https://doi.org/10.2352/ISSN.2470-1173.2021.1.VDA-329

Testing the Value of Salience in Statistical Graphs

Mark A. Livingston[a], Laura Matzen[b], Derek Brock[a], Andre Harrison[c], and Jonathan W. Decker[a]
[a]Naval Research Laboratory; Washington, DC, USA; [b]Sandia National Laboratories; Albuquerque, NM, USA;
[c]Army Research Laboratory; Adelphi, MD, USA

Abstract

Expert advice and conventional wisdom say that important information within a statistical graph should be more salient than the other components. If readers are able to find relevant information quickly, in theory, they should perform better on corresponding response tasks. To our knowledge, this premise has not been thoroughly tested. We designed two types of salient cues to draw attention to task-relevant information within statistical graphs. One type primarily relied on text labels and the other on color highlights. The utility of these manipulations was assessed with groups of questions that varied from easy to hard. We found main effects from the use of our salient cues. Error and response time were reduced, and the portion of eye fixations near the key information increased. An interaction between the cues and the difficulty of the questions was also observed. In addition, participants were given a baseline skills test, and we report the corresponding effects. We discuss our experimental design, our results, and implications for future work with salience in statistical graphs.

***Keywords**: Charts, Diagrams, and Plots; Perception; Cognition; Salient Cues; Highlights; Human-Subjects Quantitative Studies*

Introduction

Saliency is widely considered useful for emphasizing to graph readers the particular message that a graph's author wishes to convey. In this work, we interpret "salient" to mean that an element has a value in some aspect of its appearance (e.g. color, intensity, contrast, shape) that is unique among the elements of the graph. This, in theory, should attract the human perceptual system. There is ample evidence that salient cues in fact draw attention to graph components. However, there is little direct evidence that cues improve or speed up comprehension of the message.

The Value of Salience in Graphs

Statistical graphs are widely used; thus, proper use of salience in their presentation is described in the literature of multiple disciplines. In perceptual psychology, Kosslyn [15] states *The Principle of Salience* (PoS) in his eight Psychological Principles of Effective Graphs:

> The most visually striking aspects of a display will **draw attention** to them (as did this bold type), and hence they should signal the most important information. All visual properties are relative, and thus what counts as "visually striking" depends on the properties of the display as a whole.

This manifests itself in more specific guidelines; here are several examples (all quoted from [15]). "Use more salient labels to label more general components of the display." "The label for the display as a whole should be more salient than the labels for any parts." "Only 25% of wedges in a pie graph should be exploded (if too many are emphasized, the PoS is being violated)." This advice (of not emphasizing too many items) is repeated multiple times. He warns that a bar graph should not vary the salience of individual bars arbitrarily. He gives an example line graph where salience is varied intentionally to draw attention to one line out of three. He recommends to ensure that error bars do not make less stable points more salient than the stable ones (because error bars will be longer when the point is less certain). He advises to ensure that best-fit lines in scatterplots are discriminable and more salient than the points. In discussing color, Kosslyn advises to "make the most important content the most salient." He notes that having every Nth gridline (e.g. 10th) stand out can be helpful, but no gridlines (emphasized or not) should obscure the data. Pinker [25] arrives at similar advice by starting with the reader's goal. A goal invokes graph schema and gestalt processes, wherein salience will determine the likelihood of encoding a graph feature. Encoded features are used to build an interpretation of the information conveyed by the graph; thus, salient items are more likely to influence the interpretation of the graph.

In cartography, Bertin [2] recommends widely-cited guidelines for presentation of graphics. Hegarty et al. [11] study maps and draw the *salience principle* from Bertin. Their formulation suggests that visual attributes such as color or line orientation should be ordered so that important information is visually salient compared to contextual information.

In education, McCrudden and Rapp [22] define *selection* as focusing or directing attention to information in an instructional message. They note that

> [I]f attention is not allocated toward important information, it will not be consciously processed. Similarly, if attention is allocated toward interesting but unimportant information, those contents can disrupt the coherence of the main instructional message.

They define *signaling* as the use of cues to increase the salience of important information. In the work presented here, we focus on the use of two such choices, text labels and color, for this purpose.

Our Contribution

We study the efficacy of making task-relevant information in a statistical graph salient via the use of text labels, a common but understudied practice. We study a broader range of statistical graphs than we generally see in the literature, and we explicitly evaluate different levels of task difficulty. Our work is similar to existing studies on the role of salience in visual information tasks in that we use color and intensity manipulations, and we study expertise as a possible interacting factor.

Despite the above and many similar statements in the literature, we see a surprising lack of direct evidence for improved understanding of statistical graphs in the presence of salient cues. If salient cues help convey the message a graph's author intends for the reader to grasp, then we should be able to gather evidence that supports this. We undertook to design and conduct an experiment to provide direct evidence. We first describe some related work, then detail our experimental design. We present statistical results and discuss the interpretation of these results, along with some potential limitations and extensions to our work.

Related Work

As evidenced by the passages above, there is a widely-held belief in the value of salience in various visualization contexts. In this section, we start our review of related work with user studies and then consider applicable results from eye tracking in visualization. We found only a few user studies in the literature, and they are from diverse application contexts. Statistical graphs are a small portion of the related work.

Graphs and Maps

Carenini et al. [4] selected four *interventions* for bar graphs (drawing on [23]): bolding (borders around bars), de-emphasizing (desaturation), adding reference lines, and adding connected arrows. They studied these in an experiment similar to ours. The first two interventions helped identify individual bars; the other two assisted in identifying bars to be compared. They also varied the onset of the intervention; it was either present at the start of a trial, or it was added to the graph after the graph and then the question appeared in sequence. They also measured users' perceptual speed, visual working memory, verbal working memory, (self-reported) expertise, and locus of control. They used two classes of task: a task to retrieve a value and compare to a group average, and a task to aggregate comparison results. They found that more complex tasks took longer and induced more errors, and delayed onset of the cue induced longer response times. They found that the de-emphasis intervention was best, with bolding and connected arrow next; reference line was no better (statistically) than no intervention. Those with high perceptual speed performed better; those with low verbal working memory performed more poorly. Toker and Conati [27] later analyzed eye tracking data from this experiment. They found that those with low perceptual speed spent more time looking at labels. Those with low visual working memory spent more time looking at the answer choices and button to submit the answer. Those with low verbal working memory needed more time, as they lingered over the legend and question and other textual elements of the graph. Although there is some overlap with the cues we used, we study more diverse cues, graphs, and tasks.

Bera [1] examined two ways colors were used improperly in bar graphs. He defined *overuse of colors* as attracting attention through changes in color between adjacent bars when these changes carried no meaning (a violation of Kosslyn's PoS). He defined *misuse of colors* as attracting attention through color contrast to areas that are not relevant to a task. Both poor designs increased the number and duration of fixations on graph components not relevant to the task. They also delayed the time to the first fixation on the relevant components. While he noted that these fixation patterns would induce greater cognitive effort, he did not find that the increased effort affected performance. Based on his data, we believe Bera used a minimum duration to identify fixations of at least 200 ms. We set a threshold of 100 ms; this minimum has shown to increase the accounted portion of time when viewing a complex geometric stimulus [21]. We also studied types of statistical graphs that Bera did not.

Klippel et al. [13] asked observers to rate subjective similarity of star plot glyphs. They found that the introduction of color reduced the influence of shape on the classifications. Without the color, a salient shape characteristic ("has one spike") was the dominant classification criterion. With the color, more participants classified shapes with "one spike" into different classes. Star plots are an advanced type of statistical graph. We focus our study on mostly simpler graphs, a range of reading tasks, and the use of text labels as well as color.

Madsen et al. [19] had participants answer physics problems using diagrams and unlabeled graphs. People who answered correctly spent a higher percentage of time looking at the relevant areas of the diagram or graph. However, in a larger follow-up study [20], they found no effect of salience manipulations (via luminance contrast) and no interaction between these manipulations and prior knowledge of participants. They also did not find an effect of the manipulations on the percentage of fixation time relative to the relevant area. They postulate that the time window (first two seconds) may have been too long to capture the perceptual effect before cognitive processes (whether correct or incorrect) exerted a larger influence than the perceptual mechanism. Some of their graphs and tasks are similar to some of ours.

Hegarty et al. [11] provide further evidence that proper use of salience can affect understanding of complex visual representations (weather maps). Salience was manipulated by the saturation of the color map for temperature and the thickness of isobars for pressure. In their first experiment, changes in salience affected accuracy only *after* participants learned the meteorology principles; eye fixations were primarily directed by task and domain knowledge. In their second experiment, there was no evidence that participants were drawn to visually salient areas of the maps. They conclude that their "research provides evidence for one principle of cartography and graphics design more generally: Good displays should make task-relevant information salient in a display" (cf. [15]). They also provide evidence for the mechanism of this advantage. Attention was primarily affected by top-down knowledge, and the visual design affected performance through facilitation of the processing of visual features that conveyed task-relevant information. Cartographic maps and the specific use of saturation to create salience are only a small part of our study. Our use of text labels would likely not work well on maps.

Diagrams

In Duncker's radiation problem, readers must surmise from a diagram how to irradiate a tumor without damaging intervening tissue. Using an eye tracker, Grant and Spivey [9] observed an empirical pattern of gazes to the critical area of this diagram that discriminated among readers who hit upon the solution and those who did not. In follow-on work, they found that using an animation to draw attention to this area doubled the rate of success in comparison to giving readers the original static version of the diagram or a version in which a noncritical feature was animated. They proposed that the guided visual attention induced by

the critical animation likely facilitated the correct insight.

Thomas and Lleras [26] replicated this work and added an attempt to force this eye movement pattern with a secondary task. They found that only an order of animated cues that forced the gaze to *cross* the diagram led to greater success. That is, cueing the same locations in a sequence that progressed *around* the outer portion diagram were not sufficient; the gaze had to move in the pattern that was analogous to the solution. Few participants claimed to notice the connection; most appear to have had their thinking influenced by the eye movements covertly.

Lowe and Boucheix [17] found no difference in learning from an animated representation of a piano mechanism using attention-directing cues. They did find a difference in cue obedience (i.e. strict following of the bottom-up cues). They defined *color cueing* as increasing salience of elements which the animation's author wanted to direct attention. They defined *anti-cueing* as lowering salience of elements not desired to be a focus of attention. At the start of an animation, color cueing was more successful at drawing attention than anti-cueing was at limiting the drawing of (incorrect) attention. This difference faded as the animation progressed over time. Although these diagram problems are interesting, they are not directly applicable to our work.

Other Applications of Salience

Kong and Agrawala [14] describe a system to add user-specified graphical overlays to graphs. Overlays included reference structures such as gridlines, highlights such as outlines, redundant encodings such as numerical data labels, summary statistics, and descriptive text annotations.

In designing a system to automatically understand bar charts, Elzer et al. [5] applied the salience of elements as one of the "communicative signals in the graphic itself." They noted highlights in the form of color, shade, or texture, as well as labels used as annotations. They note other sources of salience, such as the tallest bar or the right-most due to being the most recent data on a time-based axis. Of note for us, they adopted "a GOMS-like approach to estimate the relative effort involved in performing a task." Independently, we utilized a Goals, Operators, Methods, and Selections (GOMS) model [3] to identify the most difficult step in completing a task and thus most in need of the reader's attention.

Eye Tracking and Visualizations

Gegenfurtner et al. [7] reviewed eye tracking research that investigated differences in expertise on the comprehension of visual representations. They discuss three theories to account for these differences. They develop their argument by then stating findings regarding gaze that would support each theory. Finally, they assess the accumulated evidence. They found the most complete support for the hypothesis that experts optimize the amount of information processed by neglecting task-irrelevant and, whenever possible, redundant information. Of note for our findings is that they conclude that task complexity modulates the difference between expertise; smaller differences were found for less complex rather than more complex tasks. However, there is scant overlap between the visual representations they reviewed and our representations. Just two (out of 73) of the studies they reviewed included cartographic maps. We will consider this finding in our Discussion in the light of the evidence we contribute.

Harsh et al. [10] tracked users' gaze as they answered science questions using graphs. They found that novices spent a greater percentage of their time reading the question and the answer choices than experts. They further found that experts followed their planned first three steps in reading a graph (two common approaches were: title/caption–variables–data and variables–title/caption–question). While non-experts planned the same progressions, they did not follow this, and those with lower expertise (even within the novice category) were farther from their plan, according to the gaze data.

Experimental Design

Our fundamental goal was to look for evidence for two effects of visually salient cues in statistical graphs. First, do these cues in fact draw readers' attention to targeted areas of information? Second, does their presence improve readers' performance on corresponding response tasks? Our design goals included using validated tests of graph comprehension, increasing statistical power by using a within-subjects design, and measuring fixation data to yield insight about the mechanisms that led to the hypothesized performance improvement. We discuss the design in detail in the rest of this section. Each stimulus consisted of a graph image, the text of the question to be answered, and between two and four answer choices. The stimuli were presented on a desktop PC running Windows 10, using custom software and a Dell U2412MB 24-inch monitor running at 1920x1200 resolution at 60 Hz. Eye tracking data was received from a GazePoint GP3HD eye tracker running over USB3 at 150 Hz. Synchronization of the eye tracker data and the stimulus was accomplished through the Windows `QueryPerformanceCounter` function; this clock tick data is returned with the eye tracker data. We estimated the mean viewing distance as 63 cm, yielding approximately 46.4 pixels per degree of visual angle (averaging vertically and horizontally).

Subject Procedures and Characteristics

After giving informed consent, participants sat down and adjusted the chair, keyboard, and mouse to their comfort. The experimenter then adjusted the eye tracker to capture the participants' eyes, which they saw on the feedback display of the eye tracker's control software. They were asked not to move significantly, in order that their eyes would remain in view of the eye tracker. The GazePoint's nine-point calibration procedure was then run. During this, a white dot swept back and forth across the screen, pausing at each of nine points that form a 3x3 grid. The participant was instructed to focus on the dot the whole time. After this ran, the control software reported how many of the nine points were successfully calibrated; if any point was not calibrated, the procedure was run again. This was necessary only for one participant, and only once.

Next, we started our data logging and custom stimulus software. Our stimulus software first displayed five test points to determine the accuracy of the eye tracking data; specifically, we needed a tolerance for a fixation point to be considered within a region of interest (described below). On the five test points across all subjects, we saw an average of 124 pixels of error, which equals 2.7° of visual angle. We felt this threshold would be too permissive, but we wanted to make sure that we were able to get sufficient tracking data. Thus we chose 2.0° of visual angle as our tolerance for eye tracking data, as described below.

We recorded data from 28 participants (19 male, 9 female).

The range of ages was 22-70, with a mean and median of 43. We grouped highest educational degree into Bachelor's or Associate's degrees (10 participants), Master's or Professional (9), or Doctorate (9). These demographics were used as independent variables in analyses as described in Results. Most participants were native English speakers (including two bilingual); the four who were not native speakers had a minimum of 30 years speaking English. All instructions and questions were in English.

Independent Variables and Stimuli

After calibrating the eye tracker and testing the data quality, we gave each participant a baseline test of graph literacy skill. For this, we used the Graph Literacy Scale [6] (GLS). It asks questions involving bar graphs (four questions), a pie graph (two questions), line graphs (five questions), and an icon array (also called a pictograph or pictogram, two questions). Designed for health care communication, GLS focuses on health care scenarios. But we felt the graph types would yield an accurate test of expertise for VLAT. Both the skills test and the VLAT graphs in the study focus on domains that we anticipated being outside our participants' areas of expertise (which varied). We had hoped to use the GLS to separate our participants into groups of high and low expertise with graphs. However, as noted in the Results, we do not see this separation in our data.

The second and final phase of data collection entailed three visual manipulations of groups of graph-reading tasks (of varying difficulty). Each trial featured a graph and a corresponding question about its data. We chose to adopt all the questions from the Visualization Literacy Assessment Test (VLAT) [16]. In the control manipulation, the graphs were unmodified from their presentation in VLAT. In the other two manipulations, certain task-relevant components in each graph were emphasized with salient cues. Figure 1 shows the control and the two manipulations for one question relating to one graph. The VLAT has many graph types: bar graph (including stacked and 100% stacked types), histogram, line graph, area chart (including stacked type), pie chart, scatterplot, bubble chart, treemap, and a chloropleth map. We edited the text of some questions and answers lightly in ways that we believed would improve their clarity for our participants. In the next two subsections, we describe the salient cues and their construction, as well as the independent variables created as a result. The VLAT has a defined set of question types; we used this as an independent variable in our analysis as well.

Visually Salient Cues

First, we divided the VLAT's questions into three groups. The goal in this division was to create groups of approximately equal overall difficulty. This enabled us to use a within-subjects design to test the effects of different types of salient cues. The first issue was that the VLAT has 53 questions; to make the number of questions divisible by three, we chose to add a question. Inspecting the data published with the VLAT, we observed that the area chart did not have a question of type "Make Comparisons" (using the terminology of [16]). We created a question to fill this slot. To determine approximately equal difficulty for the three groups, we used the VLAT's *item difficulty index*, which is the "portion of the test takers who answered the item correctly" [16] in a pilot study. We estimated the item difficulty index for our new question by comparison to similar questions and the graph type. Based on our participants' performance, we are confident that our estimate was reasonable. We then assigned questions to the groups manually, achieving nearly equal summary statistics for the item difficulty index values in each of the three groups (Table 1), as well as a nearly equal distribution of graph types among the three groups. This enabled us to compare three approaches to salient cues, since the set of questions each participant saw with each salient cue type was of approximately equal (overall) difficulty.

Table 1. *Summary statistics of the item difficulty index for the three groups into which we divided queries, showing that the overall difficulty of the three groups was approximately equivalent. Each group contained 18 queries.*

Group	Mean	Std. Dev.	Minimum	Maximum
1	0.65056	0.25708	0.20	0.98
2	0.65222	0.26906	0.24	0.98
3	0.64778	0.26249	0.15	1.00

Next, we created a GOMS model [3] for each of the 54 questions. The goal was dictated by the question: find the correct answer. The method generally involved a search for the relevant information. The operators varied depending on the type of question and the type of graph. For example, finding the maximum on a line graph required searching for the highest point along the dependent axis. Searching for the maximum of a single variable in a stacked bar graph required searching for the tallest segment of the proper series. Our choice of operators assumed that our readers knew how to decode the visual metaphors employed by each graph type; this assumption may have been incorrect for some graph types for at least some users. When faced with a selection, we chose the one we believed to be the most direct path to a solution. Our GOMS model for the question in Figure 1 is:

1. Read the graph title.
2. Read the dependent (vertical) axis title.
3. Locate (start of) second half on independent axis (July).
4. Scan up to data point for July.
5. Scan the graph to the right of July to determine the pattern.

In the Discussion, we present evidence that our readers did not follow our models, notably ignoring the title on many, most, or even all questions.

Finally, we identified graph components that we felt were central to accomplishing the task of answering the question according to our GOMS model. (We note that different selections in the GOMS model may lead to different graph components being identified.) In the above example, we chose step 3. Again, the graph as it appeared in VLAT served as a control. We drew two sets (modes) of cues designed to draw attention to these critical components[1]. One set, which we shall refer to as *text* cues, relied primarily on text labels. Clearly, these have a spatial extent which sometimes was deliberately used to enhance the cue's ability to assist with steps in the GOMS model. The second set, which we call *color* cues, relied primarily on colored shapes to draw attention to components. Often, these shapes were shapes present in the visual representation; other shapes were drawn over the visual representation. An example of each cue type appears in Figure 1; note that although they have similar positions, the cues are not normalized for intensity or salience by any metric. The independent variable Salience Type, with values "No cues," "Text cues,"

[1] Interested parties may contact the first author for the full set of cues.

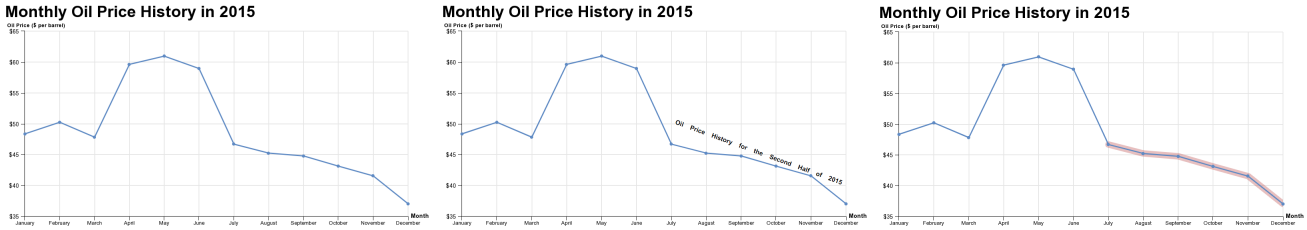


Figure 1. *We designed visually salient cues for the questions in the Visualization Literacy Assessment Test (VLAT) [16]. At the left is a graph with oil prices (y-axis) for the twelve months (x-axis labels) of 2015, shown with the original line graph from VLAT. In the center is the graph with a text cue added which reads "Oil Price History for the Second Half of 2015." At right is the graph with a color cue added to the data for the second half of 2015. These salient cues help direct the reader's attention to the components of the graph needed to answer the question "Over the course of the second half of 2015, the price of a barrel of oil was...?" with answer choices of "increasing," "decreasing," and "staying the same." We measured the effect of visually salient cues like these on the error rate, response time, and fixations of graph readers.*

and "Color cues," refers to these cues. Most graph drawing was done through the creation of specifications in the `HighCharts` language[2]; some was done directly in Adobe Photoshop[3]. We also created two sets of examples of all three cue modes; these were shown, with explanations, to participants to serve as a tutorial, immediately prior to the main task.

Our cues were designed to focus attention on the smallest area that would enable the reader to take the next step in our GOMS model of the solution process. One could argue that highlighting a specific step in the GOMS model would obviate the need to perform any steps prior to it. This would make our choice of highlighting dependent on the order of our GOMS model, which in many cases is not a uniquely valid solution. One could argue that highlighting makes it plainly obvious what the answer is. However, we reasoned that the entire purpose of providing salient cues on top of a graph is to answer a question that the graph author is implicitly placing in front of the reader and demanding that the reader answer. Thus, the entire purpose of a salient cue could be considered to be the author forcing the reader to make a particular interpretation of the graph. Thus, many of our cues in fact do this. Some of our cues could be criticized as not following standard practice. Although we are not aware of a single standard for highlighting elements of a graph, it is true that certain common practices do not appear in our cues and some uncommon ones do. For example, one common highlight for a bar graph is to assign a unique color to a bar of interest. But this cue may not focus the reader directly on the *value* of that bar (which is indicated by the top of the bar only). Thus some of our designs may seem unusual, such as circling the top of a bar. We felt such unusual cues would focus the reader on the perceptual task that most directly leads to accomplishing the steps of our GOMS model. Other cues, such as highlighting an entire sector in a pie graph (which gives its value by the angular size), are more in keeping with conceptions of "standard" practice.

Estimated Question Difficulty

While the mean difficulty of the three manipulations each participant worked through was approximately equal, we saw from the item difficulty index that certain questions were easier than others. We wanted to see if the salient cues interacted with the difficulty of the questions. Thus, we used Jenks' natural breaks [12] on the item difficulty index to partition questions into three groups. This assigned 24 questions to the "Easy" group, the next 15 to the "Medium" group, and the final 15 to the "Hard" group. The measure of quality of the partition is *goodness of variance fit*, which has a range of $[0..1]$. This partition yielded 0.8997 (Table 2), considered to be an excellent division. The partition into groups with equal mean difficulty enabled us to compare performance across the modes of salient cues. This division by difficulty enabled us to study the interaction of the cues with the difficulty of the question.

Table 2. *The division of questions into groups yielded approximately equal difficulty for the three groups. IDI is the item difficulty index in [16]; our added question was set to 0.55. Class Deviation for each class is the sum of squared deviations from class mean and for All is the sum of squared deviations for the array mean. GVF is the goodness of variance fit (used to evaluate Jenks' natural breaks), which indicates an excellent division.*

Class	IDI range	Class Deviation	GVF
Easy	0.75-1.00	0.14318	
Medium	0.47-0.72	0.08729	
Hard	0.15-0.44	0.12309	
All		3.52570	0.89972

Region of Interest Construction

To facilitate analysis of gaze data, we need to define a region of interest (ROI) for each question. We constructed regions with the following procedure. First, we manually selected pixels that were part of the text or color cues. This was done with a combination of the `Magic Wand` and `Rectangular Marquee` tools in Adobe Photoshop. Next, still using Photoshop, the selections for the text and color cues were merged (set union) into a single selection. Finally, this single selection for each question was expanded by 93 pixels (46.4×2, rounded) to give us the 2.0° of visual angle we needed for the error tolerance for our eye tracking data. For a few graphs, the color cue consisted of the outline of a large region (e.g. a sector on the pie chart). In these cases, the above procedure resulted in the ROI having a "hole" consisting of pixels that were surrounded by the region but not within 93 pixels of the boundary (and thus inside the ROI). We opted to fill in these holes (make them part of the ROI), so that a participant who fixated on the center of such a region was considered to have looked at a rel-

[2] `https://api.highcharts.com/highcharts/`
[3] `https://www.adobe.com/products/photoshop.html`

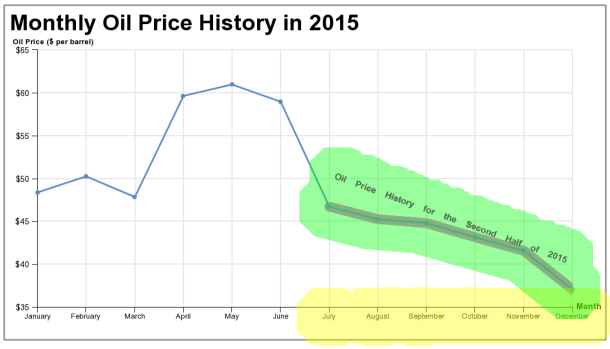


Figure 2. *Gaze-related regions of interest (ROIs) for the analysis of participants' eye tracking data in the three manipulations of the graph reading task shown in Figure 1. These brightly colored ROIs should not be confused with the salient color cues designed for our study. The number and duration of participants' gaze fixations (longer than 100 ms) falling within the color ROIs in this figure were counted. The green region is the primary ROI; the yellow is the secondary ROI. The size of the ROIs shows the error tolerance (2.0° of visual angle) we chose based on tests of our eye tracking system's performance. The gray frame around the graph was not present during the study; it represents the image boundary to illustrate how ROIs may need to exceed the image boundary. Also note that the primary and secondary ROI overlap near the bottom right of the graph; a fixation in this overlap region was considered to belong to the primary ROI.*

evant portion of the graph. Had we not done this, the center of a large, relevant region would have been considered as irrelevant to the question as the plain white background of the graph.

Our salient cues rarely direct attention to all the graph components in our GOMS model for the question. Thus, we labeled the ROIs described in the previous paragraph "primary ROIs" and created a set of "secondary ROIs." Secondary ROIs included graph components that were (potentially) applicable to methods to solve the problem, but were not highlighted by our salient cues. Most often, these included portions of one or both axes (with associated labels) as well as a portion of the legend (if present). When questions arose about the inclusion criteria, the first three authors resolved the differences through discussion. These primary and secondary ROIs enabled the dependent measures related to fixations in the ROIs[4] (Figure 2).

The ROIs were created on images that spanned the entire screen, even though the graph images were smaller. This enabled us to define ROIs that expanded beyond the image boundaries. This allowed us to include fixations that were within the tracker error tolerance of a cue but not within the boundary of the displayed graph image (as often happened for components in secondary ROIs, such as axes and legends).

Dependent Variables

For each data trial, we recorded the (binary) response error, response time, data about the question, and the hardware clock tick counter at stimulus onset and at the time of the response. Eye tracking data was recorded per subject into a log file. As a post-process, this log file was split by stimulus, using the clock tick data to identify beginning and ending data records for each stimulus, and processed to identify fixations. When searching for fixations, we ignored data records for which the tracker reported that either the left or right eye contained invalid data. In concert with the eye tracking error noted above, we used a dispersion metric of 2.0° of visual angle to determine whether a consecutive sequence of valid data records was a fixation. Invalid data records did not end or reduce the time measured for a fixation that otherwise met the criteria; invalid records were treated as if they simply did not exist. We recorded the location, duration, and inclusion in a single ROI for each fixation. In cases where a primary and secondary ROI both contained a fixation (e.g. a salient cue near a relevant axis label), the primary ROI was considered to have been fixated. Each participant saw each question once (on one salience manipulation), so we gathered data from $54 * 28 = 1512$ trials.

Hypotheses

We made the following hypotheses regarding our independent and dependent variables.

1. Error will be lower with salient cues present than without.
2. Response times will be lower with salient cues than without.
3. Question difficulty will interact with the salient cues, with a greater reduction in the error for Hard questions.
4. Question difficulty will interact with the salient cues, with a greater reduction in the response time for Hard questions.
5. Graph literacy will interact with the salient cues, with greater reduction in error for readers with low graph literacy.
6. Fixations will more often occur in the primary ROIs with salient cues present than without.

Hypothesis 6 is perhaps the most obvious and follows directly from much of the Related Work. Although only some of the related work would support Hypotheses 1 and 2, we believed that the visually salient cues would both keep participants on a correct solution process and move them through it faster. Since the cues came directly from the GOMS models for a solution, it followed that the cues would have these positive effects. Similarly, we believed the benefits would be greater when the difficulty of the question mandated greater skill from the graph reader(especially in relation to the skill level the graph reader possessed). Hypotheses 3, 4, and 5 all follow from this belief.

Results

We analyzed the results using the `ezANOVA` package in `R`. We report p-values with Greenhouse–Geisser corrections where needed and effect sizes. Post-hoc correlated t-tests were conducted by hand using intermediate values calculated in a spreadsheet and standard formulas [18]. Error was given a binary value for each trial, and all trials were analyzed. Response time data were analyzed only for those trials on which the response was correct. While this is common practice, it is often not well-justified. The concern is that incorrect answers may indicate a lack of effort on the part of participants and thus not be indicative of a process of working towards a solution. Nielsen and Wilms [24] point out that this is more likely to be true when the response accuracy is near ceiling (i.e. almost perfect). We do not think this is true of all the tasks embodied in the VLAT. However, we did instruct participants that they must answer each question; we removed the option (present in VLAT) to "Skip this question." Thus it is entirely possible that participants read a question, decided that they did not

[4] Interested parties may contact the first author for the ROIs.

know, and simply guessed. Of the 212 errors we recorded (14% of trials), approximately half (102) were by participants whose mean response time did not appear statistically faster on correct responses than on incorrect responses. This behavior could be consistent with guessing. Therefore, in the analysis presented here, we removed incorrect trials when analyzing response time. We note that one could make an analogous argument about fixation data; if the response was incorrect, perhaps the participant did not make a serious effort to identify and process the graph components. However, because we are interested in attention that is controlled below the level of conscious effort, we included incorrect trials in our analysis of gaze data. We leave for future work the application of models for differentiating effort on a per-trial basis [28]. We analyze the percentage of fixations and percentage of time spent in ROIs using the same statistical tests.

Use of visually salient cues had a main effect on the error: $F(2,54) = 2.587$, $p = 0.014$, $\eta^2 = 0.077$. We saw a small reduction in error with both types of cues when compared to no cues (17.5% error). Post-hoc testing showed that color cues (11.1%) were significantly better ($t(27) = 2.875$, $p = 0.008$) and text cues (13.5%) were marginally better ($t(27) = 1.795$, $p = 0.084$). Salient cues also had a small but statistically significant main effect on the response time: $F(2,54) = 4.426$, $p = 0.017$, $\eta^2 = 0.055$. Responses were slightly faster with both the color cues (17.6 sec, $t(27) = 2.786$, $p = 0.010$) and the text cues (18.3 sec, $t(27) = 2.121$, $p = 0.043$) than with no salient cues (21.5 sec). These effects support Hypotheses 1 and 2, respectively.

More interesting is the interaction of the salient cues with the question difficulty (Figure 3). There was a small but significant interaction between question difficulty and the use of cues for error: $F(4,108) = 4.688$, $p = 0.004$, $\eta^2 = 0.050$. This supports Hypothesis 3. We did not find a corresponding interaction between difficulty and salient cues for response time: $F(4,100) = 1.018$, $p = 0.390$. Thus we cannot support Hypothesis 4.

We found a main effect of graph literacy on error: $F(1,26) = 5.058$, $p = 0.033$, $\eta^2 = 0.163$. We found a marginal effect on the response time: $F(1,26) = 4.142$, $p = 0.052$, $\eta^2 = 0.137$. Those with very high graph literacy were slightly more accurate and faster than those with high graph literacy. (See Discussion for the reasoning behind these two classes.) However, we did not see an interaction between graph literacy and the use of cues, for either error: $F(2,78) = 1.598$, $p = 0.209$, or response time: $F(2,78) = 0.394$, $p = 0.676$, so we cannot support Hypothesis 5.

With regard to the gaze data, we found a significant main effect of salient cues on the percentage of fixations in primary ROIs: $F(2,48) = 7.166$, $p = 0.003$, $\eta^2 = 0.118$, as well as on the percentage of fixation time spent in primary ROIs: $F(2,48) = 7.237$, $p = 0.002$, $\eta^2 = 0.111$. Both the text cues ($t(24) = 4.339$, $p < 0.001$) and color cues ($t(24) = 2.423$, $p = 0.023$) increased the percentage of fixations that occurred in primary ROIs and the percentage of fixated time (with nearly identical t-tests) that was spent in primary ROIs (Figure 4). This supports Hypothesis 6. We see that when the graphs were presented with text cues, participants made a significantly *smaller* portion of fixations ($t(24) = 2.281$, $p = 0.032$) and spent a significantly *lower* portion of time ($t(24) = 2.397$, $p = 0.025$) in secondary ROIs (Figure 5).

We did not observe a main effect of educational degree or gender. Age had a marginal effect on error; the 45-54 age bracket generated slightly more error than other brackets. There was a significant main effect of age on response time. The younger a bracket's ages, the faster the response time. These results are in line with typical effects of aging [8] on attention, working memory, long-term memory (we note this includes mathematical procedures), and perception; we do not consider them further.

Discussion

We demonstrated improvement from visually salient, task-relevant cues with a wide variety of graphs, using both text cues and color cues, on error, response time, proportion of fixations, and proportion of fixated time. Intuitively, cues that are designed to draw visual attention to the relevant portion(s) of a visual representation ought to lead to improved performance on corresponding queries through increased attention. To our knowledge, however, this has not been shown with a broad range of types of statistical graphs. Previous work showed benefits of cues akin to our color cues [1, 4] and hypothesized benefits of highlighting text labels [27]. Our results offer empirical evidence for this hypothesis.

While we observed significant differences, we note that the effect sizes were quite modest. So although we can support Hypotheses 1 and 2, we do not find that the improvement is so large that it is certain to have an impact in all applications. Perhaps more notable is the significant interaction between our salient cues and the problem difficulty classification levels we assigned. Although the effect size is small, the lower error on Hard questions with the color cues (25.9%) versus either no cues (42.6%) or text cues (40.7%) is of potential value in many applications. This result supports Hypothesis 3. Although we are hesitant to create guidelines for graph authoring based on a single experiment, we think the most likely practical guideline that may eventually emerge from our research is a recommendation to add salient cues for difficult tasks. The lack of support for Hypothesis 4 is disappointing, but the main effect of adding visually salient, task-related cues on response time is sufficiently interesting and of potential value on its own. Color cues were on average 19.4% faster than no cues, while text cues were 15.0% faster.

To truly understand the differences will require a deeper understanding of the reasons for success or failure of the individual cues we developed. We do not claim that our cues were optimal by any metric. We drew inspiration from various sources of statistical graphics (technical publications, news media, government reports, school textbooks, etc.). We did not survey the literature for example cues or make our cues (cue distribution) representative of what we found. Such an endeavor would be interesting, albeit tedious, and subject to questions of how well the collected samples represent a particular application. We hoped that the baseline graph literacy skills test would yield insight into how well the cues worked for various types of graph readers. But we did not have sufficient diversity of reader expertise to study the effect of skill level. Our readers all demonstrated a high degree of graph literacy per the median split criterion defined for the GLS [6]. Perhaps because of this, we were unable to support Hypothesis 5. It is entirely possible that readers of low graph literacy skill (who are especially important to measure for fields like education) may yet prove to be helped more by these types of cues.

We were able to support Hypothesis 6; we drew readers' attention proportionally more to the primary ROIs, analogous to results with weather maps [11]. This also seems to be a sort of contrapositive result to the results of Bera [1]. We also see it

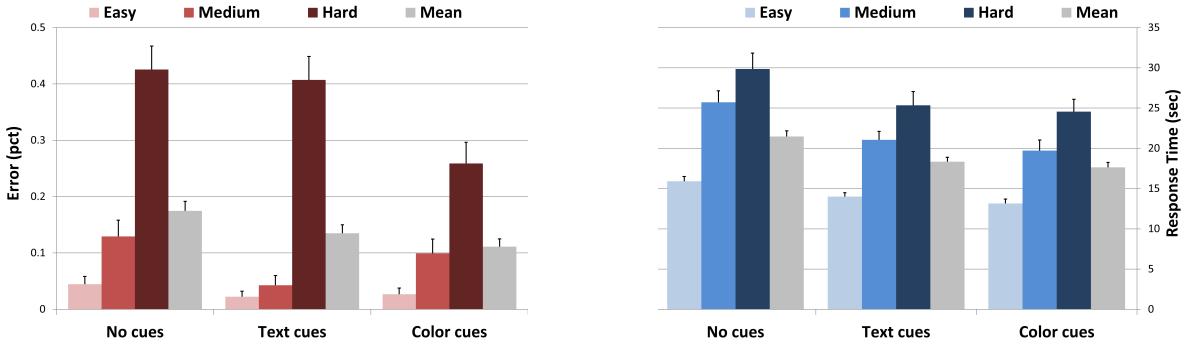


Figure 3. *Left: Error by the type of salient cues and the difficulty of the questions, with the mean for each cue type in gray. The main effect can be seen by the gray bars, whereas the interaction is most notable in the text and color cues for the medium and hard questions. Right: Response time by the type of salient cues and the difficulty of the questions, with the mean for each salient cue type in gray. The main effect can be seen by the gray bars. There was no significant interaction between salient cues and question difficulty for response time.*

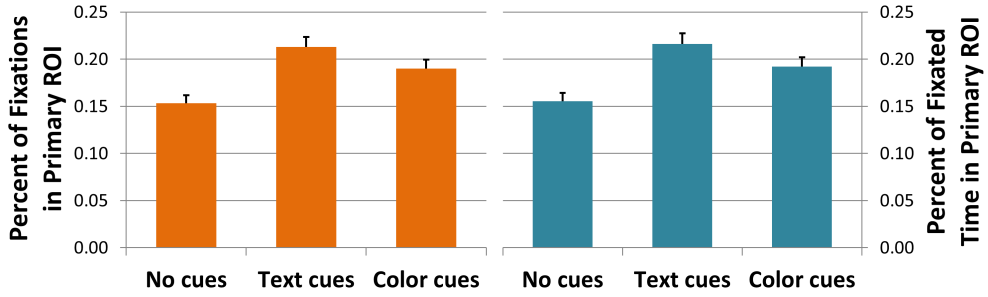


Figure 4. *Left: Percentage of fixations that occurred in primary regions of interest (ROIs) for each of the three salient cue conditions. Right: Percentage of fixation time spent in primary ROIs for each of the three cue conditions. Participants fixated proportionally more and proportionally longer in primary ROIs with both the text and color cues.*

as in concert with the results with animated salient cues for diagrams [9, 26]. One curious observation is that our readers spent very little time looking at the graph title; a few were never observed to have fixated on it at all (on any question). This may have occurred because we showed only one graph and one question at a time; readers may have felt no need to validate the graph's subject through its title. Also, only twelve graphs were used for 54 questions, so familiarity could have reduced the number of fixations even for those who initially read the title. A more complicated task might alter this behavior. We leave for future work an analysis of sequences of fixations; however, we did not ask users to announce their plan for the graph-reading task (as [10] did), although we intend to do this in future studies.

Looking across studies that reported results with respect to task performance, we see a potential pattern by difficulty that echoes the finding of Gegenfurtner et al. [7] with visual representations other than statistical graphs. Here, we consider the studies we reviewed in Related Work. Grant and Spivey [9] and Thomas and Lleras [26] found some evidence of improved performance on a single task that is known to be challenging without assistance; they saw successful responses between 20% and 37% with varied groups (and expertise) on a free-response question. Hegarty et al. [11] found that performance on a task requiring expertise was affected only when participants had training on the task. Their two-alternative, forced-choice task yielded a proportion correct just above chance (when given without instruction). The first study by Madsen et al. [19] analyzed questions that were answered correctly 51% of the time in a pilot study of university students with at least one course on the topics in the questions. Counter to the other findings, those who answered incorrectly appeared to err due to top-down misconception of the task rather than bottom-up distraction of perceptually salient areas. Their follow-up study [20] did not give sufficient data to indicate the difficulty of their tasks. The tasks in VLAT vary greatly in difficulty, with an average of 65% correct in the VLAT pilot data [16]. Bera's tasks were notably easier, with approximately 90% correct responses in his study [1]. Bera did not find an effect on performance on his bar graph tasks. Carenini et al. [4] found greater differences on the simple tasks, but found differences on both simple and complex tasks. They also noted a potential ceiling effect, with 91.4% correct responses overall. Although Madsen et al.'s and Carenini et al.'s results do not appear to fit, we could still ask whether a reliable comparison might show that this (partial) order of difficulty we've hypothesized is reasonable for a greater class of visual representations than Gegenfurtner et al. reviewed. (We note that the work of Grant and Spivey is the only overlap between their review and our discussion.) A true comparison of task difficulty does not exist and would be challenging to design. But if we and Gegenfurtner et al. are correct, then a mostly consistent pattern of harder questions (for the participants, given their expertise) yielding a performance difference from visually salient cues could be discerned. This is an important avenue for future research. We advocate for the use of GOMS or similar modeling approaches to help resolve the influence of problem difficulty.

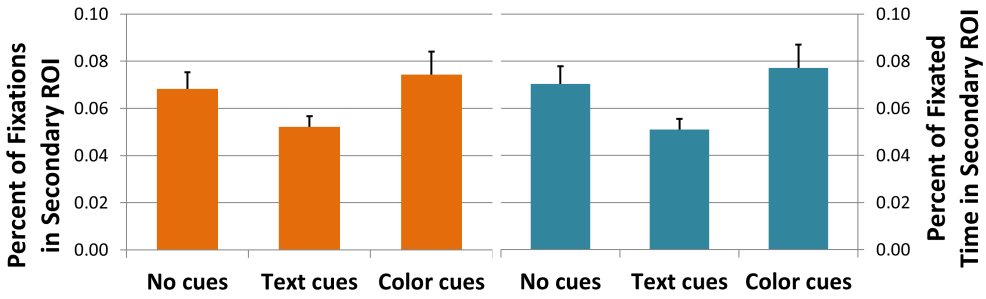


Figure 5. *Left: Percentage of fixations that occurred in secondary regions of interest (ROIs) for each of the three salient cue conditions. Right: Percentage of fixation time spent in secondary ROIs for each of the three cue conditions. Participants fixated proportionally less and for proportionally less time in secondary ROIs with the text salient cues.*

We believe it is reasonable to have expected an interaction between question difficulty and the use of salient cues (Hypotheses 3 and 4). In particular, we anticipated cueing would reduce the nominal degree of effort required for more difficult questions. We conducted analyses of error and response time as a function of the graph type and as a function of the question type. While we did not find significant results, we note that the group sizes are in some cases very small. Such unbalanced designs are unlikely to yield statistical significance. This was not a design goal of our study, but it could potentially be a design goal and an interesting variable for future studies focused on the various graph types and question types we inherited from VLAT [16] or other classifications, such as the often-used types of Bertin [2]. Error shows some differences, and so does response time, but they are clearly not in concert (Table 3). Some cues led to faster response and increased error (color cues for clustering tasks), whereas both text and color cues reduced error but not response time when finding extremes. A similar pattern is observed for the task of finding trends. Despite the lack of statistical significance, Table 3 could reasonably lead us to form hypotheses about effects for particular question types and build experimental designs that rigorously test such hypotheses. A similar argument could be made regarding the graph type, although the analogous table to Table 3 has even smaller patterns that would lead one to hypothesize about an interaction between uses of salient cues and the graph type.

It seems reasonable to conclude that the explanation for performance with the text cues was that our participants fixated proportionally more and proportionally longer in the primary ROIs. This result seems consistent with the results with maps [11] and diagrams [9, 26, 17]. Most importantly for comparison to our work, it is generally consistent with the results for bar graphs [1] and star plots [13]. However, we again note the variety of modes of salience (color and text in our work, versus color, luminance, line thickness, and motion/animation in other work). All these cues should lead readers of visual media to fixate more on the emphasized visual elements. Furthermore, we are not surprised by the slightly greater percentages for the text cues over the color cues; it generally requires more fixations (and thus longer total duration) to comprehend text than geometric shapes [21]. However, we found no correlation between fixations and error, nor betweeen fixations and response time, when the color cues were present. The lack of a consistent effect leads to speculation as to what other influences there are on the error and response time. Hegarty et al.'s [11] and Madsen et al.'s [19] observation that top-down knowledge of the task context was the primary influence on attention may reflect the influences on our graphs and tasks.

Conclusion

We found evidence that two types of visually salient cues can improve the accuracy and response time in graph-reading tasks. This is important evidence that documents the effect of two commonly-used types of cues: text labels and colored shapes. It appears that often, but not always, a mechanism driving this improvement was an increased portion of fixations and fixation time on the intended ROIs. We also found that this effect interacts with the difficulty level of the reading task, wherein the improvement was greater for more difficult questions when using color cues. We did not observe an effect of graph literacy skill within our pool of participants, although this may be due in part to the high degree of skill in our sample population. Nor did we see an interaction between graph literacy skill and salient cues, though one might reasonably hypothesize that cues would be more helpful to novices. However, any position on this conjecture is premature, and we have identified this as an avenue for future work. With the data we have collected, one could perhaps analyze thoroughly to see how to improve the salient cues we gave. We leave this for future work as well.

In conclusion, we have shown that salient, straightforward, task-relevant cues in statistical graphs can improve specific aspects of graph-reading task performance. The sources of this result in the general population cannot be fully differentiated on the basis of our findings. Our approach offers a structured framework for studying additional factors at play in the effective design of graph-reading tasks.

Acknowledgments

The authors wish to thank Joseph Coyne, Ciara Sibley, Noelle Brown, Cyrus Foroughi, the anonymous reviewers, and the anonymous study volunteers. This work was supported in part by the Naval Research Laboratory Base Program.

References

[1] Palash Bera. How colors in business dashboards affect users' decision making. *Comm. of the ACM*, 59(4):50–57, April 2016.

[2] Jacques Bertin with Marc Barbut et al. *Sémiologie Graphique: Les diagrammes, les réseaux, les cartes;, revised edition*. Gauthier-Villars, 1973. Translated as "Semiology of Graphics" by William J. Berg, U. of Wisconsin Press, 1983.

Table 3. *Error and response time broken down by a classification of question types. We collapsed some question types as listed in VLAT [16] into a single category in order to achieve a within-subjects analysis. No significant effects were found, but there are differences in the performance improvements (or lack thereof) between the categories.*

	Error (pct)			Response Time (sec)		
Question Type(s)	No cues	Text cues	Color cues	No cues	Text cues	Color cues
Find Clusters / Find Anomalies	0.081	0.026	0.108	29.272	22.933	17.936
Make Comparisons	0.198	0.205	0.155	23.541	19.714	18.685
Find Extremum	0.117	0.035	0.009	17.178	15.609	15.456
Find Correlations/Trends	0.167	0.087	0.065	19.399	19.126	17.711
Retrieve Value	0.240	0.215	0.188	23.113	20.217	18.293

[3] Stuart Card, Thomas P. Moran, and Allen Newell. *The Psychology of Human Computer Interaction*. Lawrence Erlbaum Assoc., 1983.

[4] Giuseppe Carenini, Cristina Conati, Enamul Hoque, Ben Steichen, Dereck Toker, and James Enns. Highlighting interventions and user differences: Informing adaptive information visualization support. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1835–1844, April 2014.

[5] Stephanie Elzer, Sandra Carberry, and Ingrid Zukerman. The automated understanding of simple bar charts. *Artificial Intelligence*, 175(2):526–555, February 2011.

[6] Mirta Galesic and Rocio Garcia-Retamero. Graph literacy: A cross-cultural comparison. *Medical Decision Making*, 31(3):444–457, May/June 2011.

[7] Andreas Gegenfurtner, Erno Lehtinen, and Roger Säljö. Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, 23:523–552, December 2011.

[8] Elizabeth L. Glisky. Changes in cognitive functions in human aging. In *Brain Aging: Models, Methods, and Mechanisms*, chapter 1. CRC Press/Taylor & Francis, 2007.

[9] Elizabeth R. Grant and Michael J. Spivey. Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, 14(5):462–466, September 2003.

[10] Joseph A. Harsh, Molly Campillo, Caylin Murray, Christina Myers, John Nguyen, and Adam V. Maltese. "seeing" data like an expert: An eye-tracking study using graphical data representations. *CBE–Life Sciences Education*, 18, Fall 2019.

[11] Mary Hegarty, Matt S. Canham, and Sara I. Fabrikant. 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, 2010.

[12] George F. Jenks. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7:186–190, 1967.

[13] Alexander Klippel, Frank Hardisty, Rui Li, and Chris Weaver. Colour-enhanced star plot glyphs – can salient shape characteristics be overcome? *Cartographica: The Intl. Journal for Geographic Information and Geovisualization*, 44(3):217–231, Fall 2009.

[14] Nicholas Kong and Maneesh Agrawala. Graphical overlays: Using layered elements to aid chart reading. *IEEE Trans. on Visualization and Computer Graphics*, 18(12):2631–2638, December 2012.

[15] Stephen M. Kosslyn. *Graph Design for the Eye and Mind.* Oxford University Press, 2006.

[16] Sukwon Lee, Sung Hee Kim, and Bum Chul Kwon. VLAT: Development of a visualization literacy assessment test. *IEEE Trans. on Visualization and Computer Graphics*, 23(1):551–560, January 2017.

[17] Richard Lowe and Jean Michel Boucheix. Cueing complex animations: Does direction of attention foster learning processes? *Learning and Instruction*, 21:650–663, 2011.

[18] Richard Lowry. *Concepts and Applications of Inferential Statistics*. `http://vassarstats.net/textbook/`, Accessed 30 June 2020.

[19] Adrian M. Madsen, Adam M. Larson, Lester C. Loschky, and N. Sanjay Rebello. Differences in visual attention between those who correctly and incorrectly answer physics problems. *Physical Review Special Topics – Physics Education Research*, 8(1), 2012.

[20] Adrian M. Madsen, Amy Rouinfar, Adam Larson, Lester Loschky, and N. Sanjay Rebello. Do perceptually salient elements in physics problems influence students' eye movements and answer choices? In *Physics Education Research Conf., American Institute of Physics Conf. Proc. 1513*, pages 274–277, 2013.

[21] Barry R. Manor and Evian Gordon. Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of Neuroscience Methods*, 128(1–2):85–93, September 2003.

[22] Matthew T. McCrudden and David N. Rapp. How visual displays affect cognitive processing. *Educ. Psychol. Rev*, 29:623–639, 2017.

[23] Vibhu O. Mittal. Visual prompts and graphical design: A framework for exploring the design space of 2-d charts and graphs. In *American Association for Artificial Intelligence Proceedings*, pages 57–63, July 1997.

[24] Simon Nielsen and Inge L. Wilms. Cognitive aging on latent constructs for visual processing capacity: A novel structural equation modeling framework with causal assumptions based on a theory of visual attention. *Frontiers in Psychology*, 5(1596), January 2015.

[25] Steven Pinker. A theory of graph comprehension. In *Artificial Intelligence and the Future of Testing*, chapter 4, pages 73–126. Lawrence Erlbaum Assoc., 1990.

[26] Laura E. Thomas and Alejandro Lleras. Moving eye and moving thought: On the spatial compatibility between eye movements and cognition. *Psychonomic Bulletin & Review*, 14(4):663–668, 2007.

[27] Dereck Toker and Cristina Conati. Eye tracking to understand user differences in visualization processing with highlighting interventions. In *Intl. Conf. on User Modeling, Adaptation, and Personalization, LNCS Vol. 8538*, pages 219–230, 2014.

[28] Steven L. Wise and Christine E. DeMars. An application of item response time: The effort-moderated irt model. *Journal of Educational Measurement*, 43(1):19–38, Spring 2006.

Author Biography

Mark A. Livingston, mark.livingston@nrl.navy.mil, is a computer scientist at the Naval Research Laboratory in Washington, DC, USA. He received his Ph.D. in computer science from the University of North Carolina at Chapel Hill. His research focuses on human factors of interactive graphics systems, recently focused on comprehension and complexity of statistical graphs, plus virtual environments and augmented reality.

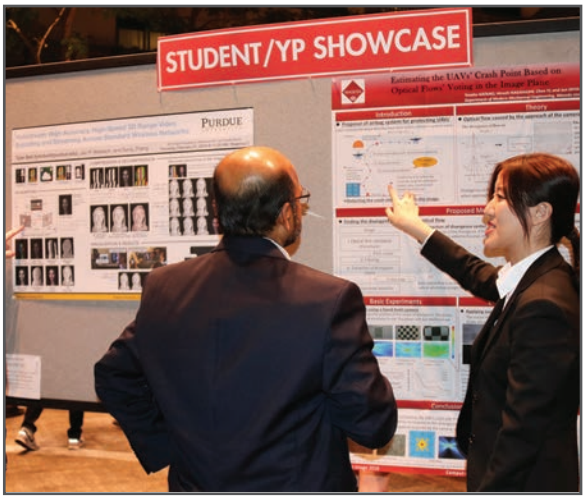
JOIN US AT THE NEXT EI!

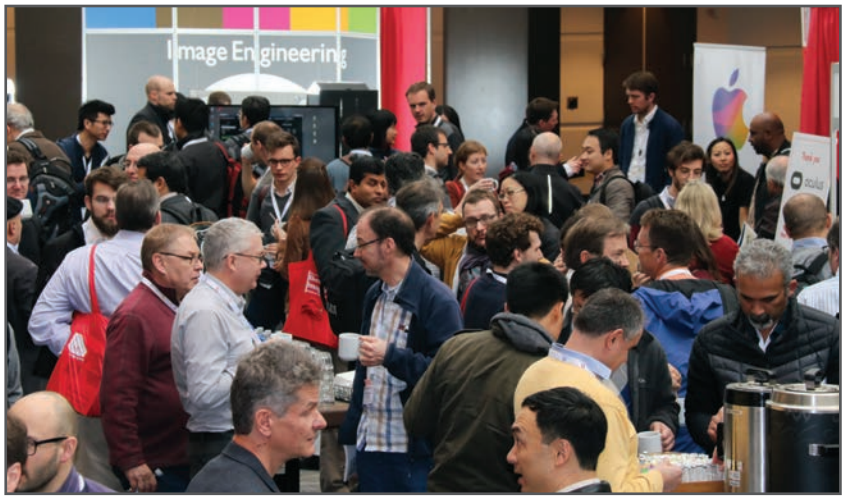
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org