JOURNAL OF GEOSCIENCE EDUCATION **61**, 68–80 (2013)

The Effects of Scaling Cues and Interactivity on a Viewer's Ability to Estimate the Size of Features Shown on Outcrop Imagery

Cari L. Johnson,[1,a] Ian L. Semple,[2] and Sarah H. Creem-Regehr[3]

ABSTRACT

The scale of features shown on outcrop photographs can be critical to geoscience interpretations, yet little is known about how well individuals estimate scale in images. This study utilizes a visualization test in which participants were asked to estimate the absolute size of several boxes shown in outcrop images using high resolution, stitched photopanoramas (Gigapans). Participants viewed two different outcrops that highlight different kinds of photographic distortion, first using static images and then with "interactive" Gigapans that permitted zooming and panning. A test group was given basic scaling cues in the form of distance to and height of the outcrops, whereas a control group completed the test without any scaling cues. Other population comparisons were investigated (e.g., gender, age, experience level, and major) but no other statistically significant population difference was observed. Therefore, scaling cues seem to invoke a primary effect at least in the first part of the exercise. Results show that scaling cues increase accuracy overall, but with wider spread and a tendency to cause overestimation of size. The control group, which was not given any scaling information, was less accurate overall and tended to underestimate the size of features. Both groups gave more accurate scale estimates with smaller standard deviations for the extension-distorted photopanorama than the compression-distorted image. Participants also generally showed improved accuracy in the second part of the test, which probably reflects the impact of interactivity, although a training effect cannot be discounted. These results suggest that nonembedded scaling cues (as opposed to physical objects denoting scale in photographs) can be useful for some individuals to estimate the size of features shown in outcrop images. Results also underscore the importance of interactivity and multiple exposures in classroom applications. © *2013 National Association of Geoscience Teachers*. [DOI: 10.5408/12-329.1]

Key words: physical scale, visualization, outcrop imagery, Gigapan

INTRODUCTION

Spatial cognition and visualization are complex but essential components of earth science education and research (e.g., Orion et al., 1997; Libarkin and Brick, 2002; Black, 2005; Kastens and Ishikawa, 2006; Piburn et al., 2011; Ormand, 2012). Earth scientists often use images to communicate scientific concepts, providing cues to establish the absolute physical scale of features shown ("hammer for scale," etc.). How effective are these kinds of scaling cues? Do observers translate and apply that information correctly? This study investigates how viewers estimate scale using outcrop imagery, specifically high resolution photopanoramas (Gigapans) that cover large swaths (>10 m) of rock exposure with various types of photographic distortion.

In geoscience, spatial representations commonly convey complex three-dimensional (3D) information using a two-dimensional (2D) plane (e.g., photos, maps, fence diagrams, seismic sections). A growing body of research addresses spatial cognition as it relates to geoscience education, including visualization skills like 2D–3D transference, shifting frame of reference, and spatial transformations (Mathewson, 1999; Kastens and Ishikawa, 2006; Kastens et al., 2009; Titus and Horsman, 2009; Kastens, 2010). In addition to these cognitive skills, recognizing the relative size of geologic features can be critical to accurate interpretations. For example, the physical hierarchy and relationships of architectural elements in fluvial systems largely dictate their interpretation. Similarly, key aspects of the size and type of paleo-river system are interpreted based on the scale of features deposited within them (e.g., bar forms; Miall, 1996).

Despite the importance of scale to many different areas of geoscience (and STEM fields in general; AAAS, 1989), relatively few studies focus on the issue of physical scale perceptions and estimates. As summarized by Jones and Taylor (2009), some of the recent and relevant literature includes studies of different conceptual divisions of space (Hegarty et al., 2006), the significance of human-scale interactions versus larger- and smaller-scale perceptions (Tretter et al., 2006a, 2006b), the impact of proportional reasoning (Jones et al., 2007), and the use of representational "rulers" such as body size (Jones et al., 2009). Observational skills related to physical scale estimates can be improved through repeated experience and practice (e.g., Charness et al., 1996). Nevertheless, there is much to learn about an individual's use of scaling cues in different contexts. Lock and Molyneaux (2006) summarized the issue as follows: "Scale is a slippery concept, one that is sometimes easy to define but often difficult to grasp . . . there is much equivocation about scale, as it is at the same time a concept, a lived experience, and an analytical framework" (p. 1; cf., Jones and Taylor, 2009).

Geoscience educators often use outcrop photos in lectures to illustrate geologic features. Photograph scale in

Received 29 May 2012; revised 9 October 2012; accepted 24 October 2012; published online 21 February 2013.

[1]Department of Geology and Geophysics, 115 S 1460 E, FASB 383, University of Utah, Salt Lake City, Utah 84112, USA
[2]Western Geco, 10001 Richmond Avenue, Houston, Texas 77042, USA
[3]Department of Psychology, University of Utah, 380 S. 1530 E., BEHS 502, Salt Lake City, Utah 84112, USA
[a]Author to whom correspondence should be addressed. Electronic mail: cari.johnson@utah.edu.

 © Nat. Assoc. Geosci. Teachers

these cases is typically depicted by showing a familiar object as a representative "ruler," by overlaying a scale bar on the image, or by making a more general informative statement (e.g., "the cliffs are 50 meters tall"). It is perhaps assumed that this information relays an accurate sense of scale, but this hypothesis has not been thoroughly tested. Eye tracking studies suggest that scale objects may act as distracters, potentially impacting novice-level viewers in particular, so there are important pedagogical implications (Coyan et al., 2010; Morton, 2010). For example, students may focus more on the scale object than the geologic subject the instructor is trying to convey. Furthermore, "familiar" objects are not necessarily interpreted the same way by different individuals: vegetation (e.g., forest trees versus desert bushes) might be inferred differently depending on an individual's background. Previous experience thus influences spatial understanding in an effect known as representational correspondence (Biederman, 1972; Chabris and Kosslyn, 2005; Ishikawa and Kastens, 2005; Jones and Taylor, 2009).

An additional complication to the issue of image scale is that distortion in photographs can create misrepresentations. There are two main types of perspective distortion common in geologic images: (1) extension distortion, which results in a forelengthening effect, and (2) compression distortion, which results in a foreshortening effect (Pratt, 1978). Extension distortion occurs when a wide angle photograph is taken of a subject close to the camera. This distortion makes objects closer to the camera appear larger in size relative to objects that are farther away. Conversely, compression distortion occurs when a photograph is taken of objects far away from the camera, as with panoramas and telephoto images. In this case, objects distant from the camera appear large relative to those that are closer. This effect greatly reduces the viewer's ability to judge distance and size (Hegarty et al., 2006).

Finally, interactivity has emerged as another key component to understanding geoscience imagery. Interactivity, defined in this context as some kind of user-driven manipulation of the image (e.g., zooming in/out, panning across), has also been cited as important for other kinds of visualization tests (Reynolds et al., 2005). However, interactivity is also poorly understood with respect to its effectiveness in improving spatial skills (Keehner et al., 2008). This is potentially an important pedagogical tool, given that interactive visualization labs are now found on many university campuses. These labs can range from simple stereographic projection systems (i.e., Geowall systems; Kelly and Riggs, 2006) to fully immersive virtual environments (Lin et al., 2000). Other interactive visualization methods used to improve spatial thinking include shaded topographic displays, satellite maps, and block diagrams that can be rotated and turned partially transparent to permit penetrative visualization of the block interior (Piburn et al., 2002; Arrowsmith et al., 2005; Piburn et al., 2005; Reynolds et al., 2005).

This study investigates the effects of scaling cues and interactivity (and/or repeat exposures) using high resolution, digital 2D photopanoramas. The main hypothesis tested is that providing scaling cues will generally result in more accurate scale estimates than not providing such cues. Outcrop images displaying different kinds of photographic distortion (extension versus compression distortion) were used to see if the effect of scaling cues varied under these scenarios. A secondary effect we investigated is whether estimates are improved by allowing for interactivity via zooming and panning the image. We also acquired various demographic data to investigate whether other population effects may be evident (gender, amount and type of previous geoscience experience, etc.).

METHODS

Photopanoramas

High resolution photopanoramas were taken from Nelson Canyon and Stone House Canyon, both offshoots of the larger Range Creek Canyon in central Utah (Fig. 1). The Flagstaff and Colton Formations (Paleocene-Eocene), featured in the panoramas, are widespread units across much of central Utah and also have many outcrop characteristics similar to other nonmarine units in the region. The panoramas were produced using a Gigapan, which is a tripod-mounted robot that takes individual photos that are then stitched together to produce a single high resolution image (Gigapix Systems, 2008). Both panoramas were taken with the camera fully zoomed (12x). The Nelson Canyon panorama was taken close to the outcrop (30 m away), producing extension distortion, while the Stone House Canyon panorama was taken from far away (~3,600 m away from the center of the image),

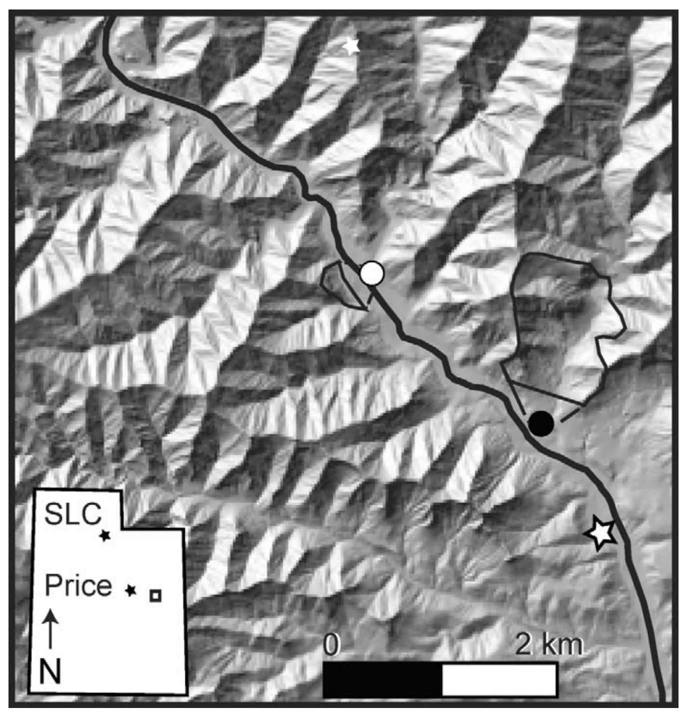

FIGURE 1: Hill-shade map of Range Creek Canyon, central Utah showing the locations of the Nelson Canyon (ExtDis; white circle) and Stone House Canyon (CompDis; black circle) panoramas. Black outlines show approximate field of view of the stitched photopanoramas (Gigapans). The DEM for this hillshade was generated using 10 m NED maps from the Utah GIS portal (Automated Geographic Reference Center, 2011). Stars on the inset map denote the cities of Salt Lake City (SLC) and Price; the small square shows the location of Range Creek Canyon.

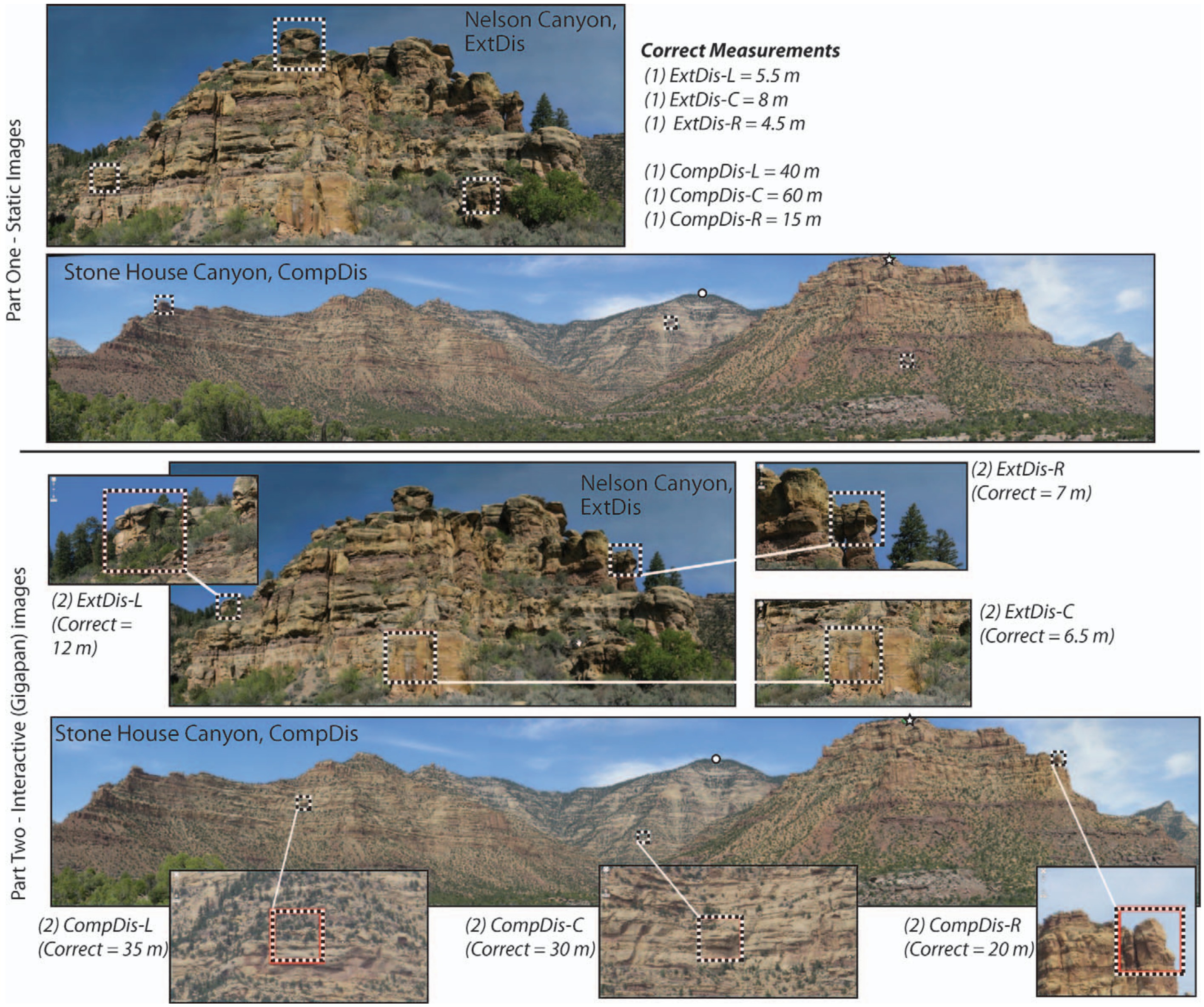

FIGURE 2: Summary of the test exercise. Part 1 (top): All participants were asked to estimate the size of three boxes shown on these static images (left, right, and center), beginning with the ExtDis panorama. Part 2 (below): All participants were asked to estimate the size boxes using interactive Gigapan images of the same outcrops. Participants could zoom in and out and pan across the images. Different boxes were used than in Part 1. The same scaling cues were again provided for the SC group. Correct answers are provided for each box (see footnote 4 on page 73 in the text for naming convention). The scaling cues provided to the SC group were as follows: ExtDis panorama: the cliffs are ~90 m (~300 ft) tall at the highest point near the center; distance from the camera to the base of the cliff is ~30 m (~100 ft). CompDis panorama: the height of the distant cliffs in the center (circle) above the camera is ~950 m (3100 ft) and they are ~3600 m (~11,800 ft or 2.2 miles) away from the camera. The height of the cliffs on the right (star) is ~400 m (1,300 ft) above the camera and their approximate distance from the camera is ~1,000 m (~3,300 ft).

producing compression distortion. For simplicity, these photopanoramas are referred to by the abbreviations "ExtDis" (extension distortion, Nelson Canyon) and "CompDis" (compression distortion, Stone House Canyon) throughout the rest of this manuscript.

Exercise Description

Participants were first given a short description of the purpose and background for the exercise and the location of the images (Fig. 1). Participants were then asked basic demographic questions, such as age, gender, degree in progress, major, etc. Completing a demographics survey before the test introduces the potential for stereotype bias (Steele and Aronson, 1995; Shih et al., 1999). However *t*-tests discussed below do not indicate any such bias in our dataset, given that population comparisons based on gender, major, experience level, and so on do not indicate statistically distinct results.

The workflow for the test exercise is summarized in Fig. 2 (additional details including the actual worksheets are in Semple, 2011). Data were collected over 6 months between March and August 2010 at the University of Utah. The test was conducted using a computer screen ranging from 16 to 18 inches diagonally. All participants ($n_{\text{total}} = 63$) first viewed static images of each outcrop, followed by interactive images via the Gigapan website, which allows for zooming in and out and panning across the photopanorama. The extension distortion image (ExtDis) was shown first, followed by the compression distortion image (CompDis) in both parts of the test. Given the size of the test population, we did not investigate possible randomization effects such as showing the CompDis photo first. While this decision may indeed have impacted the results, we note later that the estimates for the ExtDis panorama tended to be more accurate overall, so simple improvement based exclusively on training is not likely to be evident in this case.

In Part 1 of the exercise, participants were asked to estimate the height of the three boxes for each panorama using either feet or meters (their choice). Participants were given a time limit of 90 seconds to complete the estimates for each panorama, although most used less than 60 seconds. Each person was then asked to describe the process they used to estimate the size of the boxes. These responses are provided in Semple (2011); most refer to estimating the size of familiar features like trees and bushes in the panoramas. Finally, participants were asked to provide an estimate of how close they thought they were to the correct answer (i.e., a "confidence" factor). Many individuals provided their confidence answers in feet or meters, and some provided a percentage of their estimate. To normalize these values, all error estimates were converted to percentages in order to scale their accuracy prediction to the magnitude of their estimate. The different box sizes used in the test were checked in the field where possible, and cross-checked using topographic maps and Google Earth imagery. Certainly there is error associated with these measurements and approximations of "correct" (see Fig. 2), but we estimate it to be significantly less than error associated with the exercise, where participants had no means for direct measurement of scale.

For Part 2 of the exercise, the participants used an interactive format via the Gigapan website, with the ability to zoom and pan across the image (see www.gigapan.org, search for "UDOM"). After learning the controls, viewers were asked to estimate the height of three boxes on each panorama with the same time limits as before (note that the location of the three boxes varied from Part 1; Fig. 2). As before, they were asked to describe the process they used to estimate the box sizes as well as how close they thought they were to the true value.

A control group provided estimates without any scaling information (we refer to this as the no scaling cues [NS] group), whereas a test group was given some general scaling cues (the scaling cues [SC] group; Table I). The indirect scaling cues provided to the SC group included distance to and heights of the cliffs, provided in both feet and meters, in multiple places: this information is detailed in the Fig. 2 caption. We used these indirect cues rather than embedded scale bars to test whether such information is useful in place of possible distracters.

TABLE I: Demographic information of participants minus outliers ($n = 50$). All values are in percents. Average age of participants is 27 years. "Other" majors include political science, theater, business, liberal arts, environmental science, and metallurgy.

Gender	
Female	26
Male	74
Majors	
Geology or geophysics	68
Geography	12
Other	20
Degree	
PhD candidates or faculty	16
MS candidates	38
Seeking a BA or BS degree	46
No scaling cues (NS)	52
Females in NS group	23
Geoscience majors in NS group	81
Undergraduates in NS group	42
Scaling cues (SC)	48
Females in SC group	29
Geoscience majors in SC group	54
Undergraduates in SC group	50

RESULTS

As introduced previously, our primary goal in this study was to determine how inclusion of a scaling cue would influence absolute judgments of size, including accuracy as well as self-reported confidence or error. Furthermore, we investigated whether these effects are modulated by the distance of the images portrayed, interactivity with the images, and the demographics (particularly gender and expertise) of the population tested. Below we describe relevant descriptive and statistical analyses in the context of these questions.

All estimates (Table IIA; Semple, 2011) are presented in meters, converted from feet where necessary. Any participant who reported at least one estimate that exceeded two times the standard deviation (of the whole group averages for each box) was identified as an outlier and not included in subsequent analyses; 13 of the original 63 participants were removed in this manner so these are not included in the following statistical analysis ($n = 50$ after outliers; 21% of the original test population removed). Although this screening procedure eliminated a large part of the population from further analysis, the 2x standard deviation filter shows reasonable consistency between averages and medians for all groups. Using a 1x standard deviation filter would have decimated the test population, and using 3x standard deviation gave unnecessarily large ranges. The outlier participants removed in this manner demonstrated no obvious demographic similarities with one another that would indicate a prediction of such estimates. However, 9 of the 13 outliers were from the SC group (4 from the NS group), indicating wider spread given scaling cues. Most of

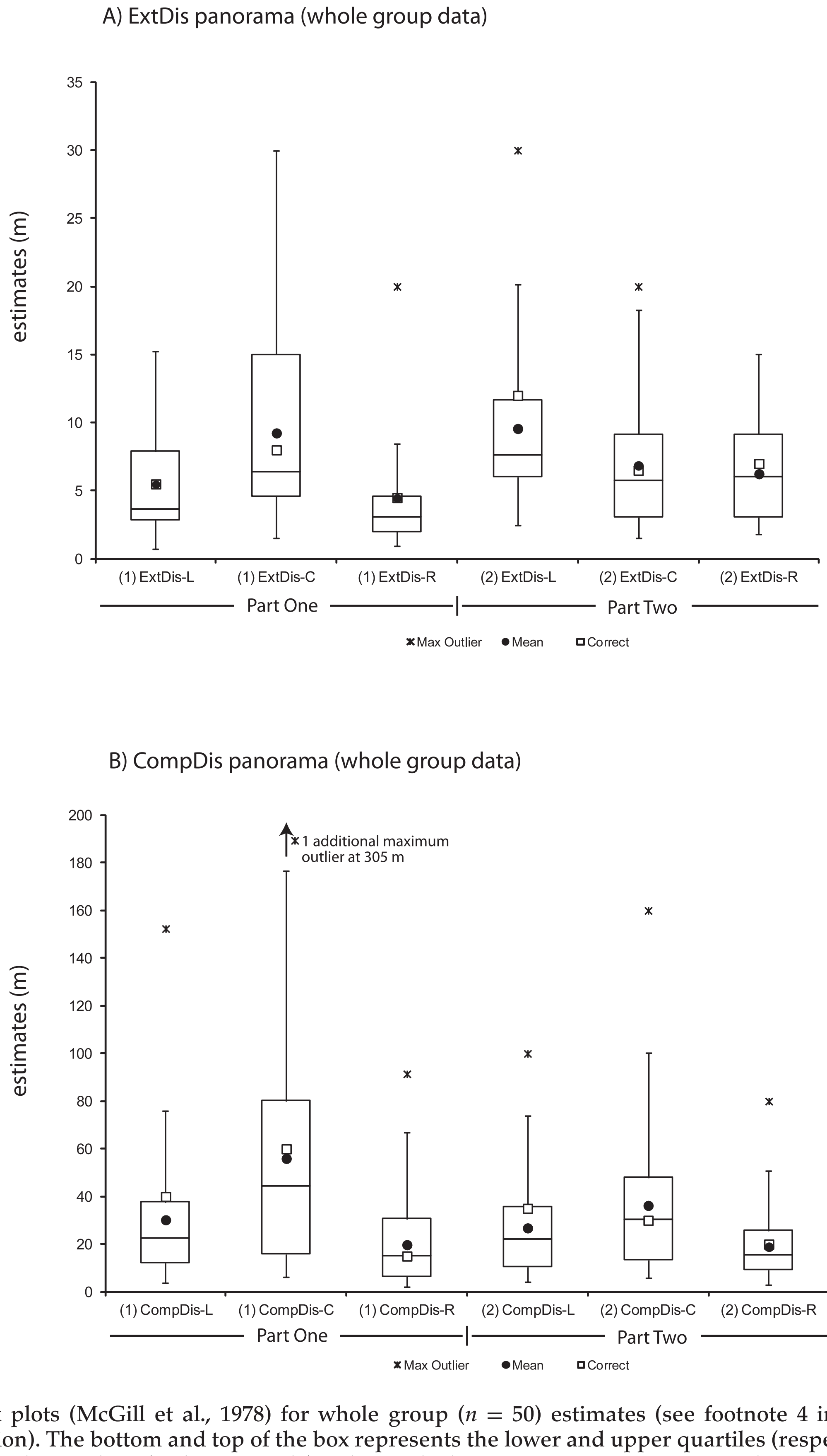

FIGURE 3: Box plots (McGill et al., 1978) for whole group (n = 50) estimates (see footnote 4 in the text for box naming convention). The bottom and top of the box represents the lower and upper quartiles (respectively), the band in the middle is the median (50th percentile) value. Whiskers represent maximum and minimum values, with maximum outliers also noted (there were no minimum outliers). Mean values and correct values are also plotted for each box (black circle and white square, respectively).

the 2x standard deviation outliers were based on overestimates relative to the whole group means, a trend which is observed for the SC group in general, as discussed below. Unless otherwise stated, the following discussion refers to the filtered dataset ($n = 50$).

Whole Population Demographics and Results

The test population ($n = 50$) had an average age of 27 years, and was 74% male (Table I). Almost half (46%) were undergraduate students, 38% were master's candidates, and 16% were doctoral candidates or faculty. The population included mostly geoscience majors (68%), with 12% from geography and 20% from other disciplines (Table I), More than half of the participants (58%) answered in feet, 36% answered in meters, whereas 6% used a combination of both. There does not appear to be a clear population-based predictor of what unit system was used. All estimates were converted to meters before being evaluated.

Overall, the whole group means were within 10% of correct for 6 of the 12 boxes, and within 20% correct for all but one box (Fig. 3; Table IIA and B). Outliers on the box plots were all overestimates, similar to the trend of the 2x standard deviation outliers that were filtered out from the original test population. The average and median estimates from the whole group for both parts of the test tended to be more accurate and show less spread for the ExtDis image than the CompDis image. This indicates, as expected, that scale is more difficult to estimate using more distant images. Whole group answers improved in accuracy and show decreasing spread for most boxes between parts one and two of the test. This may reflect the effect of interactivity but also could represent a training effect due to repeat exposure, as will be discussed later.

To investigate population differences, independent *t*-tests (two-tailed and unequal variance) compared the means of the NS versus SC groups, male versus female gender groups, undergraduate versus graduate student plus faculty groups, and geoscience (geology or geophysics) versus nongeoscience major groups (Table III). Using a cutoff of *p* values less than 0.05 to indicate statistically distinct populations, only the NS versus SC group indicated statistically significant differences and only for Part 1 of the test. Additional population differences may become evident with more robust sample sets, but based on this preliminary dataset, the primary population distinction is based on first exposure (Part 1 of the test) of the NS versus SC groups. The following discussion therefore focuses on this comparison.

Scaling Cues Group (SC) vs. No Scaling Cues (NS) Group: Estimates

Beginning with similar between-group trends, both the SC ($n = 24$) and NS ($n = 26$) groups generally show convergence, with increasing accuracy and decreasing standard deviations, from Part 1 to Part 2 of the test (i.e., comparing within the CompDis and ExtDis panoramas; Fig. 4, Table IIA). However, the NS group did show an increase in spread for four of the boxes in Part 2 compared to Part 1 (Table IIA). Both groups also had greater standard deviations, along with higher mean and median estimates, for the CompDis panorama than the ExtDis panorama.

Between-group distinctions include consistently higher estimates from the SC group compared to the NS group on all box estimates, for both mean and median comparisons (Fig. 2; Table IIA); one minor exception is (2)CompDis-L box[4], where the averages were basically equal (1% difference). The SC group exhibited a greater standard deviation than the NS group for all boxes in Part 1 of the test, but only for the first two boxes in the second part of the test. As noted previously, out of the original test population, participants from the SC group were more than twice as likely to be identified as a >2x standard deviation outlier, which indicates greater spread in the SC group estimates. Whereas both groups generally showed convergence and less spread in Part 2 of the test, this effect is more significant in the SC group, particularly for the CompDis panoramas.

The SC group and the NS group were subequal (to each other) in terms of gender split and percent undergraduates (Table I). By chance, the NS group was more heavily represented by geoscience majors than the SC group (81% geoscience in the NS group versus 54% in the SC group). The population-comparison *t*-tests discussed previously did not show a significant difference related to experience type (geoscience versus other majors, all p values > 0.05; Table III). To investigate this further, we completed an analysis of variance (ANOVA) based on $2 \times 2 \times 2$ (Scale group = NS vs. SC; Discipline = geoscience vs. other; Test part = one vs. two) for each box on each photopanorama. These results (Table IV) underscore the initial *t*-test results, showing that scaling information changed estimates for most of the photopanoramas and boxes particularly in Part 1 of the test (as revealed by the part x scale group interaction). The analysis also confirms a significant change in mean estimates across Part 1 and Part 2, supporting an effect of interactivity. However, there were no evident effects based on discipline. Therefore, despite the heavier influence of geoscience majors in the NS group, we interpret the influence of scaling cues to be the primary effect rather than discipline. Future studies might further investigate additional population effects.

Scaling Cues Group vs. No Scaling Cues Group: Accuracy

Table IIA summarizes differences from the correct answer as decimal percents (estimate/correct answer; i.e., 1 = "perfect" answer, 0.50 ratio means that the group average or median underestimated by 50% of correct, whereas a 1.50 ratio indicates overestimation by 50%). Accuracy is shown graphically in Fig. 5, normalized to zero = correct. The NS group consistently underestimated box sizes on all median values and all but one average value. The SC group overestimated 10 of the 12 boxes based on group average values versus 7 of the 12 boxes based on group median values. The SC group was more accurate than the NS group on all but one of their group estimates based on median values, (1)CompDis-R. A second box, (2)CompDis-C, showed no statistical difference in accuracy between the SC and NS groups' median values. However, based on average values the results are more mixed (Fig. 5b). An overall "score" based on all difference-from-correct ratios for each panorama also indicates that the SC group estimated more accurately overall (Table IIB). The SC group showed

[4] Notation for specific boxes (see Fig. 2) used in the text is as follows: (part 1 or 2)Panorama (ExtDis or CompDis)-box (left [L], center [C], or right [R]). For example, "(2)CompDis-L" refers to Part 2, CompDis panorama, left box.

TABLE IIA: Data summary.

Whole group $n = 50$	Part 1 (1) ExtDis-L	(1) ExtDis-C	(1) ExtDis-R	(1) CompDis-L	(1) CompDis-C	(1) CompDis-R
"Correct" answers (m)	5.5	8.0	4.5	40.0	60.0	15.0
Average (m)	5.48	9.24	4.46	30.23	55.97	19.78
Median (m)	3.66	6.40	3.05	22.19	44.20	15.00
St dev (m)	4.12	6.85	4.37	27.28	52.93	17.26
St dev (% of average)	75	74	98	90	95	87
Decimal % of Correct (individual estimate/correct) (1 = perfect)						
Average	1.00	1.16	0.99	0.76	0.93	1.32
Median	0.67	0.80	0.68	0.55	0.74	1.00
St dev	0.75	0.86	0.97	0.68	0.88	1.15
NS versus SC Group Comparisons						
NS group average (m)	4.29	6.79	3.13	20.27	34.37	13.61
NS group median (m)	3.02	5.00	2.13	14.12	27.19	8.05
NS group st dev (m)	3.82	6.32	3.71	15.08	28.02	12.11
NS group st dev (% of average)	89	93	119	74	82	89
SC group average (m)	6.77	11.91	5.90	41.01	79.37	26.46
SC group median (m)	6.55	10.67	4.29	32.74	80.00	30.24
SC group st dev (m)	4.12	6.50	4.63	33.25	63.37	19.65
SC group st dev (% of average)	61	55	78	81	80	74
Decimal % of Correct (estimate/correct) (1 = perfect)						
NS group average (m)	0.78	0.85	0.70	0.51	0.57	0.91
NS group median (m)	0.55	0.63	0.47	0.35	0.45	0.54
SC group average (m)	1.23	1.49	1.31	1.03	1.32	1.76
SC group median (m)	1.19	1.33	0.95	0.82	1.33	2.02

even more improvement from Part 1 to Part 2 of the test compared to the NS group.

Scaling Cues Group vs. No Scaling Cues Group: Confidence Factors

As mentioned previously, participants were asked to estimate how close they thought they were to the correct answer after each panorama in both parts of the exercise. Participants typically answered in meters or feet, some answered in percent (Semple, 2011). These self-reported "error" estimates (i.e., confidence factors) were normalized to the appropriate meter value and then converted to decimal percent relative to that individual's size estimates for each panorama (average of all three boxes) for both parts of the test (Table V). In this case, a smaller number represents higher confidence. In other words, if an individual reported a box-size estimate of 10 m with a 1 m error range, their converted confidence ratio would be 0.1. In some cases, error reports were extremely high (e.g., 10 m estimate with 10–20 m uncertainty), resulting in confidence ratio scores of 1 or greater. Of course, it is highly unlikely that these individuals actually thought that the size of the boxes could be 0 m or even less, but we include these results for comparisons of relative confidence between and within groups (Fig. 6).

Six of the participants included in the post-outlier analysis ($n = 50$) only gave qualitative responses to this part of the exercise (e.g., "I am not very confident"). Four of these were from the NS group, two from the SC group. These responses were not included in the following analysis of confidence ($n = 44$). Error estimates were reported for the

TABLE IIB: Panorama "scores" based on average or median values for all boxes; group average or median/correct (1 = perfect).

	(1) ExtDis	(2) ExtDis	ExtDis Part 1 to 2 Change	(1) CompDis	(2) CompDis	CompDis Part 1 to 2 Change	Overall Score (all estimates)
NS group average	0.77	0.81	−0.03	0.66	0.89	−0.23	0.78
NS group median	0.55	0.64	−0.09	0.45	0.60	−0.15	0.56
SC group average	1.34	1.03	0.32	1.37	1.07	0.30	1.20
SC group median	1.16	0.92	0.24	1.39	1.06	0.33	1.13

TABLE IIA: Extended.

Whole group $n = 50$	Part 2					
	(2) ExtDis-L	(2) ExtDis-C	(2) ExtDis-R	(2) CompDis-L	(2) CompDis-C	(2) CompDis-R
"Correct" answers (m)	12.0	6.5	7.0	35.0	30.0	20.0
Average (m)	9.57	6.83	6.25	26.81	36.26	19.05
Median (m)	7.62	5.74	6.00	22.10	30.24	15.24
St dev (m)	6.05	4.78	3.54	21.56	30.59	14.54
St dev (% of average)	63	70	57	80	84	76
Decimal % of Correct (individual estimate/correct) (1 = perfect)						
Average	0.80	1.05	0.89	0.77	1.21	0.95
Median	0.64	0.88	0.86	0.63	1.01	0.76
St dev	0.50	0.73	0.51	0.62	1.02	0.73
NS versus SC Group Comparisons						
NS group average (m)	9.39	5.66	5.40	26.93	31.22	17.29
NS group median (m)	7.62	4.29	4.29	19.05	20.57	11.33
NS group st dev (m)	6.07	4.15	3.55	26.09	32.00	17.16
NS group st dev (% of average)	65	73	66	97	102	99
SC group average (m)	9.76	8.10	7.16	26.69	41.72	20.96
SC group median (m)	9.15	7.31	6.10	25.45	40.00	22.20
SC group st dev (m)	6.15	5.17	3.35	15.82	28.65	11.07
SC group st dev (% of average)	63	64	47	59	69	53
Decimal % of Correct (estimate/correct) (1 = perfect)						
NS group average (m)	0.78	0.87	0.77	0.77	1.04	0.86
NS group median (m)	0.64	0.66	0.61	0.54	0.69	0.57
SC group average (m)	0.81	1.25	1.02	0.76	1.39	1.05
SC group median (m)	0.76	1.12	0.87	0.73	1.33	1.11

whole panorama (ExtDis or CompDis) rather than individual boxes, so the same confidence estimate was applied to all boxes in each panorama and the results shown as box and group averages (Table V; Fig. 6).

An independent *t*-test indicates that the NS and SC groups are not statistically different in terms of their self-reported error estimates (only one comparison, (2)ExtDis (NS group), has a p value < 0.05; Table III). Nevertheless, a 2 $\times$ 2 $\times$ 2 repeated measures ANOVA (Photopanorama; Test part; Scale group) showed some interesting results (Table V). A significant main effect of photopanorama ($p = 0.001$) indicates that both the NS and SC groups thought that they were more accurate, relative to their own average estimates, in the CompDis panorama than in the ExtDis panorama (Fig. 6). A panorama x part interaction ($p = 0.048$), revealed increased confidence from parts one to two of the test, but

TABLE III: *T*-test population comparisons showing computed *p* values—bold numbers less than 0.05 indicate statistical differentiation.

	ExtDis-L	ExtDis-C	ExtDis-R	CompDis-L	CompDis-C	CompDis-R
Part 1						
(34) Geo vs. (16) Other	0.096	0.051	0.232	0.057	0.182	0.137
(13) Female vs. (37) Male	0.609	0.850	0.348	0.129	0.322	0.372
Undergrad (22) vs. Grad (28)	0.071	0.080	0.883	0.073	0.106	0.056
NS vs. SC	**0.033**	**0.007**	**0.025**	**0.009**	**0.003**	**0.009**
Part 2						
(34) Geo vs. (16) Other	0.184	0.175	**0.024**	0.393	0.178	0.184
(13) Female vs. (37) Male	0.559	0.505	0.696	0.153	0.186	0.432
Undergrad (22) vs. Grad (28)	0.527	0.718	0.460	0.928	0.755	0.455
NS vs SC	0.829	0.073	0.077	0.969	0.227	0.370

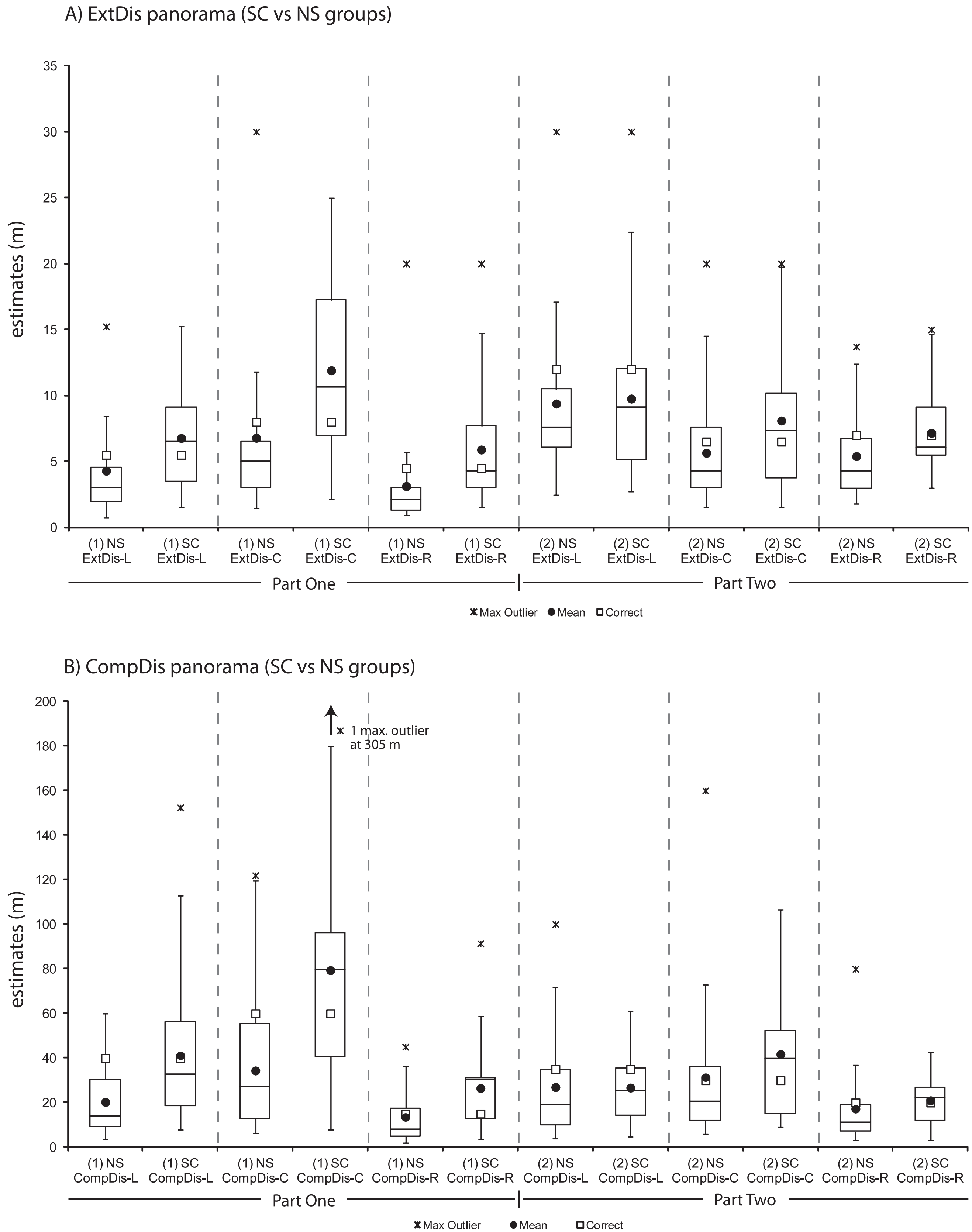

FIGURE 4: Box plots (McGill et al., 1978; see Fig. 1 caption) comparing SC and NS group median estimates for ExtDis (A) and CompDis (B) panoramas. Dashed gray lines separate results from each box used in the exercise.

TABLE IV: p values resulting from 2 (Scale Group) × 2 (Discipline) × 2 (Part) ANOVA; p values less than .05 in bold.

	ExtDis-L	ExtDis-C	ExtDis-R	CompDis-L	CompDis-C	CompDis-R
Part (1 vs. 2)	**0.001**	**0.016**	**0.001**	0.416	**0.004**	0.951
Scale group (NS vs. SC)	0.561	**0.036**	**0.036**	0.248	**0.035**	0.219
Discipline (Geo vs. Other)	0.087	0.152	0.156	0.083	0.299	0.141
Part × scale group	**0.044**	0.081	0.108	**0.014**	**0.018**	0.092
Part × discipline	0.201	0.718	0.182	0.487	0.844	0.548
Group × discipline	0.911	0.775	0.283	0.944	0.965	0.194

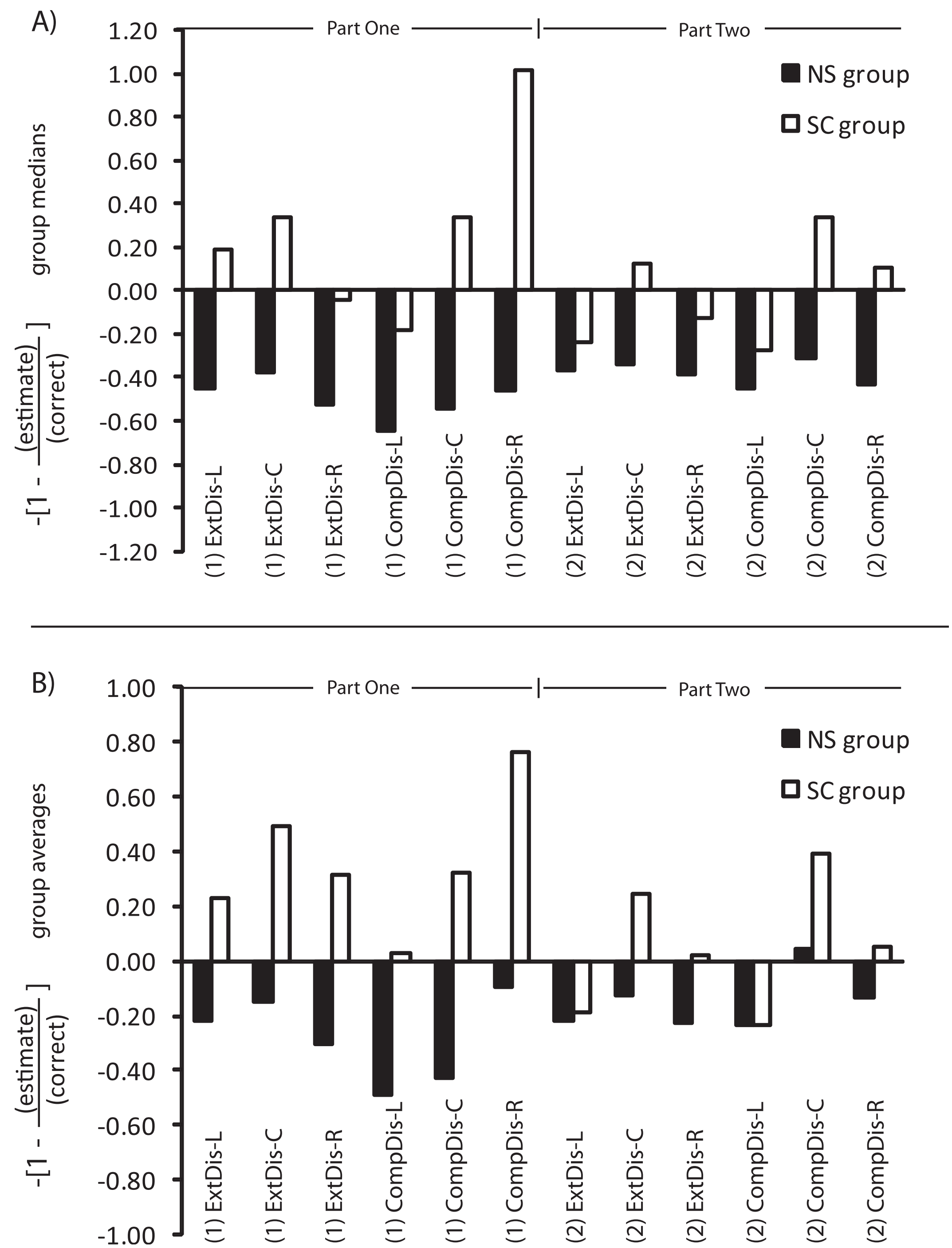

FIGURE 5: Comparison of accuracy for the SC and NC group, plotted as normalized ratio [-(1-(estimate/correct))] for group median values (A) and group means (B). Zero represents a perfect estimate, whereas negative numbers are underestimates.

TABLE V: Confidence ratios based on ([reported error] / [box size estimate]), averaged by group for each panorama.

	(1) ExtDis	(2) ExtDis	(1) CompDis	(2) CompDis
NS group mean	1.44	0.46	0.31	0.25
NS group st dev	2.61	0.17	0.37	0.15
SC group mean	1.58	1.46	0.38	0.59
SC group st dev	2.26	0.58	1.86	1.06
Difference SC minus NS	0.14	1.00	0.07	0.35
Change from Part 1 to 2:	ExtDis		CompDis	
NS Group	0.98		0.06	
SC Group	0.12		−0.22	
p Values, T-test NS vs SC Confidence ratios (bold = <0.05)	0.85	0.59	**0.02**	0.14

showed that this effect varied with the type of panorama and scale group. Notably, the SC group, CompDis panorama, reported 22% less confidence from Part 2 compared to Part 1 and the biggest confidence change occurred for the NS group in the ExtDis panorama (1.44 in Part 1 to 0.46 in Part 2; Table V).

DISCUSSION

Our main hypothesis in designing this study was that indirect scaling cues would improve visual estimates of scale on outcrop imagery. Indirect cues (i.e., approximate distance to and height of the outcrops) were used rather than embedding a specific scaled feature like a meter stick in order to test how well this information can be applied, as well as to avoid potential distracters (Jones et al., 2009; Coyan et al., 2010; Morton, 2010). We expected that the CompDis panorama would prove more difficult for scale estimates than the ExtDis panorama, because more distant features tend to be harder to estimate (Holway and Boring, 1941; Gilinsky, 1951; Stroebel and Zakia, 1993). Secondary predictions were that interactivity would improve scaling estimates as well as confidence based on user-reported error ranges. Other possible group effects were considered but, given the small anticipated size of the dataset, no specific predictions were made.

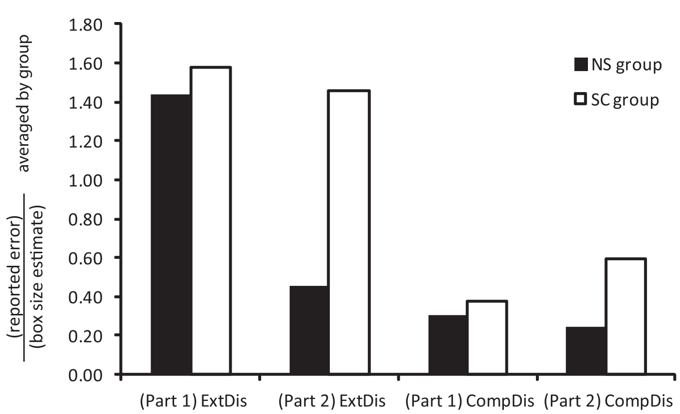

FIGURE 6: SC versus NS group confidence estimates in decimal form ([user reported error] / [user answer]), averaged by group for each panorama. Large numbers represent less "confidence," or higher estimates of error, relative to the actual estimates of box size.

Results largely confirm these basic predictions, although with several less obvious but potentially significant implications for the estimation of scale in geoscience imagery. As a whole (the test population minus 2x standard deviation outliers; $n = 50$), estimates were reasonably accurate, with $\sim^{2}/_{3}$ of the answers falling within 25% error range of correct (based on average values; Table IIA). Outliers, those filtered out by the 2x standard deviation filter as well as remaining outliers (Fig. 3), tended to be overestimates relative to the means, suggesting a general relationship between lower accuracy and overestimation. Global effects observed in the whole population as well as within the SC and NS group comparisons are that the CompDis panorama estimates tended to be less accurate than the ExtDis estimates, and that accuracy overall tended to improve (and standard deviations decrease) from Part 1 to Part 2 of the exercise. Confidence based on error estimates also tended to improve throughout the test in most cases.

The increase in accuracy from Part 1 to Part 2 could arguably reflect a training effect due to second exposure (Gibson and Bergman, 1954; Wagman et al., 2008). However, while the same photopanoramas were used in both parts of the test, different boxes (in different locations) were shown for Part 2, so the participants were not estimating size of features in the same locations. Nearly all participants cited "zooming in and out" in Part 2 as part of their method for estimating box size. Furthermore the SC versus NS subgroups actually showed different degrees of change (improvement in accuracy) from Part 1 to Part 2: the SC group showed more improvement compared to the NS group (Table IIB), suggesting that the combination of scaling cues plus interactivity influenced the results. Thus, while a training effect cannot be discounted, at a minimum it appears to be acting in combination with the interactive capability to improve accuracy (Charness et al., 1996; Jones and Taylor, 2009).

Population comparison tests indicate the only statistically significant group distinction was between the SC and NS groups, and that the two were only significantly different in Part 1 of the exercise. ANOVA analysis confirms basic t-test comparisons and also indicates that, although the NS group was much more heavily represented by geoscience majors, the primary effect reflects whether scaling cues were given rather than study discipline. It is particularly notable that the SC group consistently gave larger estimates of box sizes (usually overestimating from the correct answer) compared to the NS group (Fig. 4). This result suggests that, without any additional scale information, individuals tend to think that outcropping geologic features are smaller than they really are. The SC group estimates were generally closer to correct based on median values, implying that the scaling cues were in fact being used effectively, although with wider spread and more outliers than the NS group. Thus, it appears that embedded physical scale bars in images (i.e., rock hammer or person) are not necessarily critical to effectively conveying scale; i.e., indirect cues such as distance to and height of outcrop can also be effective for some individuals. However, most of the 2x standard deviation outliers were from the SC group, so the effectiveness of such cues appears to be highly variable individually. It is also interesting to note that the NS group tended to be more confident, though less accurate and with less spread, in their answers than the SC group. Furthermore, both groups reported higher confidence in their estimates for the CompDis panorama than the ExtDis panorama, even though the CompDis estimates were less accurate. This result suggests that lack of scaling information may give a false sense of confidence, and vice versa.

CONCLUSIONS AND IMPLICATIONS

This study involves a simple exercise conducted with a relatively small group. The implications are thus presented with some caution. Nevertheless, it is clear that sense-of-scale cognition is under studied but potentially quite important in geoscience education and research. Therefore follow-up and expanded studies are needed to fully investigate these preliminary interpretations. The key conclusions from this study are that scaling cues such as distance to and height of large outcrops do tend improve size estimates of outcrop features, and thus can be useful (and nondistracting) tools in place of, or perhaps in addition to, physical scale bars placed in the photograph. Whereas the scale cues appear to be effective for some, the SC group showed more spread and outliers, so this may be a strongly individual effect. Scaling cues also tended to result in higher estimates of error (less confidence) than no scale information. Therefore, in a classroom or research setting it might be useful to actually test and confirm accuracy while viewing images. If large and distorted outcrop images are shown without scaling cues, one might assume that most viewers will tend to underestimate their size based on these results. Finally, we argue that interactivity provided by zooming in/out and panning across high resolution photopanoramas (Gigapans) was useful for both groups in improving accuracy, particularly for the image with compression distortion.

Acknowledgments

The authors wish to thank Professors Lisa Stright and David Chapman at the University of Utah for their presubmission reviews and input, and two anonymous reviewers plus Associate Editor Clark for thorough reviews of two versions of this manuscript. This research was supported by a grant from the National Science Foundation to C. Johnson (NSF-DUE-0736053). This research was conducted with approval by the University of Utah Institutional Review Board.

REFERENCES

American Association for the Advancement of Science (AAAS). 1989. Science for all Americans. Washington, DC: AAAS.

Arrowsmith, C., Counihan, A., and McGreevy, D. 2005. Development of a multi-scaled virtual fieldtrip for the teaching and learning of geospatial science. *International Journal of Education and Development using Information and Communication Technology*, 1(3):42–56.

Automated Geographic Reference Center. 2011. Utah GIS Portal. Available at http://gis.utah.gov/ (accessed 5 January 2011).

Biederman, I. 1972. Perceiving real-world scenes. *Science*, 77(4043):77–80.

Black, A.A. 2005. Spatial ability and Earth science conceptual understanding. *Journal of Geoscience Education*, 53:402–414.

Chabris, C.F., and Kosslyn, S.M. 2005. Representational correspondence as a basic principle of diagram design. *In* Tergan, S.O. and Keller, T., eds., Knowledge and information visualization. Berlin, Germany: Springer-Verlag, p. 36–57.

Charness, N., Krampe, N., and Mahr, U. 1996. The role of practice and coaching in entrepreneurial skill domains: An international comparison of life-span chess skill acquisition. *In* Ericsson, K., ed., The road to excellence. Mahwah, NJ: Erlbaum, p. 51–88.

Coyan, J., Busch, M., and Reynolds, S. 2010. Using eye trackers to evaluate the effectiveness of signaling to promote the disembedding of geologic features in photographs. *In* Frick, A., Nardi, D., and Ratliff, K., eds., Spatial Cognition 2010: Doctoral Colloquium SFB/TR8 Report No. 025-07/2010. Bremen, Germany: Universitat Bremen and Universitat Freiburg, p. 15–19.

Gibson, E.J., and Bergman, R. 1954. The effect of training on absolute estimation of distance over ground. *Journal of Experimental Psychology*, 48:473–482.

Gigapix Systems. 2008. Available at http://www.gigapansystems.com/ (accessed 9 May 2009).

Gilinsky, A.S. 1951. Perceived size and distance in visual space. *Psychological Review*, 58:460–482.

Hegarty, M., Montello, D., Richardson, A., Ishikawa, T., and Lovelace, K. 2006. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, 34:151–176.

Holway, A.H., and Boring, E.G. 1941. Determinants of apparent visual size with distance variant. *The American Journal of Psychology*, 54:21–37.

Ishikawa, T., and Kastens, K. 2005. Why some students have trouble with maps and other spatial representations. *Journal of Geoscience Education*, 53:184–197.

Jones, M.G., and Taylor, A. 2009. Developing a sense of scale: Looking backward. *Journal of Research in Science Teaching*, 46:460–475.

Jones, M.G., Taylor, A., Minogue, J., Broadwell, B., Wiebe, E., and Carter, G. 2007. Understanding scale: Powers of ten. *Journal of Science Education and Technology Education*, 16:191–202.

Jones, M.G., Taylor, A., and Broadwell, B. 2009. Estimating linear size and scale: Body rulers. *International Journal of Science Education*, 31:1495–1509.

Kastens, K. 2010. Commentary: Object and spatial visualizations in geosciences. *Journal of Geoscience Educatio*n, 58:52–57.

Kastens, K.A., and Ishikawa, T. 2006. Spatial thinking in the geosciences and cognitive sciences: A cross-disciplinary look at the intersection of the two fields. *In* Manduca, C.A., and Mogk, D.W., eds., Earth and mind: How geologists think and learn about the Earth. *Geological Society of America Special Paper,* 413:53–76.

Kastens, K., Manduca, C., Cervato, C., Frodeman, R., Goodwin, C., Liben, L., Mogk, D., Spangler, T., Stillings, N., and Titus, S. 2009. How geoscientists think and learn. *Eos*, 90:265–272.

Keehner, M., Hegarty, M., Cohen, C., Khooshabeh, P., and Montello, D.R. 2008. Spatial reasoning with external visualizations: What matters is what you see, not whether you interact. *Cognitive Science*, 32:1099–1132.

Kelly, M., and Riggs, N. 2006. Use of a virtual environment in the GeoWall to increase student confidence and performance during field mapping: An example from an introductory-level field class. *Journal of Geoscience Education*, 54:158–164.

Libarkin, J.C., and Brick, C. 2002. Research methodologies in science education: Visualization and the geosciences. *Journal of Geoscience Education*, 50:449–455.

Lin, C.-R., Loftin, R.B., and Nelson, H.R. 2000. Interaction with geoscience data in an immersive environment. *Virtual Reality, IEEE Proceedings*. p. 55–62.

Lock, G., and Molyneaux, B. (eds.). 2006. Confronting scale in archeology. New York: Springer.

Mathewson, J.H. 1999. Visual-spatial thinking: An aspect of science overlooked by educators. *Science Education*, 83:33–54.

Miall, A.D. 1996. The geology of fluvial deposits. Berlin, Germany: Springer.

McGill, R., Tukey, J.W., and Larsen, W.A. 1978. Variations of box plots. *The American Statistician,* 32:12–16.

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

Orion, N., Ben-Chaim, D., and Kali, Y. 1997. Relationship between Earth-science education and spatial visualization. *Journal of Geoscience Education*, 45:129–132.

Ormand, C. 2012. Learning to think spatially. *In The Trenches (the news magazine of the National Association of Geoscience Teachers),* 2(2):10–11.

Piburn, M., Reynolds, S., Leedy, D., McAuliffe, C., Birk, J., and Johnson, J. 2002. "The hidden earth: Visualization of geologic features and their subsurface geometry." Paper presented at the National Association for Research in Science Teaching Annual Meeting, New Orleans, LA, April 2002.

Piburn, M.D., van der Hoeven, Kraft, K., and Pacheco, H. 2011. A new century for geoscience education research. Prepared for the National Academies Board on Science Education, Committee on the Status, Contributions, and Future Directions of Discipline-Based Education Research, p. 24. Available at http://www7.nationalacademies.org/bose/DBER_Piburn_October_Paper.pdf (accessed 25 September 2012).

Piburn, M.D., Reynolds, S.J., McAuliffe, C., Leedy, D.E., Birk, J.P., and Johnson, J.K. 2005. The role of visualization in learning from computer based images. *Journal of Science Education*, 27:513–527.

Pratt, W.K. 1978. Digital image processing. New York: John Wiley & Sons.

Reynolds, S., Johnson, J., Piburn, M., Leedy, D., Coyan, J., and Busch, M. 2005. Visualization in undergraduate geology courses. *In* Gilbert, J., ed., Visualization in science education. Dordrecht, The Netherlands: Springer Press, p. 253–266.

Semple, I. 2011. Sedimentary geology and geoscience education: A combined study of sequence stratigraphy and spatial visualization using outcrop examples from the Cretaceous Tertiary of Utah. [M.S. Thesis]. Salt Lake City: University of Utah, p. 190.

Shih, M., Pittinsky, T.L., and Ambady, N. 1999. Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, 10(1):80–83.

Steele, C.M., and Aronson, J. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5):797–811.

Stroebel, L., and Zakia, R. 1993. The focal encyclopedia of photography. Woburn, MA: Butterworth-Heinemann Publishing.

Titus, S., and Horsman, E. 2009. Characterizing and improving spatial visualization skills. *Journal of Geoscience Education*, 57:242–254.

Tretter, T., Jones, G., Andre, T., Negishi, A., and Minogue, J. 2006a. Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena. *Journal of Research in Science Teaching*, 43:282–319.

Tretter, T.R., Jones, M.G., and Minogue, J. 2006b. Accuracy of scale conceptions in science: Mental maneuverings across many orders of spatial magnitude. *Journal of Research in Science Teaching*, 43:1061–1085.

Wagman, J.B., McBride, D.M., and Trefzger, A.J. 2008. Perceptual experience and posttest improvements in perceptual accuracy and consistency. *Perception and Psychophysics*, 70:1060–1067.