bar was similarly divided into three equal sections that roughly corresponded to the salinity distribution across the three AOIs in the map: low, medium, and high. Map latitude and longitude scales, each hypothesis array option, and each vertical data profile were defined as separate AOIs. The color bar for the data profiles was divided into AOIs in the same way as the map color bar.

Analysis of data expertise

Transcripts and gaze records were analyzed post hoc for observable manifestations of data expertise in action. A single scale was created from those items that had sufficient variability to discriminate between participants, as well as high internal reliability and strong correlations. The relationships between the resulting data expertise scale and performance on a summative data reasoning task and spatial thinking tasks were assessed.

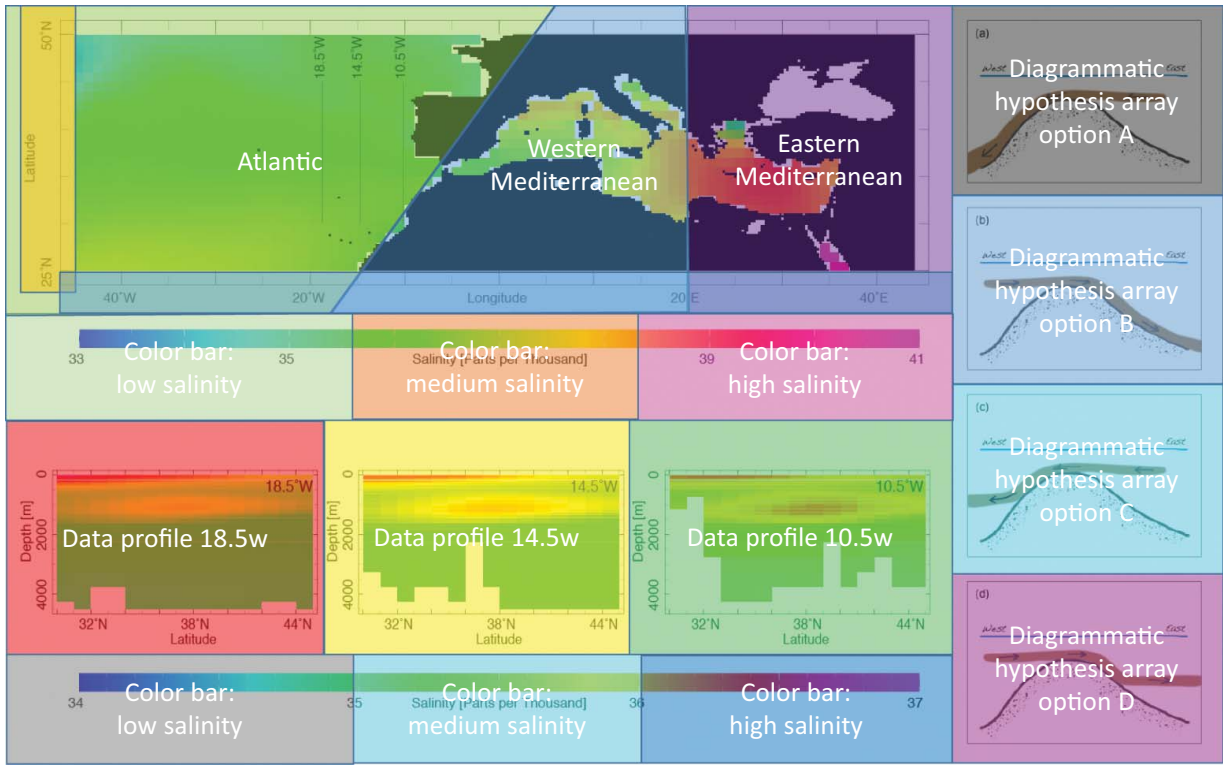


Figure 2. Screen divided into areas of interest on summative task.

Results

In-class portion: Performance on spatial reasoning measures

All students were assessed in class on two spatial reasoning measures: water-level and a modified version of the geologic block model test. Students represented the water-level within 5 degrees of horizontal (the correct orientation) for 56.67% ($SD = 33.33\%$) of the items, and were within 10 degrees of horizontal for 75.83% ($SD = 24.17\%$) of the items. The mean score on the modified geologic block test was 49.13% correct ($SD = 17.25\%$). Participants in the laboratory portion did not differ significantly from the students who took only the in-class portion on either spatial thinking measure ($p > .05$). This suggests that the sample who volunteered for the laboratory portion is not qualitatively different from the overall class.

Laboratory portion: Opportunities to reason about data

The following section is organized by research question. First, we document students' looking and behavioral strategies (research question 1) by characterizing looking patterns from the initial 20 seconds of free exploration and then analyzing students' performance, verbal responses, and spontaneous behaviors, in the order of presentation in the investigator protocol (Figure 1). Then, we examine how observed looking and behavioral strategies relate to one another to form a data expertise continuum (research question 2). Finally, we assess how the data expertise continuum is related to performance and behaviors during a summative task (research question 3).

Research question 1: What is the portfolio of students' looking and behavioral strategies when interpreting complex data-based visualizations, and how common are each of these strategies?

Characterization of looking patterns in the initial 20 seconds of free exploration

To begin the study, participants were presented with just the salinity map of the Mediterranean Sea and adjacent areas and given 20 seconds to freely explore the image. In order to assess looking patterns, we considered the gaze sequences of each participant. To visualize these sequences, we plotted where each participant was looking (defined by the AOIs) along the y-axis and time along the x-axis using Excel, referred to here as a *strategy diagram* (Figure 3). Patterns were identified by reasoning about what looking behaviors we thought would be useful based on our own experience as data interpreters and by looking

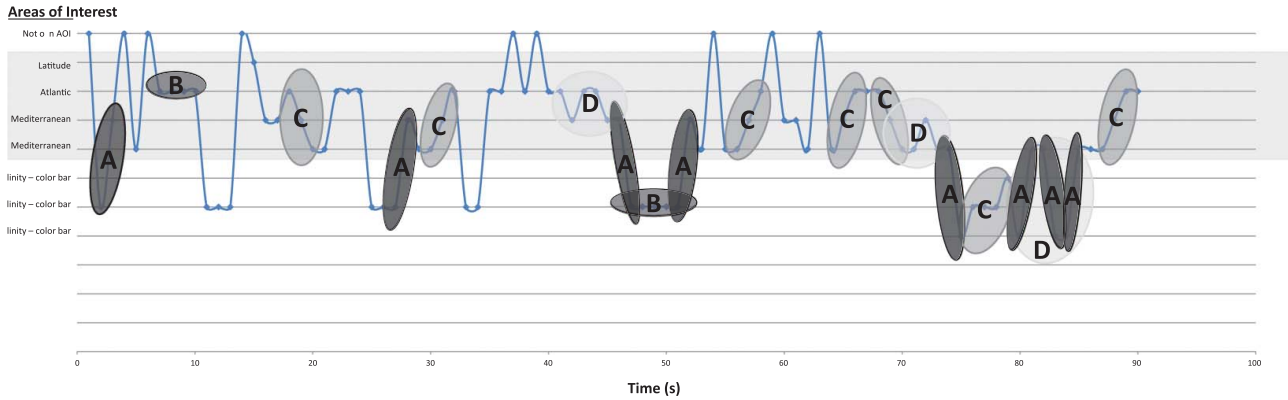


Figure 3. Example of a strategy diagram: our methodology to identify eye movement patterns. This example is taken from a portion of the protocol when only the map is visible. The dots show how the participant's gaze moved from AOI to AOI (vertical axis) across time (horizontal axis). Although presented as continuous data, it is important to note that the AOIs listed along the vertical axis is actually discrete data; thus, the line passing over a listed AOI does not imply the participants' gaze correspondingly traversed passed that AOI. Rather, the participant's gaze moved directly from dot to dot. Characteristic gaze patterns are labeled: A = coordination, B = lingering, C = scanning gradient, D = comparison.

behaviors evident in the strategy diagrams. Using this approach, we identified four looking strategies (coordination, comparison, scanning, and lingering) that were widespread among the participants (Figure 4). A majority of participants also examined cartographic elements. These looking patterns are described below (see Table 1 for a summary of looking pattern frequencies). It is important to note, however, that given the high degree of freedom of this bottom-up-approach, there may be additional patterns that were not identified.

Coordination of the map and color bar. All but one participant coordinated the map with the color bar, which is defined as looking between specific sections of the color bar and the corresponding area of the map (Figure 4A). For example, a participant would be categorized as coordinating the map and the color bar if he or she looked from the high salinity in the color bar to the high salinity in the eastern Mediterranean. There was not a significant difference in direction of coordination: map to color bar vs. color bar to map ($p > .05$). Such coordination between corresponding sections of the map and color bar is significantly more frequent than looking between noncorresponding sections of the map and color bar (e.g., looking from the high salinity section of the color bar to the Atlantic, which has low salinity), even when controlling for the distances between the map and the color bar ($ts(32) > 4.08$, $ps < .05$).

We hypothesize that participants who rarely coordinate the color bar with the map may have a poor understanding of the map's contents because they did not have the opportunity to integrate the important pieces of information; these participants may not have enough map/data skills to know that coordination is a useful strategy. At the other extreme, participants who coordinate the map and color bar many times may be having a hard time understanding, and therefore continue to coordinate the map and color bar in an attempt to figure something out. Participants who coordinate the map and color bar somewhere in between these extremes may have a good understanding; they coordinated the map and color bar once, understood, and so there was no reason to go back and coordinate again. It would take a minimum of coordinating the map and color bar three times to coordinate each section once, and six times to coordinate each section twice, thus providing a framework for cutoff scores. Seven participants coordinated the map and color bar less than three times, seven coordinated more than six times (a total of 14 participants with poor coordination strategies), and the remaining 19 participants coordinated between three and six times, exhibiting what we considered to be a good coordination strategy.

Comparison of any areas of interest (AOIs). A majority (25 of 33) of participants compared areas of the image (Figure 4B). Comparison is defined as looking at a place in the image, moving to a second area, back to the original area, and finally returning again to the second area. To note, participants could compare the map and the color bar in addition to coordinating these areas by looking back and forth between the respective areas of interest the required number of times.

Scanning along the salinity gradient. All participants scanned along the salinity gradient, which is defined as looking at an area with low salinity, then medium salinity, and finally high salinity, or vice versa (Figure 4C).

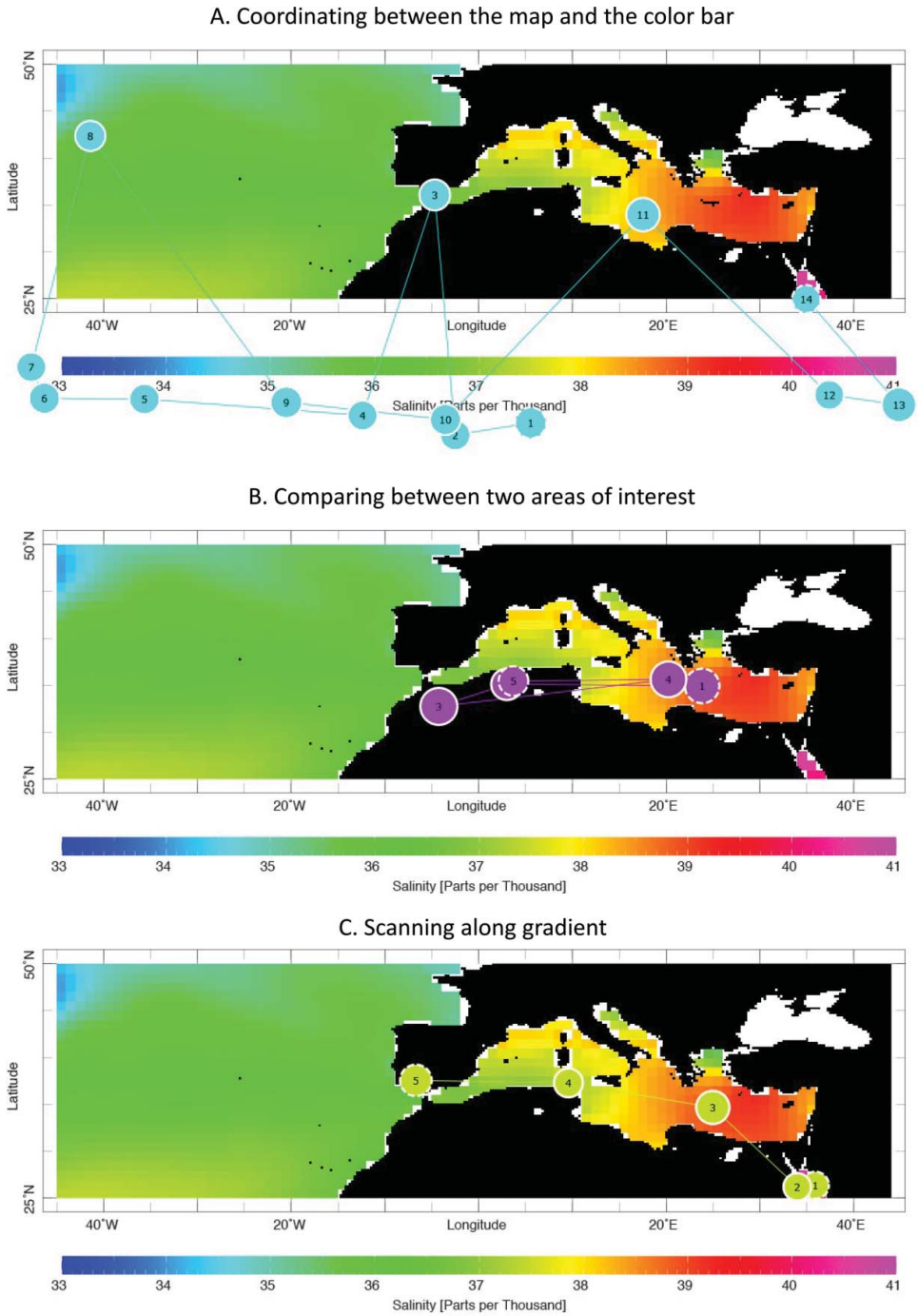


Figure 4. Gaze plots depicting examples of commonly observed looking patterns during participant examination of the map. Each circle represents a fixation; the larger the circle the longer the fixation. The number within each circle shows the order of the fixations, and the line between circles shows the eye movement path. A. Coordinating between the map and the color bar. B. Comparing between two areas of interest. C. Scanning along the gradient.

Scanning along the gradient was the most common looking pattern, making up 15.6% (west to east) and 14.6% (east to west) of total observed looking patterns. Participants did not scan primarily in one direction (e.g., west to east vs. east to west); they were equally likely to scan in either direction ($p > .05$). Participants were not as likely to follow a discontinuous sequence, such as looking from high to low salinity and then to medium salinity, which would reflect looking at the same information but not following along the gradient.

Lingering. Lingering is defined as more than three successive fixations in a given location. All but two participants lingered.

Table 1. All elements identified in interview protocol, their frequency, and how they were included in the data expertise scale.

Individual elements from interview protocol	Percentage of participants exhibited a strong understanding	How element was included in data expertise scale: Not included because a majority of participants either demonstrated a poor understanding or did not explicitly demonstrate their understanding	Not included because a majority of participants demonstrated a strong understanding	Included because participants exhibited a range of responses
Characterization of looking patterns in the initial 20 seconds of free exploration				
Coordination of the map and color bar	58			√
Comparison of any areas of interest (AOIs)	76[a]		√	
Scanning along the salinity gradient	100[a]		√	
Lingering	94[a]	Ranking could not be established because lingering may reflect both strong and weak understandings.		
Cartographic elements	100[a]		√	
Analysis of students' performance and verbal responses				
What is the image?	94		√	
How do you know what image is?	88		√	
Where is image?	27			√
Representation of land in the salinity map	49			√
2D salinity gradient model aligned with data	55			√
Characterization of reasoning about 2D salinity gradient model	See Figure 5 for groupings of shared ideas	Ranking could not be established because all responses included plausibly relevant observations but were nonnormative explanations.		
Verbal hypothesis array performance	6	√		
Initial diagrammatic hypothesis array performance	9	√		
Where to collect data	42			√
Verbally compared data profile(s)	85		√	
Representation of seafloor in the data profile(s)	15[a]	√		
Cumulative scientific argumentation score	36			√
Switched diagrammatic hypothesis array choice after data	85[b]		√	

Note: [a]Percentage of participants who exhibited behavior; [b]64% of students did not discuss the representation of seafloor in data profile; [c]three participants appropriately did not switch from their initially correct answer.

Cartographic elements. During the initial 20 seconds of free exploration of the image, all 33 of the participants looked at one or more cartographic element, with 33 looking at the color bar, 29 looking at latitude scale, and 31 looking at longitude scale. This finding stands in contrast to a companion study of a college population who were complete novices in Earth science data interpretation (Kastens et al., 2016); fewer than 38% of those novices looked at the latitude or longitude scales on maps of geoscience data.

Analysis of students' performance and verbal responses

What is the image?. To begin the study, participants were shown an image of a salinity map of the Mediterranean and surrounding areas (Figure 1A). After 20 seconds of free exploration, they were asked what the image is. Thirty-one of the participants correctly identified the image as a salinity map (although they did not necessarily specify where). The two remaining participants stated they did not know what the image was.

How do you know what the image is?. After the participants were asked to identify the image, they are asked how they knew what the image was. Twenty-nine participants identified at least one aspect from the image to explain how they knew what the image was of (e.g., the salinity label in the color bar). One participant was unable to provide evidence from the image to support her response, because she was uncertain what the image was. The other participant stated, "I just guessed based on the fact with salinity and I guess we've been studying salinity in water recently." There was one participant who initially did not identify the image as a salinity map, but during her response to this question realized the

image was a salinity map and thus was able to provide evidence.

Where is image?. Participants were asked where the image is located. Although all participants at this point had identified the image as a salinity map, 20 participants were unable to identify the map's location. Nine participants correctly identified at least one location within the image. For example, they might have referenced Italy, Spain, or the Mediterranean Sea. Three participants did not respond.

Representation of land in the salinity map. In the salinity map, the color black represented land. Although 31 of 33 participants correctly identified the image as a salinity map, nine held a misconception about what the black represented. Four participants inverted the black and color representations, identifying color as representing land and black representing water. One participant believed black represented areas with no salinity data but did not recognize the black as representing land, two could not identify what black represented, and two did not understand what any of the colors represented.

2D salinity gradient model. Participants were presented with the salinity map (Figure 1), and were asked what process they thought led to the observed pattern. A majority (28 of 33) described plausibly relevant observations and/or Earth processes to explain the observed salinity gradient, whereas five did not respond. These arguments were analyzed in two ways: is the 2D salinity gradient model aligned with data, and characterization of reasoning about 2D salinity gradient model.

Participant responses were coded for whether their model of the salinity gradient was compatible with the data (2D salinity gradient model aligned with data). For example, one participant suggested, "Maybe it [salt] got trapped in there because it's [Mediterranean] really closed off but still part of the ocean." Although this response may account for increased salinity in the Mediterranean as a whole, it does not explain the gradient across the Mediterranean Sea. Eleven participants produced models that did not account for all available data.

Participant responses with shared ideas were grouped together (characterization of reasoning about 2D salinity gradient model). Figure 5 presents the resulting categories of process ideas in the context of the full set of plausible hypotheses identified by the researchers. Figure 5 hierarchically structures how these claims about process could be substantiated by certain observations and reasoning.

One of the most frequent claims argued that salt is trapped in the Mediterranean Sea, such as in the example provided above ($n = 7$). However, this claim does not attach a mechanism to the claim. The other most frequent claim suggested that salt is being increased in the east relative to the other areas ($n = 7$). Here, participants identified runoff from the land as the mechanism, citing either natural sources or human activity.

Five participants noted there is less movement of water in the Mediterranean Sea compared to the Atlantic Ocean, and three participants noted there is less space, both hypothesizing these factors cause the salinity gradient. It is interesting to note that less movement is an inference and less space is actually an observation rather than a process or mechanism; a further reasoning step would be required to convert these ideas into a process capable of causing the observed salinity gradient.

Four participants hypothesized that the water in the western Mediterranean has been diluted by rain. The remaining participant responses combined the ideas described above ($n = 4$) or presented a unique model ($n = 4$).

Although none of these proposed explanations (or "process ideas") exactly matches the scientifically normative explanation, they are all grounded in the observations and all call upon legitimate Earth processes to explain the observations.

Verbal hypothesis array performance. Half of the participants received a verbal array, outlining four possible hypotheses, describing the two-dimensional surface salinity gradient across the Mediterranean Sea (Figure 1B). Three of the response options were compatible with the direction of the salinity gradient observed in the data, whereas one response option was not: "Salt was added to the western Mediterranean" (salt added to the western Mediterranean would have resulted in a gradient from saltier in the west to less salty in the east, the opposite of the empirical observations). Although a majority of the participants chose an incorrect response option, none chose the response option that was incompatible with the data (option C). This suggests the participants were able to construct or recognize a basic interpretation of what is possible given what the data represent. Nine participants indicated that salt was being added to the eastern Mediterranean (option D), five indicated that fresh water was being added to the western Mediterranean (option A), and two correctly indicated that fresh water is being removed from the eastern Mediterranean (option B).

Initial diagrammatic hypothesis array performance. All participants were asked to choose a diagrammatic hypothesis array response option, which had four hypotheses describing the three-dimensional movement

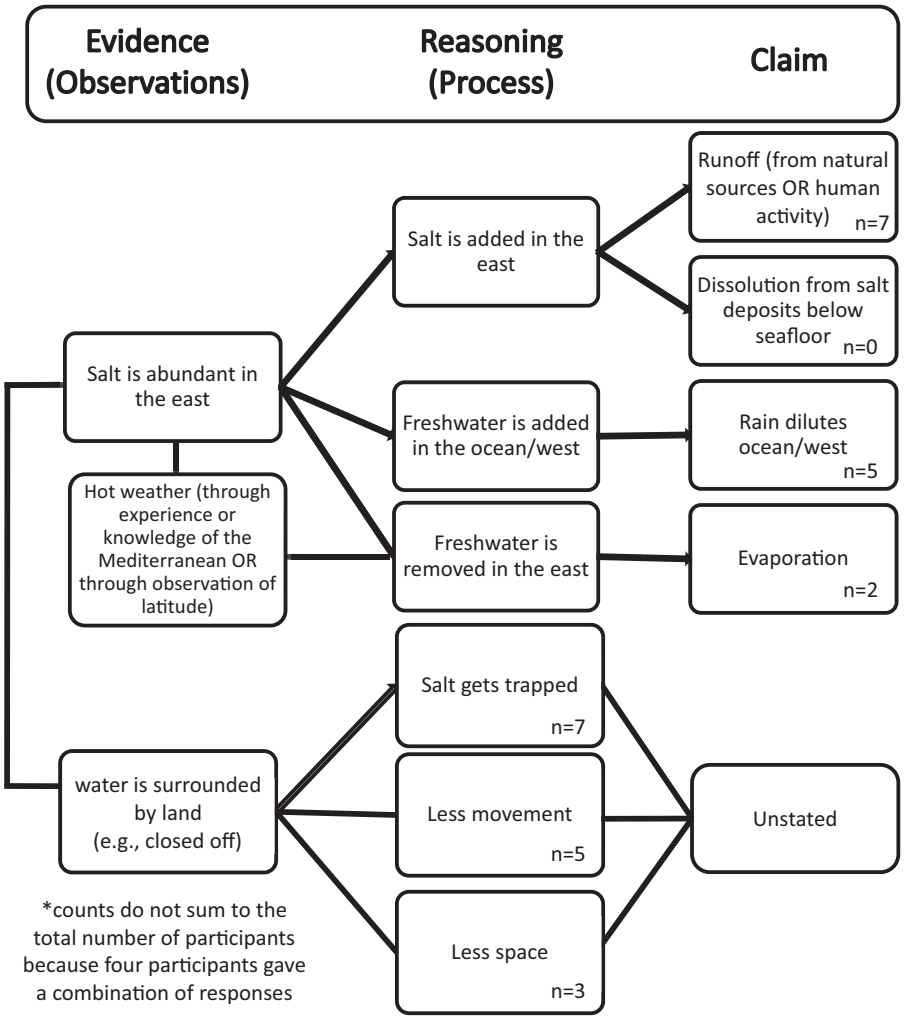


Figure 5. 2D salinity gradient model concept flow chart. Shows potential reasoning from an observation (first column), to reasoning (second column), to claim (third column). *n* represents the number of participant responses that include that response.

of salinity across the Atlantic and Mediterranean Sea (Figure 1C). Participants then saw and described three vertical data profiles (Figure 1D), and were asked once again to choose a diagrammatic hypothesis array option. Figure 6 summarizes response patterns before and after the presentation of data profiles. Notably, 24 participants initially chose a response option in which salinity traveled along the ocean floor (options A or B), with 16 of these participants expressing their reasoning that salt water is heavier than fresh water. This reasoning is aligned with what the participants were learning about in class regarding the flow of salty and fresh water in estuaries, but this happens not to be true for the more complex case of the Mediterranean. The most frequent first response ($n = 19$) was incorrect response option B, which depicts salt flowing from the Atlantic into the Mediterranean Sea along the seafloor.

Where to collect data. After reasoning about the first diagrammatic array, participants were asked where they would collect data to help them choose the correct response option. A scientifically normative response would indicate that samples should be taken at multiple depths and at multiple locations along an east–west transect. Fourteen participants correctly identified locations within both the vertical water column (e.g., middle of water column) and horizontal dimension. For example, one participant stated, "You would want to collect the salinity and the correlation of salinity and depth; that will narrow it down and then. I don't really know how you can test direction. I guess you can measure it at different data points that are west versus east." Eighteen participants identified only one location within the vertical water column (e.g., the seafloor). One participant did not respond.

Verbally compared data profiles. Participants were sequentially presented with three data profiles, and asked to describe each data profile. When describing the second and third data profile, 28 participants made explicit

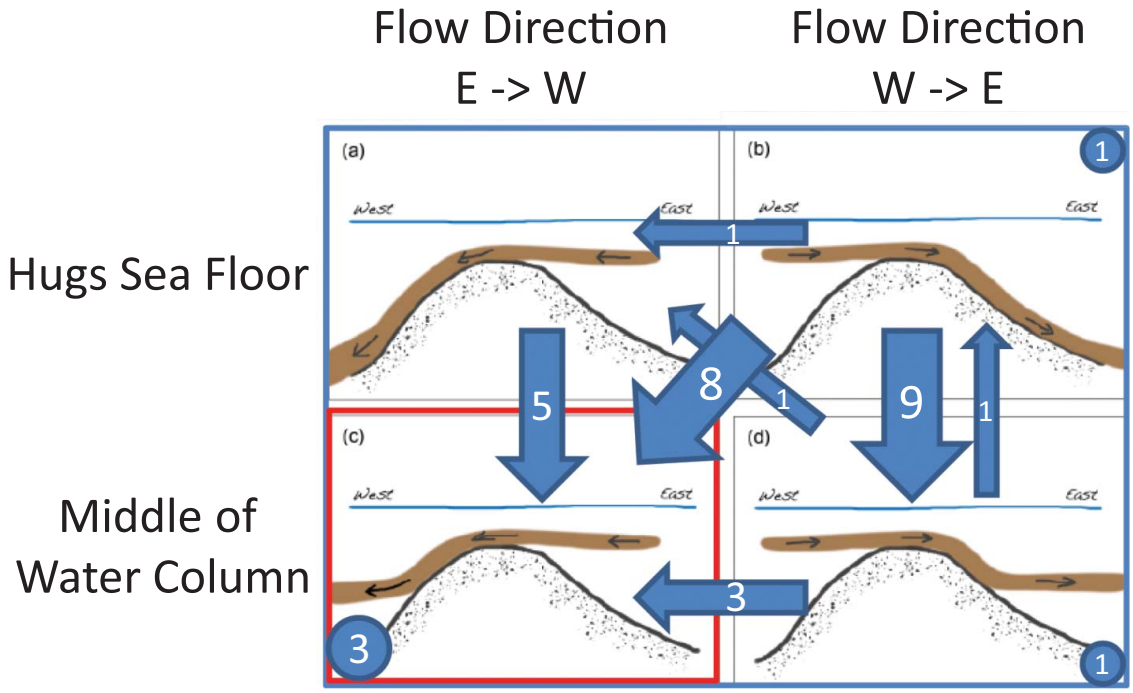


Figure 6. Pattern of responses before and after the presentation of data on the diagrammatic hypothesis array. Arrows represent movement from initial response to response after the presentation of data. Circles represent participants who did not change their response. Numbers indicate how many participants exhibited each pattern. The red box marks the correct answer.

verbal comparisons between data profiles. For example, one stated, "It [the second data profile] is basically very similar to the other graph except the red spot is less pronounced. It is a bit more faded which means that there is less salinity in that area."

Representation of seafloor in the data profile(s). In the data profiles the seafloor and sub-seafloor were represented by the color white. In their description of the data profiles, 12 participants explicitly demonstrate either an understanding (five) or a misconception (seven) of what the white color represented. An example of a misconception includes one participant's response: "The very bottom is white, and I'm guessing that that means that it is either data hasn;t been collected or it's just not salty at all." The remaining 21 participants did not discuss the white color or mention a white form in their response.

Cumulative scientific argumentation score. Participants were asked to hypothesize why a salinity gradient was present in the map, as well as to provide an initial hypothesis for the direction and location in the water column in which salinity is distributed (steps 1A and 1C of Figure 1). In each case, participants provided reasoning to explain their hypotheses. Strong scientific argumentation is categorized as having a claim, evidence, and reasoning that links the claim with the evidence through a scientific principle or model, and potentially a rebuttal for alternative explanations (McNeil & Krajcik, 2012).

Participants' responses were coded for the presence of a claim, evidence, reasoning, and rebuttal, as defined by (McNeil & Krajcik, 2012). A claim is defined as a statement of conclusion that answers an original question or problem. Evidence is defined as empirical observations or data that supports the claim. The data/information needs to be relevant, accurate, and sufficient to support the claim. Whereas (McNeil & Krajcik, 2012) identified one type of reasoning, we differentiate between two types of reasoning. The lower-level reasoning invokes a mechanism, process, principle, or model as defense for the claim. More sophisticated reasoning explains *how* the mechanism, process, principle, or model serves to logically connect or link data/evidence with the claim. For example, lower-level reasoning about the vertical data profiles might include a statement that saltier water is moving along the middle of the water column (claim), the observation that the saltier water in the vertical profiles is in the middle of the water column (evidence), and a relevant process statement that saltier water is heavier than fresher water. Higher-level reasoning would further explain that because saltier water is heavier than fresher water it would therefore sink within the water column. Rebuttal is defined as recognizing, rejecting, and providing reasoning for rejecting an alternative explanation. A rebuttal would include spontaneously identifying that

saltier water would not be towards the top of the water column because saltier water is heavier than fresh water (rebuttal).

Responses were scored on a 0–5 point scale: having no claim (0); having a claim (1); having claim plus evidence (2); having claim, evidence, plus reasoning that states a relevant principle/model/ mechanism/process (3); having claim, evidence, plus reasoning that provides a logical linkage (4); and having a claim, evidence, reasoning with linkage, plus rebuttal (5). If a response had some elements, but not all, of a given score, the participant was assigned the lowest score for what he or she provided. For example, a response containing a claim, evidence, and reasoning would be scored as a 3. However, if the response only contained a claim and reasoning, but no evidence, it would be scored as a 1 (claim only). This occurred on 10% of responses. Twelve participants had scores of 3 or higher, indicating they had strong scientific reasoning, whereas the remaining 21 participants had scores of 0 to 2, indicating they did not have strong scientific reasoning. Note that the scientific argumentation score did not include responses to the summative task.

Switched diagrammatic hypothesis array choice after data. After the presentation of the three data profiles, participants were asked to choose from the diagrammatic hypothesis array response options for a second time. Three participants correctly kept the correct response option (C), two kept their incorrect response, and 28 switched their response choice. This suggests these participants are willing to change their mental model to fit new data. Notably, 26 of 28 switched to a better response (from bottom to middle of water column, from W→E to E→W direction of flow, or both). Of those participants who initially chose response option B, eight switched to correct answer C (salinity flows from Mediterranean to Atlantic in middle of water column), whereas nine switched to incorrect response option D (salinity flows from Atlantic to Mediterranean along seafloor).

Research Question 2: Within this portfolio, which looking and behavioral strategies are associated with greater and lesser levels of data expertise?

Data expertise continuum

Emergent in the interview was the opportunity to explore a new construct: data expertise. Students progressed through the interview, ending in a summative task, a final question in which they were asked to choose a diagrammatic hypothesis array option and justify that choice for a second time after seeing the three data profiles. Each of the behaviors and understandings exhibited throughout the interview could have contributed to competence on this summative task. Thus, instances of reasoning about data were identified throughout the interview (excluding performance on, and behaviors during, the summative task), and examined for how well they fit together as a single scale (Table 1).

Development of data expertise scale. Two elements were not considered for inclusion in the data expertise scale because it was not clear if the behavior reflected a strong versus weak understanding: lingering in AOIs and characterization of reasoning about 2D salinity gradient model. Lingering in an AOI could reflect the student was having trouble understanding, was having higher-level thoughts about that AOI, or the mind wandered altogether. In the characterization of reasoning about 2D salinity gradient model, we were unable to rank the accuracy of responses, because they all included plausibly relevant outcomes and were nonnormative scientific explanations. We note, however, most of the participants exhibited these behaviors, suggesting their inclusion in the scale might not have added useful information.

A number of elements were not included in the data expertise scale because of their limited variability. For the following elements, a large majority of participants demonstrated a strong understanding (e.g., over 94% were able to correctly identify the image as a salinity map): *what is the image, how do you know what the image is, switched diagrammatic hypothesis array choice after data*, and *compared data profiles*. For the following elements, a large majority of participants demonstrated a weak understanding (e.g., only three chose the correct response option for the initial diagrammatic hypothesis array) or a majority did not explicitly demonstrate their level of understanding: *diagrammatic hypothesis array choice, verbal hypothesis array choice*, and *representation of the seafloor in the data profile*.

Between these two extremes, however, there were six elements of data expertise that had sufficient variability to discriminate among participants. These were *coordination of the map and color bar, representation of land in salinity map, 2D salinity gradient model aligned with data, where to collect data, scientific argumentation*, and *where the image is*. For these elements, participants were scored either a 0 to indicate they did not understand/did not do well on that element or a 1 to indicate they did understand and did well (as defined above). See Table 2 for the distribution of performance on these six elements.

A Rasch analysis (Rasch, 1960/1980) was conducted to assess the appropriateness of summing these individual elements into a single data expertise scale. The six items had an internal reliability of 0.82, which suggests that they are consistent in measuring the same construct (above 0.7 is acceptable). The test statistic (infit zstd),

Table 2. Distribution of participant responses for each element considered for inclusive in the data expertise scale.

ID	Coordination of the map and color bar	Representation of land in the salinity map	2D salinity gradient model aligned with data	Where to collect data	Cumulative scientific argumentation score	Where is image?	Raw data expertise score
9	1	1	1	1	1	1	6
20	1	1	1	1	1	1	6
17	1	1	1	1	1		5
30	0	1	1	1	0	1	4
32	1		1	1	1	0	4
19	1	1	1	1	0	0	4
31	1	1	0	1	1	0	4
33	1	1	0	0	1	1	4
6	1	0	1	1	1	0	4
1	1	1		0	0	1	3
3	0		1	1	1		3
11	0	1	0	1	0	1	3
13	1	1		0	0	1	3
15	0	1	1	1	0	0	3
24	1	1	0	0	0	1	3
2	1	1		0	0	1	3
18	0	0	1	0	1	0	2
23	0	1	1	0	0	0	2
4	1	1	0	0	0	0	2
8	0	0	1	1	0	0	2
14	1	0	1	0	0		2
16	0	0	1	0	1	0	2
25	0	1	0	0	0	1	2
26	1		1	0	0	0	2
29	0		1	1	0	0	2
12	0			1	1	0	2
22	0	0	0	0	1	0	1
7	1		0	0	0	0	1
21	1				0	0	1
10	1		0	0	0	0	1
28	1	0	0	0	0	0	1
5	0			0	0	0	0
27	0	0	0	0	0	0	0

Note: Each column represents a potential element of data expertise. Each row represents an individual participant. Dark gray (score of 1) indicates the participant demonstrated strong understanding within that element. Light gray (score of 0) indicates the participant demonstrated a weak understanding within that element. White (no score) indicates where participants' understanding was not explicitly demonstrated. The participant with the highest data expertise score is located in the top row and extending downward to the participant with the lowest expertise score in the bottom row.

which can be interpreted as a z-score, was 0.8. This means that the items were within 0.8 standard deviations from the model, suggesting a strong fit (an infit zstd between –2.0 and 2.0 is acceptable).

Points were summed to yield a data expertise scale (0–6). See Table 2 for the distribution of data expertise across participants. Scores were converted to percentages, so instances in which participants did not explicitly demonstrate their level of understanding did not influence their score. Importantly, participant scores had a normal distribution, ranging from 0% (demonstrated a weak understanding on all elements) to 100% (demonstrated a strong understanding on all elements; Figure 7). Neither spatial thinking task was predictive of data expertise.

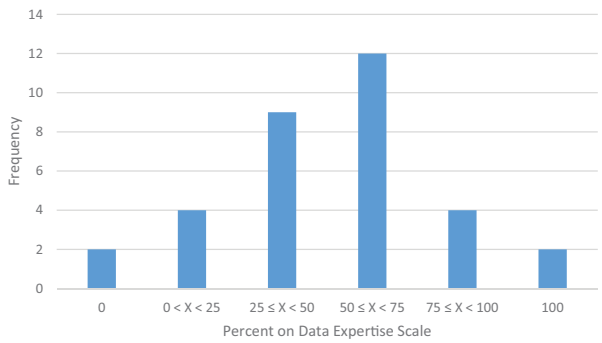


Figure 7. Distribution of participants on data expertise scale.

Research Question 3: Do students who have higher levels of data expertise perform better or differently than students with lower levels of data expertise when completing interpretations that are more challenging?

As noted previously, the data expertise scale was developed by examining instances throughout the interview, which led to a summative task. The summative task was to identify the correct diagrammatic hypothesis array option after viewing all available data and information. Recall that 19 participants initially chose diagrammatic response option B, with eight switching to the correct option C and nine switching to incorrect option D. Of these participants who initially chose B and then switched, participants with a

high data expertise score were significantly more likely to switch to the correct response option compared to those with a low data expertise score ($t(15) = 2.99$, $p = .009$). For this analysis, high and low data expertise was determined using a median split.

Data expertise is inversely correlated with how long it took to respond to the summative task, choosing a diagrammatic hypothesis array option after being presented with all available data ($r = .35$, $p = .04$); as data expertise increases, the length of time to respond decreases. This correlation is true of both how long it took participants to begin speaking and how long after they began speaking until they chose a response option (Table 3). Participants with lower data expertise took on average four times longer to begin speaking and nine times longer to choose a response after beginning to speak.

There was also a difference in looking patterns on the summative task between participants with low data expertise and high data expertise. High data expertise is correlated with increased looking time at the data when choosing a diagrammatic response option after seeing the vertical profiles ($r = .45$, $p < .01$) and decreased looking time at the color bar (mean time ($r = .36$, $p = .02$) compared to low data expertise. Table 4 shows mean looking times and illustrates this overall pattern with data from two participants as examples, both of whom ultimately chose response option C, but one with high and one with low data expertise. The two example "heat maps" show the accumulated fixation duration relative to the total viewing time (referred to as "relative duration"). This accounts for the different overall looking times of individual participants. No normalization filters were used. Overall, there is a tendency for participants with high data expertise to spend more time looking at data and for participants with low data expertise to spend more time looking at the color bar. Neither spatial thinking task was predictive of behaviors during the summative task.

Discussion

The current study provides a richly nuanced portrait of students' reasoning about simple and more complex tasks using a data-based visualization from professionally collected data (research question 1). Many of these students seem to be in an interesting transitional state. They are successfully decoding and describing data, successfully accessing relevant bits of knowledge about the Earth systems, successfully recognizing significant patterns in the visually available data, and successfully generating some claims/evidence/reasoning (e.g., Figure 5). Yet they are not necessarily pulling all of these elements together to generate scientifically normative explanations. In terms of scientific knowledge integration, these students have many of the ingredients, but they are not necessarily successfully combining them.

Table 3. Mean time to respond to diagrammatic hypothesis array after data based on a median split.

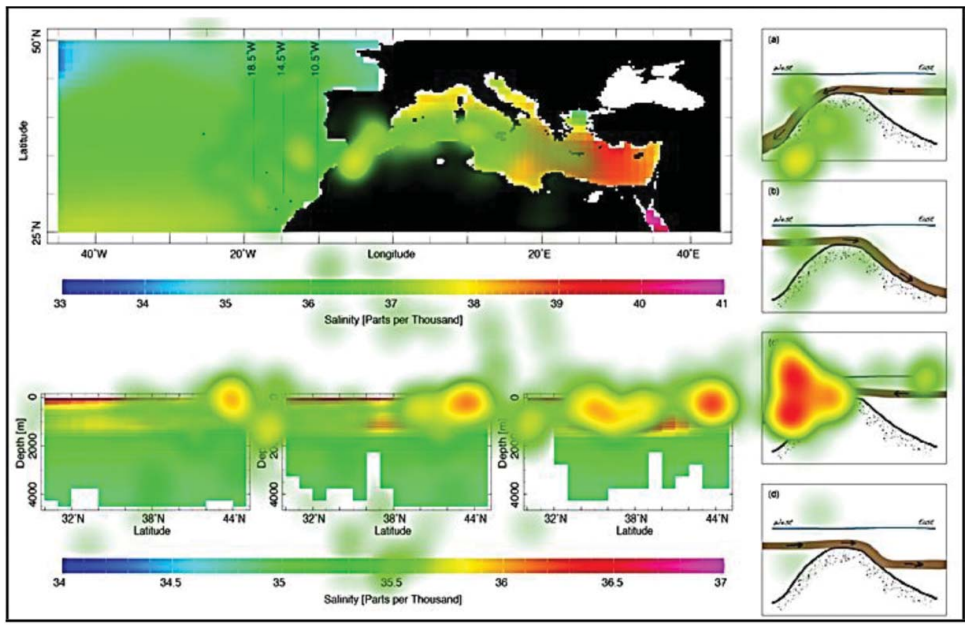
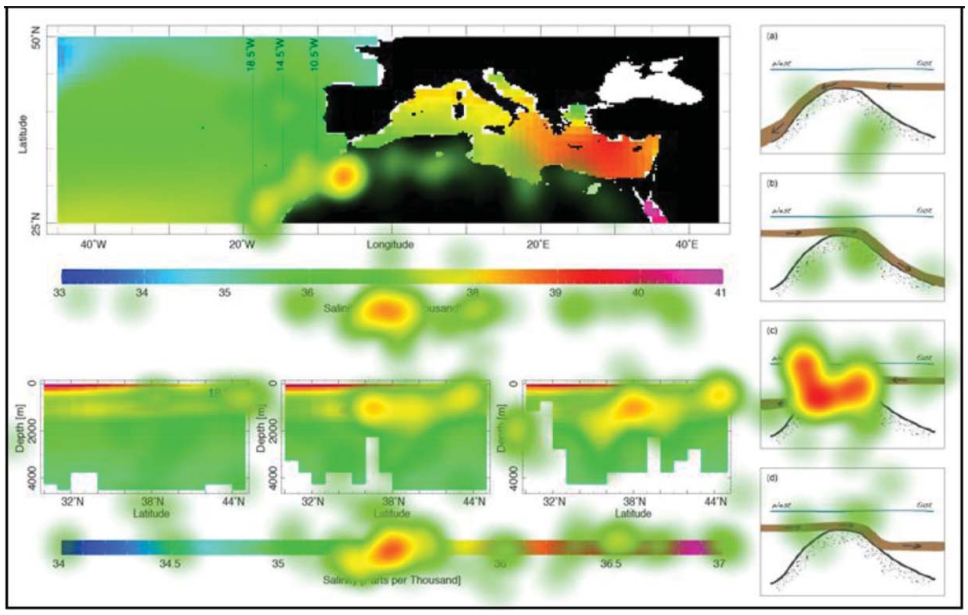
	Time to begin speaking(s)	Time after speaking to pick a response option(s)	Total time from question to response(s)
High data expertise score	3.8	4.4	8.2
Low data expertise score	16.2	39.2	55.4

This population of students is no longer complete novices; they have substantial knowledge about how to read scientific visualizations and reason about relevant processes, and are willing to change their mental model in light of new data (Figure 6). This is in stark contrast to other studies examining novices' ability to read data-based visualizations. For example, many novices are unable to successfully identify what a data-based visualization represents (Cid, Lopez, & Lazarus, 2009; Swanson & Kastens, 2011), do not look at important cartographic elements and labels (Kastens et al., 2016), and focus on irrelevant but perceptually salient aspects of visualizations (Lowe, 1999) such as objects included for scale (Coyan, Busch, & Reynolds, 2010; Morton, 2010). Additionally, many students are unwilling to accept the scientifically normative model when they already have a naïve explanation of the phenomenon (Chi, 2005).

Data expertise continuum

Conceptualizing skill in reasoning and interacting with data along a data expertise continuum is important both theoretically, in characterizing a progression from novice to expert, and practically, in developing teaching and assessment materials. The current study identified six looking and behavioral strategies within our interview sessions that may be related to a data expertise scale (research question 2). From the six elements that emerged in this study, we suggest the following broader habits of mind that may generalize across a wider range of data tasks: (a) coordinate data with a diagrammatic model, (b) compare and contrast across data representations from different times or places, and (c) extract 3D structure from multiple 2D representations. Such skills may be thought of as being a part of representational competence, which is the set of skills required to use external representations for problem solving (Kozma & Russell, 1997; Nathan, Stephens, Masarik, Alibali, &

Table 4. Average looking times for participants with high vs. low data expertise during summative task, with examples from one high and one low data expertise individual.

Data expertise score	Mean looking time in seconds (standard deviation)		Example of looking patterns
	Data section	Color bar	
High data expertise	23.62 (5.46)	9.08 (5.15)	
Low data expertise	17.31 (2.34)	14.24 (2.43)	

Note: Duration in heat map is represented as a scale from green (relatively shorter duration) to red (relatively longer duration). Although both participants looked along the salinity gradient and chose the same correct response, the participant with high data expertise spent a greater proportion of time looking at the data and the participant with low data expertise spent a greater proportion of time looking at the color bars.

Koedinger, 2002; Wu, Krajcik, & Soloway, 2001). Such tasks are by no means limited to this particular data set, and may represent a subset of the metrics that may be useful in developing a general data expertise scale. Additionally, certain data reasoning tasks and skills may be diagnostic only at limited locations on the data expertise continuum. For example, in the current study, most participants switched their response after seeing additional data, most compared the data profiles, and all looked at cartographic elements. Because there was little to no variation, these particular behaviors were not diagnostic for this intermediate population. However, these behaviors might have been diagnostic for students lower on the data expertise continuum.

Our final research question examined how the data expertise continuum scale related to performance and behaviors on a summative task (research question 3). Higher data expertise was related to switching to the normative scientific explanation on the summative task, providing validity to the data expertise scale. Participants with weaker data expertise took longer to begin and to complete their response on the summative task. Participants with weaker data expertise also spent relatively more time looking at the color bar, whereas participants with higher data expertise spent relatively more time looking at the data. Taken together, this suggests that participants with stronger data expertise may possess better and more efficient strategies for reasoning.

Limitations of the study

While the participants reasoned about the data-based visualizations, their eye-movements were recorded. This eye-tracking methodology was useful in identifying common looking patterns and strategies. However, there were also a number of limitations. Where a participant looks is likely influenced by task demands (e.g., What has the participant been asked to do, and what are they talking about?). Where participants loos may also not correspond to what they are thinking about in that moment. For example, we noticed that participants spent a lot of time fixating on the middle of the screen (which was straight ahead), perhaps drifting into a series of thoughts or no thoughts at all. Thus, researchers using eye-tracking methodologies should take care not to place theoretically important AOIs in the middle of the screen.

The aim of the hypothesis array is to scaffold student thinking around relevant dimensions. Researchers and educators interested in using hypothesis arrays should consider how the structure of the hypothesis array may influence students' reasoning. For example, the verbal hypothesis array used in the current study included three response options that suggested something was added and only one response option that suggested something was taken away. This could have led students to engage in a test-taking strategy (do not chose the option that is different from all the rest) rather than reasoning about the data. Finally, the current study included only female students in one class at one institution. Thus, it is possible that different findings would emerge among a broader student sample.

Educational implications

In guiding students from complete novice to expert, it is essential to know what skills they bring to bear as they progress along this continuum. In this way, instructors and curriculum developers can prepare appropriate and meaningful lessons for targeted skill levels. The current study identified an intermediate population of students and characterized their reasoning.

One area of difficulty these students faced was being able to think about something that was not there. For example, in the verbal hypothesis array, only two participants chose the response option that something was removed; the majority of participants chose response options that indicated that either fresh water or salt was added. Additionally, 25% to 30% of the participants did not understand that the black in the salinity map and the white in the data profiles, respectively, represented an absence of data. Having difficulty reasoning about the absence of something is consistent with findings from the field of perception, showing it is easier to locate an object based on the presence of a feature rather than its absence (Treisman & Gormican, 1988; Treisman & Souther, 1985). People similarly reason differently about missing versus present concepts and actions (Spranca, Minsk, & Baron, 1991). Thus, it may be particularly useful to explicitly identify absences of phenomena within a given data set, and actively scaffold student reasoning around the reason for its absence.

Another stumbling block for these students arose when the available data were not sufficient to adequately constrain the interpretation. In the current study, all of the participants chose and defended a response option when they were first shown the diagrammatic hypothesis array, despite the fact that they had not yet been provided with enough information to determine the correct answer. The participants did not recognize that the data they had at the time of this question was ambiguous. A better response would have been to say there was not enough information available to answer the question, but no student did so. Students may have been habituated to expect that all school questions should be answerable. Future research should examine the utility of first

asking students if a question can be answered given the data in hand. Such an approach may promote thinking about what kind of data is required to reason about scientific processes, and could be incorporated into other similar approaches. For example, the knowledge integration approach advocates for scaffolding students to evaluate evidence, articulate their thinking, and connect ideas (e.g., Clark & Linn, 2003/2009; Linn, 2000; Linn, Clark, & Slotta, 2003), and the backward fading scaffolding approach explicitly scaffolds students to identify alignments and misalignments between evidence and claim (Slater, Slater, & Lyons, 2010; Slater, Slater, & Shaner, 2008).

Ability to generate explanatory models, to compare the behavior of a model with the behavior of the Earth as captured in data, and to evaluate the veracity of competing models is an essential skill for a geoscientist—but it is difficult to teach or assess. Although the students in the current study did not produce scientifically normative models, the course professor noted that they nevertheless generated substantially more sophisticated reasoning and models of Earth processes than was typical. We believe our use of hypothesis arrays as part of our experimental design facilitated and scaffolded deeper reasoning than students could otherwise have produced. Whereas we used the hypothesis array in a research study, the same technique can be used in curriculum design and instruction (Kastens et al., 2015; Mayer et al., 2002).

Directions for further research

We have identified a broad portfolio of behaviors and strategies that were exhibited by students working at the outer limits of their data interpretation expertise, and have quantified which of these were more or less common and which were more or less strongly associated with data expertise. Some of these behaviors and strategies are presumably idiosyncratic to our particular stimuli and tasks. It would be valuable to do a similar analysis for other data types and parts of the Earth system. By comparing across studies, we could then deduce which are of broad or universal importance, and could craft instructional strategies to nurture the most valuable.

Future research should examine if a single data expertise scale is generalizable to multiple fields, or if there are multiple threads within different content domains (Wilson, 2005, 2009). If one achieves data expertise in one domain (e.g., geology), does the ability to reason about data transfer to other fields (e.g., biology, economics)? If not, are there specific data reasoning skills that transfer to other disciplines, and would they transfer only at specific points on the data expertise continuum? For example, there may be basic data-reading skills that transfer (e.g., reading labels), whereas more complex practices do not. Finally, is data expertise a stable construct within a field but across different kinds of data representation?

When a geoscientist looks at the diagrammatic models, such as those in Figure 1D with a static image showing process, the expert is likely to visualize (or otherwise mentally represent) the motion or change. The expert has the ability and the habit of mind of turning a static representation into a dynamic model, of "running" a conceptual model in his or her head, and then comparing the behavior of the model with the behavior of the Earth as captured in data. This is what we hoped the students were doing. Certainly, they were engaging with the alternative models, contrasting the models with one another, comparing the models with the data. However, the current study does not tell us whether they actually reached the point of creating and using runnable mental models. If students were generating and running models to compare to data, this would be a significant step toward developing the habit of changing one's conceptual model when faced with data that disagree with the model. Developing research protocols and then classroom assessments to detect when students are using runnable mental models remains a challenge for another day.

Acknowledgments

We would like to thank the instructors for allowing us into their classrooms and providing their intuitions regarding students' ability levels; Linda Pistolesi, for conducting the eye-tracked interviews; Frances Straccia, for working on coding of the students' reasoning; the participants, for volunteering their time; and the reviewers, for their insightful feedback.

Funding

This research was supported by National Science Foundation Grants 1138616 (Columbia University), 1331505 (Education Development Center), and 1138619 (Temple University) as part of the Fostering Interdisciplinary Research in Education (FIRE) program, the National Science Foundation Grants SBE-0541957 and SBE-1041707, which support the NSF-funded Spatial Intelligence Learning Center, SBE-1640800, and the Institute of Education Sciences Grant R305B130012 as part of the Postdoctoral Research Training Program in the Education Sciences.

References

American Association for the Advancement of Science (AAAS). (2008). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

Blumenthal, M. B., Bell, M. A., del Corral, J. C., Cousin, R., & Khomyakov, I. Y. (2014). IRI data library: Enhancing

accessibility of climate knowledge. *Earth Perspectives Transdisciplinary Enabled*, *1*(19), 1–12.

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *Journal of Learning Sciences*, *14*, 161–199. doi:10.1207/s15327809jls1402_1

Cid, X. C., Lopez, R. E., & Lazarus, S. M. (2009). Issues regarding student interpretation of color as a third dimension on graphical representations. *Journal of Geoscience Education*, *57*(5), 372–378. doi:10.5408/1.3559675

Clark, D., & Linn, M. C. (2009). Designing for knowledge integration: The impact of instructional time. *Journal of Education*, *189*(1–2), 139–158. doi:10.1177/0022057409189001-210 (Reprinted from *Journal of the Learning Sciences, 12*[4], 451–494 [2003])

Cleland, C. (2001). Historical science, experimental science, and the scientific method. *Geology*, *29*, 987–990. doi:10.1130/0091-7613(2001)029%3c0987:HSESAT%3e2.0.CO;2

Coyan, J., Busch, M., & Reynolds, S. (2010). Using eye trackers to evaluate the effectiveness of signaling to promote the disembedding of geologic features in photographs. In A. Frick, D. Nardi, & K. Ratliff (Eds.), *Spatial cognition 2010: Doctoral colloquium SFB/TR8 report No. 025-07/2010* (pp. 15–19). Bremen, Germany: Universitat Bremen and Universitat Freiburg.

Creilson, J. K., Pippin, M., Henderson, B., Ladd, I., Fishman, J., Votapkova, D., & Krpcova, I. (2008). Surface ozone measured at GLOBE schools in the Czech Republic: A demonstration of the importance of student contribution to the larger science picture. *Bulletin of the American Meteorological Society*, *89*, 505–514. doi:10.1175/BAMS-89-4-505

Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist*, *33*, 109–128. doi:10.1080/00461520.1998.9653294

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Greenberg, J., & Walsh, K. (2012). *What teacher preparation programs teach about K–12 assessment: A review* (Rev. ed.). Retrieved from http://www.nctq.org/p/publications/docs/assessment_report.pdf

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.

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. doi:10.1080/09500690701506945

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. New York, NY: Harper & Row.

Kali, Y., & Orion, N. (1996). Spatial abilities of high-school students in the perception of geologic structures. *Journal of Research in Science Teaching*, *33*, 369–391. doi:10.1002/(SICI)1098-2736(199604)33:4%3c369::AID-TEA2%3e3.0.CO;2-Q

Kastens, K. A. (2009). *The meaning of "meaning": Causes & consequences*. Retrieved from http://serc.carleton.edu/earthandmind/posts/meaning_meaning.html

Kastens, K. A. (2012). *Is the fourth paradigm really new?* Retrieved from http://serc.carleton.edu/earthandmind/posts/4thpardigm.html

Kastens, K. A. (2015). Philosophy of Earth science. In R. Gunstone (Ed.), *Encyclopedia of science education: Dordrecht, Heidelberg* (pp. 354–357). New York, NY: Springer.

Kastens, K. A., Agrawal, S., & Liben, L. S. (2009). How students and field geologists reason in integrating spatial observations from outcrops to visualize a 3-D geological structure. *International Journal of Science Education*, *31*, 365–394. doi:10.1080/09500690802595797

Kastens, K. A., & Ishikawa, T. (2006). Spatial thinking in the geosciences and cognitive sciences. In C. Manduca & D. Mogk (Eds.), *Earth and mind: How geoscientists think and learn about the complex Earth* (pp. 53–76). Boulder, CO: Geological Society of America.

Kastens, K. A., Krumhansl, R., & Baker, I. (2015). Thinking big: Transitioning your students from working with small, student-collected data sets toward "big" data. *The Science Teacher*, *82*(5), 25–31. doi:10.2505/4/tst15_082_05_25

Kastens, K. A., & Mandua, C. A. (2012). Fostering knowledge integration in geoscience education. In K. A. Kastens & C. Manduca (Eds.), *Earth & mind II: Synthesis of research on thinking and learning in the geosciences* (pp. 183–206). Boulder, CO: Geological Society of America.

Kastens, K. A., Shipley, T. F., Boone, A., & Straccia, F. (2016). What geoscience experts and novices look at, and what they see, when viewing data visualizations. *Journal of Astronomy & Earth Science Education*, *3*, 27–58.

Kozma, R., & Russell, J. (1997). Multimedia and understanding: Expert and novice responses to different representations of chemical phenomena. *Journal of Research in Science Teaching*, *34*, 949–968. doi:10.1002/(SICI)1098-2736(199711)34:9%3c949::AID-TEA7%3e3.0.CO;2-U

Linik, R. J. (2015). *After the confetti: Updates on 2014's ultimate science fair champs*. Retrieved from http://iq.intel.com/after-the-confetti-updates-on-2014s-ultimate-science-fair-champs/

Linn, M. C. (2000). Designing the knowledge integration environment. *International Journal of Science Education*, *22*(8), 781–796. doi:10.1080/095006900412275

Linn, M. C., Clark, D., & Slotta, J. D. (2003). WISE design for knowledge integration. *Science Education*, *87*, 517–538. doi:10.1002/sce.10086

Lombardi, D., Sinatra, G. M., & Nussbaum, E. M. (2013). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction*, *27*, 50–62. doi:10.1016/j.learninstruc.2013.03.001

Lowe, R. K. (1999). Extracting information from an animation during complex visual learning. *European Journal of Psychology of Education*, *14*, 225–244. doi:10.1007/BF03172967

Magnani, L. (2004). Model-based and manipulative abduction in science. *Foundations of Science*, *9*, 219–247. doi:10.1023/B:FODA.0000042841.18507.22

Mandinach, E. B., & Gummer, E. S. (2012). *Navigating the landscape of data literacy: It IS complex*. Washington, DC: WestEd/Education Northwest.

Manduca, C. A., & Kastens, K. A. (2012). Geoscience and geoscientists: Uniquely equipped to study the Earth. In K. A. Kastens, and C. Manduca (Eds.), *Earth & mind II: Synthesis of research on thinking and learning in the geosciences* (p. 1–12). Boulder, CO: Geological Society of America.

Manduca, C., & Mogk, D. W. (2002). *Using data in undergraduate science classrooms*. Retrieved from http://d32ogoqmya1dw8.cloudfront.net/files/usingdata/UsingData.pdf

Mayer, R. E., Mautone, P., & Prothero, W. (2002). Pictorial aids for learning by doing in a multimedia geology simulation game. *Journal of Educational Psychology*, *94*(1), 171–185. doi:10.1037/0022-0663.94.1.171

McNeill, K. L., & Krajcik, J. (2012). *Supporting grade 5–8 students in constructing explanations in science: The claim, evidence and reasoning framework for talk and writing*. New York, NY: Pearson Allyn & Bacon.

Merwade, V., & Ruddell, B. L. (2012). Moving university hydrology education forward with community-based geoinformatics, data and modeling resources. *Hydrology and Earth System Sciences*, *16*, 2393–2404. doi:10.5194/hess-16-2393-2012

Morton, M. C. (2010). Eyetrackers train students to see like geologists. *Earth Magazine*, October 2010, 28–33.

Nathan, M. J., Stephens, A. C., Masarik, D. K., Alibali, M. W., & Koedinger, K. R. (2002). Representational fluency in middle school: A classroom based study. In D. Mewborn, P. Szatajn, D. White, H. Wiegel, R. Bryant, & K. Nooney (Eds.), *Proceedings of the twenty-fourth annual meeting of the North American chapter of the international group for the psychology of mathematics education* (vol. 1, pp. 463–472). Columbus, OH: ERIC, Clearinghouse for Science, Mathematics, and Environmental Education.

National Research Council. (2007). Learning progressions. In R. A. Duschl, H. A. Schweingruber, and A. W. Shouse (Eds.), *Taking science to schools. learning and teaching science in grades K–8* (pp. 213–250). Washington, DC: The National Academies Press.

National Research Council. (2010). *Steps toward large-scale data integration in the sciences: Summary of a workshop. The current state of data integration in science*. Washington, DC: National Academies Press. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK45678/

National Research Council. (2012). *A framework for K–12 science education. Committee on a conceptual framework for new K–12 science education standards. Board on science education, DBASSE*. Washington, DC: National Academies Press.

National Research Council. (2014). Developing assessments for the next generation science standards. In Committee on Developing Assessments of Science Proficiency in K–12, Board on Testing and Assessment and Board on Science Education, J. W. Pellegrino, M. R. Wilson, J. A. Koenig, & A. S. Beatty (Eds.), *Division of behavioral and social sciences and education*. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press.

Oh, P. S. (2010). How can teachers help students formulate scientific hypotheses? Some strategies found in abductive inquiry activities of Earth science. *International Journal of Science Education*, *32*, 541–560. doi:10.1080/09500690903104457

Ormand, C. J., Manudca, C., Shipley, T. F., Tikoff, B., Harwood, C. L., Atit, K., & Boone, A. P. (2014). Evaluating geoscience students' spatial thinking skills in a multi-institutional classroom study. *Journal of Geoscience Education*, *62*(1), 146–154. doi:10.5408/13-027.1

Ormand, C. J., Shipley, T. F., Kiven, C., Davis, J. S., Klopfer, D., & Vrolijk, P. (2014, October). *The geologic block cross-sectioning test: Insights into the novice-expert spectrum in visual penetrative ability*. Paper presented at the Annual Meeting of the Geological Society of America, Vancouver, BC, Canada.

Pellegrino, J., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the next generation science standards*. Washington, DC: National Academies Press.

Rasch, G. (1980). *Probabilistic models for some intelligence and achievement tests* (expanded ed.). Chicago, IL: University of Chicago Press. (Original work published 1960)

Salinas, I. (2009, June). *Learning progressions in science education: Two approaches for development*. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Shipley, T. F., & Tikoff, B. (2016). Linking cognitive science and disciplinary geoscience practice: The importance of the conceptual model. In R. W. Krantz, C. J. Ormand, & B. Freeman (Eds.), *3-D structural interpretation: Earth, mind, and machine* (p. 219–237). Tulsa, OK: American Association of Petroleum Geologists memoir 111 (Hedberg Series number 6, August).

Slater, S. J., Slater, T. F., & Lyons, D. J. (2010). *Engaging in astronomical inquiry*. New York, NY: Freeman.

Slater, S. J., Slater, T. F., & Shaner, A. (2008). Impact of backwards faded scaffolding in an astronomy course for preservice elementary teachers base on inquiry. *Journal of Geoscience Education*, *56*(5), 408–416. doi:10.5408/jge_nov2008_slater_408

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*, 76–105. doi:10.1016/0022-1031(91)90011-T

Swenson, S., & Kastens, K. A. (2011). Student interpretation of a global elevation map: What it is, how it was made, and what it is useful for. In A. Feig & A. Stokes (Eds.), *Qualitative inquiry in geoscience education research* (pp. 189–211): Boulder, CO: Geological Society of America.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological Review*, *95*(1), 15–48.

Treisman, A., & Souther, J. (1985). Search asymmetry: a diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, *114*(3), 285–310.

Uttal, D., Meadow, N., Tipton, E., Hand, L., Alden, A., Warren, C., … Newcombe, N. S. (2013). The malleability of spatial

skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. doi:10.1037/a0028446

Ware, C. (2004). *Information visualization: Perception for design*. Amsterdam, The Netherlands: Elsevier.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730. doi:10.1002/tea.20318

Wolkovich, E. M., Regetz, J., & O'Connor, M. I. (2012). Advances in global change research require open science by individual researchers. *Global Change Biology*, *18*, 2102–2110. doi:10.1111/j.1365-2486.2012.02693.x

Wu, H.-K., Krajcik, J. S., & Soloway, E. (2001). Promoting conceptual understanding of chemical representations: Students' use of a visualization tool in the classroom. *Journal of Research in Science Teaching*, *38*, 821–842. doi:10.1002/tea.1033