Naval Research Laboratory

Washington, DC 20375-5320

NRL/MR/5581--19-9929

Report on a Query Generation Technique for Measuring Comprehension of Statistical Graphics

Mark A. Livingston

Information Management and Decision Architectures Branch
Information Technology Division

Derek Brock

Navy Center for Applied Research in Artificial Intelligence
Information Technology Division

Jonathan W. Decker

Information Management and Decision Architectures Branch
Information Technology Division

Dennis J. Perzanowski

Navy Center for Applied Research in Artificial Intelligence
Information Technology Division

Christopher Van Dolson
Joseph Mathews

Center for High Assurance Computer Systems
Information Technology Division

Alexander S. Lulushi

Information Management and Decision Architectures Branch
Information Technology Division

August 29, 2019

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
29-08-2019	NRL Memorandum Report	2016 – 2018

4. TITLE AND SUBTITLE

Report on a Query Generation Technique for Measuring Comprehension of Statistical Graphics

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Mark A. Livingston, Derek Brock, Jonathan W. Decker, Dennis J. Perzanowski, Christopher Van Dolson, Joseph Mathews, and Alexander S. Lulushi

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
1E32

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/MR/5581--19-9929

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

10. SPONSOR / MONITOR'S ACRONYM(S)

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In our information-driven society, there is increasing use of statistical graphics to convey information in a variety of settings, including industry, mass media, government operations, and health care. Current methods for assessing a reader's ability to comprehend statistical graphics are custom-written, not widely accepted, usable only once, and/or reliant on subjective interpretations and inferences. We have developed a method for generating queries suitable for evaluating graph comprehension capability. Our method is based on the Sentence Verification Technique (SVT), an empirically validated framework for measuring an individual's comprehension of prose material. Compared to ad hoc methods for testing graph comprehension, our technique is less subjective, requires less manual effort and subject matter expertise, and addresses the essential features of a given graph: values and relationships depicted, frames of reference, and style attributes. The SVT and our derived method combat superficial comprehension by testing what the reader has encoded, as opposed to testing the reader's ability at visual recall or ability to look up data without reaching real comprehension. We motivate and describe our query generation method and report on a pilot study using queries generated with it.

15. SUBJECT TERMS

Graph comprehension
Statistical graphics
Sentence verification technique (SVT)
Quantitative evaluation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Mark A. Livingston
Unclassified Unlimited	Unclassified Unlimited	Unclassified Unlimited	Unclassified Unlimited	44	**19b. TELEPHONE NUMBER** *(include area code)* (202) 767-0380

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

This page intentionally left blank.

1 INTRODUCTION

Statistical graphics have become ubiquitous in modern mass media, scientific and technical publications, and government reports. Thus, some consider the abilities to read, write, and design statistical graphics important for visual or even general literacy [1-3]. This notion of literacy is interrelated with an individual's ability to comprehend information. If literacy means an ability to read or write a language (including a visual one), then it is a critical step to achieve comprehension of written language. If it means to be conversant, then it means to be able to reach comprehension. To measure a person's ability, we must have a reliable and robust test of comprehension. According to Kintsch [4], "[w]e comprehend a text, understand something, by building a mental model." Comprehension research first focused on how this model was organized, progressed to consider how it was constructed, and then focused on iteration and interaction between the construction process and the resulting model. Testing methods for reading comprehension are well-established (albeit with strenuous disagreements).

There are multiple tests of graph literacy or interpretation in the literature, but none seem to be widely-used (although some were introduced recently, cf. Section 2). Standard practice is the subjective development of tests (test materials with items, questions, or queries) by experts in relevant fields, which is a time-consuming process that tends to produce a single test. The effort required to generate suitable test queries from visual information was noted as a concern long ago [5]. We overcome this challenge with an algorithmic approach for queries about information conveyed by statistical graphs.

Given the extensive use of graphs in modern communications and the interest in developing comprehension tests for graphs, an algorithmic method of constructing tests of graph comprehension would be of great value. Covering the range of forms for statistical graphics requires a large corpus of questions [6]. A single test enables graph authors to determine whether a particular graph or set of graphs is understandable by a target population (via testing with representative readers). But a battery of tests (requiring an even larger corpus of queries) could determine the parameters within a class of graphs that make an instance harder or easier to read. A series of tests could help an educator identify whether a particular individual has learned the skills necessary to read a particular type of graph. With a large base of results from such a test battery, a general level of skill required to successfully read a particular graph (akin to reading level or grade level of prose) could be assessed through the graph properties. A precise test battery could even help ascribe the resulting difficulty level to individual properties. For all these reasons, we desire a reliable, robust method of generating not just a single test of graph comprehension, but a large corpus of graph comprehension queries. Further, even test questions custom-written by experts in accordance with standard test procedures may not truly measure comprehension. Our approach is based on a reading comprehension assessment methodology designed to overcome this challenge.

Our primary goal is to develop an algorithm to generate queries to measure comprehension of statistical graphics. The technique for generating queries described in Section 3 is adapted from a validated test construction method for prose reading comprehension known as the Sentence Verification Technique (SVT) [7] and is built on its principles applied to graphs. Applicable features of the SVT are described as they become relevant.

2 RELATED WORK

Most test development strategies for graph comprehension focus on the type of tasks graph readers are asked to do, rather than the effort required to develop queries or the definition of comprehension implicit in queries. Bertin [8] introduced a task taxonomy of elementary (e.g. data extraction), intermediate (e.g. understanding trends), and overall (e.g. comparing trends) query types. This is a common choice [6, 9-13] for distribution of graph tasks, although it does not and cannot lay claim on its own to testing comprehension. In cognitive science, comprehension requires the construction of a mental model [14]; comprehension can thus only be tested by querying this mental model, which in turn requires removal of source material during queries.

Wainer [6] wrote eight multiple-choice questions for each graph on his test, based on Bertin's taxonomy, with two elementary, three intermediate, and three comprehensive tasks. Data was collected from 360 children in

Manuscript approved July 24, 2019.

grades three through five (ages 8-10 years, approximately). About one-third of the third-graders were removed from analysis for scoring at or below chance (guessing) levels. Elementary questions were easiest, and line graphs were harder than tables, bar graphs, and pie graphs. He noted that new test items would need to be developed to explore the design space of the graphs. This is a high cost considering the large design space of graphs.

The Test of Graphing in Science (TOGS) [9] was designed for science students in grades seven through twelve. Its development and use demonstrate several challenges for test development. Test items were validated by a review panel and a validation study (strategies which have been used for other tests [10,13] as well). These reviews often resulted in items being removed or re-written. The Graphing Interpretation Skills Test (GIST) [11] reused three TOGS questions rather than develop new items, decreasing the independence of tests and offering some evidence of the difficulty of writing questions. Multiple other tests later re-used questions from GIST.

Curcio [10] found that scores on her custom-designed graph comprehension test significantly correlated with measures of reading achievement, mathematics achievement, and prior knowledge of the topic, mathematical content, and graphical forms (all collected at the same time). However, our examination of her test material leads us to believe that some questions may have been answered through general knowledge rather than comprehension of the graph. To us, this argues for building a graph comprehension test that does not allow general knowledge to be useful as a method for determining the correct response. The SVT limits application of general knowledge by asking readers to verify agreement of query probes with source material, rather than asking for the truth value of query probes or for repetition of statements of facts presented in source material.

Boy et al. [15] employed a test development method based on evaluation of manually-constructed test items through item response theory (IRT) [16]; their experience illustrates the challenge of writing questions at appropriate difficulty levels for the intended audience. Their first test of line graphs provided more information about below-average examinees. A second test found discrepancies in the ability of questions to discern differences in examinees. A third test on bar graphs yielded a finding that half the questions were either too easy or too hard. While no test development framework is immune to this challenge, the SVT framework mitigates this challenge through a four-fold structure for query probes (Section 3.2).

The Graphing Inventory [12] was designed to test students' ability to comprehend, construct, and critique a range of graphs used in science education. Thus, their test includes not only items to measure general knowledge of graph features and rules for interpretation of relationships depicted on graphs, but also students' ability to integrate scientific concepts with information presented in graphs. They formed their test from an existing database of items and from new items designed to test the integration of scientific concepts. They conducted a pilot test and assembled a team of experts (middle school and high school science teachers, as well as other science education professionals) to review the responses. Misinterpreted questions were revised; items found to have content too easy or too difficult for the target grade level were identified and eliminated. Psychometric properties were test using classical test theory and IRT. The test was found to have high internal consistency and distribution of difficulty among the test items, based on field testing with 460 middle school students.

The Visualization Literacy Assessment Test (VLAT) [13] was developed according to the established procedure of test creation in psychological and educational measurement. The authors defined "visualization literacy" as "the ability to read and interpret visually represented data in and to extract information from data visualizations." The authors then developed several graphs and maps; for each visual representation, they created three to seven questions using Bertin's taxonomy. A panel of five experts deemed only 54 out of 61 questions developed to consist of a task "essential to visualization literacy." One additional item was dropped due to low discriminability found after piloting of the test with 191 volunteers. While VLAT is likely to be useful, the authors reported taking a month to develop test items from twelve source graphs, which were only then given to the expert panel for review and later tested with volunteers. As with our analysis of Curcio's test, our examination of VLAT test materials leads us to believe that some questions also may have been answered from general knowledge.

Yeh and McTigue [17] classified graphical representations on late elementary school and middle school standardized science tests into a variety of visual representations: pictorial illustrations, charts and graphs, cut-

aways and cross-sections, and hybrid representations. In the general category of charts and graphs, they found 199 items, which included items in subcategories of scale diagram, flow chart, maps, tables, and graphs/histograms. In the entire charts and graphs category, 11.6% were judged to be unnecessary for correctly answering the question (according to domain expert judgment). However, they did not provide a breakdown for each subcategory, and one could argue that only the subcategory of "graphs/histograms" fits within our current scope. (Our previous work [18] addressed both tables and node-link diagrams that can be similar in structure to flow charts.) Additionally, their unintentional development and use of questions that can be answered without knowledge of the corresponding source material exemplifies a significant motivation for our source-based, query generation technique. Moreover, this shortcoming has been observed in evaluations done for reading comprehension tests, which have found above-chance accuracy on responses to questions without reading the passage to be comprehended [19]. Whether for prose or for graph comprehension, it speaks to the challenge of writing questions that cannot be answered through general knowledge or logical deduction.

We believe these contributions and results with them show two key challenges for writing tests of graph comprehension. The process becomes quite labor-intensive, as it requires many queries to adequately test many aspects of graph comprehension. Even experts, writing subjective questions, may not realize the difficulty of a query, and it may have to be removed from the test. These challenges emphasize the need for a better way to generate test questions. We thus devised a more rigid, algorithmic query generation methodology for graph comprehension, based on the SVT.

3 NEW TECHNIQUE FOR GENERATING GRAPH QUERIES

The arguments in favor of the SVT for reading comprehension tests all apply well to visual representations of information. As noted, Royer and Cunningham [5] long ago foresaw the possibility of adapting the SVT to visual forms, but argued the difficulty of generating test material was high. We noted this difficulty while manually developing comprehension questions for a node-link diagram [18]. We saw a way to overcome this difficulty with a graph specification language, converting the challenge from one of image manipulation into a set of rules to alter a (textual) graph specification. We developed rules for governing changes to graph specifications; these changes generate the paraphrase, meaning change, and distractor query probes central to the SVT.

3.1 Graph Specification

Our clients make information dashboards for their customers. They use, and therefore we adopted, HighCharts <https://www.highcharts.com/> to build graphs. HighCharts is a JavaScript library intended to ease the addition of interactive graphs to web applications. Options for graph configuration are specified in JavaScript Object Notation (JSON). We have thus far used line graphs, bar graphs, and column graphs (a vertically oriented bar graph, which we shall call a "bar graph" as well). The JSON specification contains a hierarchical set of keys and values, enabling us to manipulate graphical elements systematically. In order to alter an image, we can merely change the values in the text specification. A sample specification and the corresponding graph are given in Table 1 and Figure 1 (next page), respectively; this graph was one of the source graphs used in a pilot study we report in Section 4.

3.2 Graph Query Definitions and Rules

Kosslyn [20] decomposed graphs into five components. The framework "sets the stage, indicating what kinds of measurements are being used and what things are being measured." For most statistical graphs, the framework consists of the axes. The content (or specifier in his earlier formulation [21]) is the representation of the data: points, lines, bars, et al. These "specify particular relations among the things represented by the framework." Labels name the variables, give titles to the graph or axes, and/or create a legend. The background holds a pattern over which other components of the display are presented. A *caption* gives a comment on the display, a short description that explains key terms, or directs the reader's attention. Our graphs thus far have a solid white *background* and no *caption*; neither of these components will be discussed further.

Table 1. A HighCharts graph specification in JavaScript Object Notation (JSON). See Figure 1 for the visual form of this graph.

```
{  chart: {   type:“bar”,
              width:800,
              height:600,
            },
   exporting: { scale:1, },
   credits: { enabled: false, },
   legend: { enabled: false, },
   colors: [ ‘rgb(153,255,153)’, ‘rgb(51,153,51)’, ‘rgb(0,102,0)’, ],
   series: [{   data:[ 67, 58, 54 ],
                name: “Landfill”,
                colorByPoint: true,
                maxPointWidth:75,
                pointPadding:0,
            }],
   title:    { style: {
                       color: “#000000”,
                       font-size: “x-large”,
                       fontWeight: “bold” },
               text: ”Percentage of Garbage going to Landfills”,
            },
   xAxis: [{  categories: “1990”, “2000”, “2010”],
              labels: { style: { color: “#000000”, font-size: “20px”, fontWeight: “bold” }, },
            }],
   yAxis: [{   linewidth:1, gridLineWidth:0, max:100, tickInterval:20,
               title: { style: { color: “#000000”,font-size: “20px”, fontWeight: “bold” },
                        text: “Percentage”, },
               labels: { style: { color: “#000000”, font-size: “16px”, fontWeight: “bold” }, },
            }],
   tooltip: { enabled: false, },
}
```

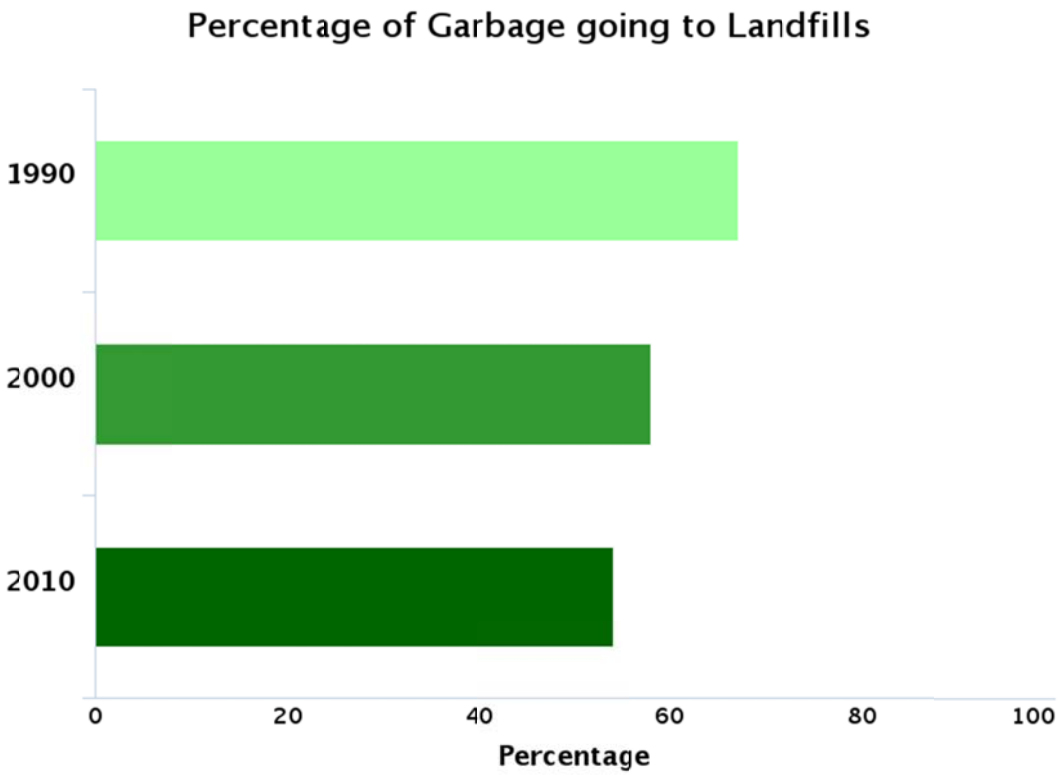

Figure 1. The bar graph corresponding to the specification in Table 1.

Kosslyn's components helped us identify information statements in graphs (analogous to sentences in prose). We next consider how to use these components to algorithmically construct query types for the SVT.

The SVT defines transformations of prose sentences into four types of query probes. Readers must identify whether a sentence probe gives information that was "stated" or "not stated" in a sentence within the source reading. However, Kosslyn's graph components do not convey complete thoughts; rather, they function akin to words in a sentence. On the other hand, "sentences" in graphs are meaningful informational statements that are coordinated, collectively, by the graph's components. A lone bar, divorced from a graph, does not convey informational, but it does when shown together with (at a minimum) a framework and labels. Two bars from the same graph convey an abstract relationship, but fail to make a meaningful informational statement – unless their display is coordinated by a framework and labels. By analogy, points and lines on line graphs require a framework and labels to join them in a construct equivalent to a sentence. When constructing a query, we do not need to include all the data in the source graph; this is analogous to the SVT using a single sentence at a time for a query. We may use one datum or multiple data, to reflect the various information statements that are shown in a graph.

Furthermore, prose sentences can be simple or complex. The analogy for this in a graph is to think of the graph's information statements as assertions that can be combined. In prose, each sentence is a clearly delineated statement; this may be a simple assertion or a complex statement composed of multiple contributing statements. A graph's informational assertions are not clearly delineated, with the exception of taking the whole graph as a statement. Thus, for example, in a bar graph with three bars, one could find perhaps seven information statements. Each bar (together with its coordinating information) is a simple statement (of a data value); three combinations of two bars, and the combination of all three bars, form compound statements (of relationships or trends). This analysis mirrors the tasks specified by Bertin [8] in reading graphs. It also gives us a way to echo these three types of tasks, which we noted above were often the basis of graph comprehension queries on existing tests in the literature; we can use the number of points in a query to emphasize whether a reader recognizes the data values, a single trend, or relationships between trends by varying the amount of data in a query probe.

With the above analysis of what constitutes simple sentence-level information in a graph, we need rules that define alterations to these information statements that come from graphs. This completes the analogy to the sentence transformations defined by Royer et al. [7]. However, there are numerous subtle features of graphs that may be altered without changing the meaning of the graph. Navigating these features is a key contribution to applying the SVT to graphs. In the following subsections, we use two source graphs (Figure 2) as examples of applying (some) rules for transformations from source graphs to graphs appropriate for the four query probe types of the SVT. We emphasize that the query probes given here are one example; a diverse set of query probes could be developed, with emphasis on a number of different features of graphs.

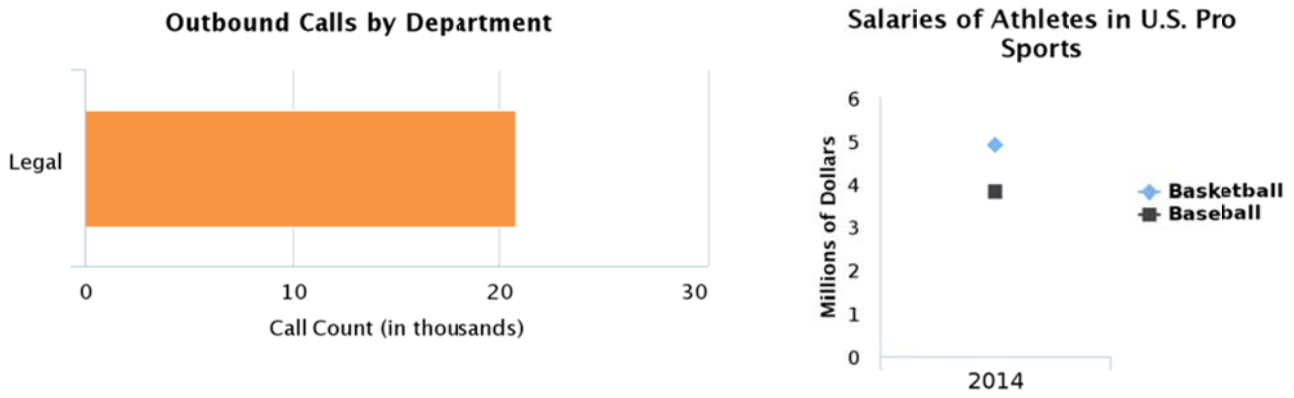

Figure 2. An example bar graph and line graph used in the tutorial instructions for the pilot study (Section 4), used here to demonstrate the four SVT query variations as we adapted them from prose sentences to these two types of statistical graphs.

3.2.1 Original Query Type

In the SVT, an original query type is defined as a verbatim copy of a sentence in the reading passage. Here we take some license with the definition of "verbatim." We assert that features of the graph that do not alter the meaning of the underlying data are not fundamental to the graph. Content may have different style parameters: colors, fill, shapes, borders, et al. Labels may be drawn in different font family, size, style, or position. We note also that the framework could theoretically be changed without altering the meaning, but this would necessarily change the syntax of the content, and Royer et al. [7] recommended avoiding such "gray areas" in queries. Figure 3 shows examples of how some of these considerations are manifested for original query probes. We note that the changes we permit for original query probes reduce the reliance on visual memory for this probe type; this marks a potential difference in our application of the SVT to visual comprehension from the properties for reading comprehension. We believe this potentially strengthens the SVT, with no concern of a negative effect.

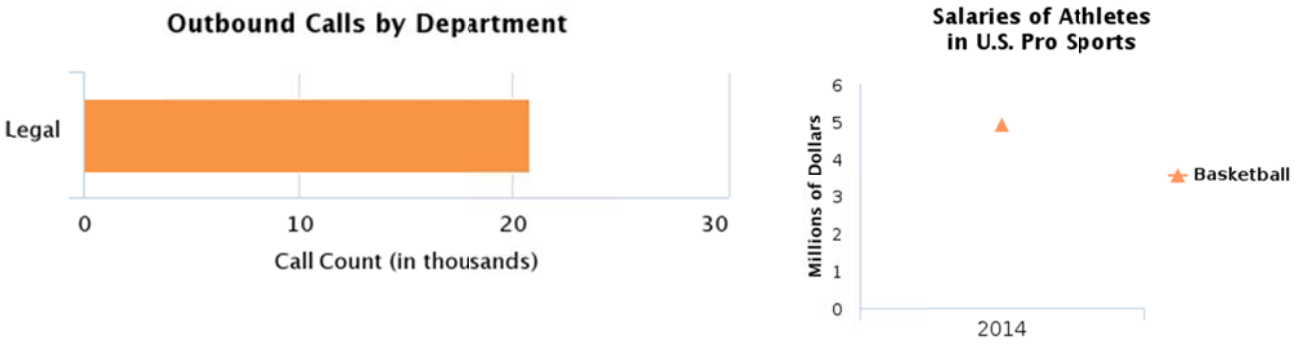

Figure 3. *Original* query probes for the source information graphs shown in Figure 2.

3.2.2 Paraphrase Query Type

An SVT paraphrase query type calls for "as many words as possible to be changed without altering the meaning or the syntactical structure of" the source sentence. All style changes permitted in an original query are also permitted in a paraphrase query (Figure 4), since these changes would not alter the meaning. We also deem rounding of the content values to be acceptable (so long as it moves the content by amounts that do not confuse the value). We argued similarly about smoothing data, but with few data points per graph, we did not adopt this. In retrospect, smoothing data is challenging and we recommend not adopting this change in combination with others. Labels may still have different style; however, a paraphrase should also change the wording of labels when possible, using synonyms or different units for numbers (e.g. converting to scientific notation, or giving numbers in thousands). On this, we must accept subjective judgments about equivalence of the words substituted into labels. As with the application of the SVT to prose, a thesaurus may mitigate this challenge, although the jargon associated with the domain of a graph could create additional complexity (and perhaps limit the applicability of the resulting test to those who can be expected to know the domain). However, with the wide use of statistical graphics, we feel that domain-specific issues are easily avoided without limiting the range of style attributes explored in a test. In the framework, we allow changes to major and minor units (denoted by gridlines and/or tick marks). As with original query types, we choose not to transpose axes, change the range of an axis or convert from linear axes to logarithmic axes. We assert that such changes alter the syntax of the graph. If we decide in the future to relax adherence to Royer's definition, then we may study whether to permit such framework changes.

3.2.3 Meaning Change Query Type

The SVT rule for a prose meaning change is to "alter one word in an original sentence such that the meaning of the sentence is changed." Since we adopt the paradigm that the "words" of a graph are the constituents in the content, labels, and framework, it follows that we should change only one constituent in a way that alters the meaning, and no further changes are permitted. Style changes to these three components are still allowed. No-

ticeable change to a datum (content) is perhaps the most obvious approach (Figure 5, left), though changes to labels (Figure 5, right) or the framework are possible ways to change meaning. One may argue that multiple data changes to maintain a trend may be permitted. We leave this issue for future work. It should be emphasized, though, that a meaning change entails a clear alteration of strictly a single "word" in the source "sentence" used as the basis of the query; the introduction of new content belongs to the distractor query type.

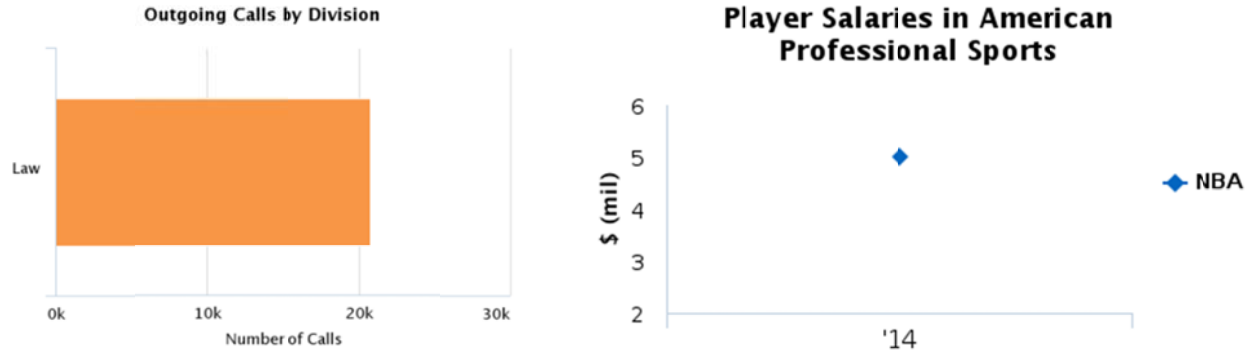

Figure 4. *Paraphrase* query probes for the source information graphs shown in Figure 2.

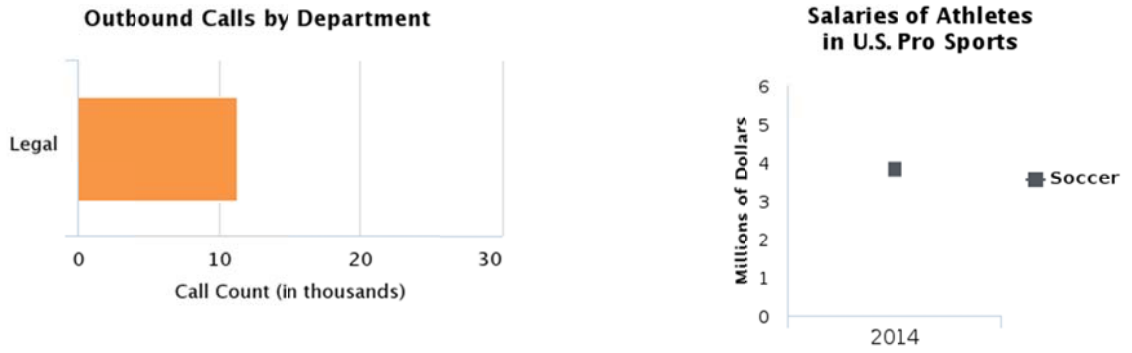

Figure 5. *Meaning change* query probes for the source information graphs shown in Figure 2. Note that this datum corresponds to the bottom point in the source graph, rather than the top point, which was used in the original and paraphrase query probes.

3.2.4 Distractor Query Type

The SVT definition of a prose *distractor* query is "a sentence that is consistent with the general theme of the source material but is unrelated to any original sentence; it should also have the same length, syntactical structure, and conceptual complexity as sentences in the source material." This tells us that we may make multiple changes of the type we may make for a *meaning change*, or introduce new material (Figure 6). However, we must limit ourselves to changes that stay within the topic of the source graph.

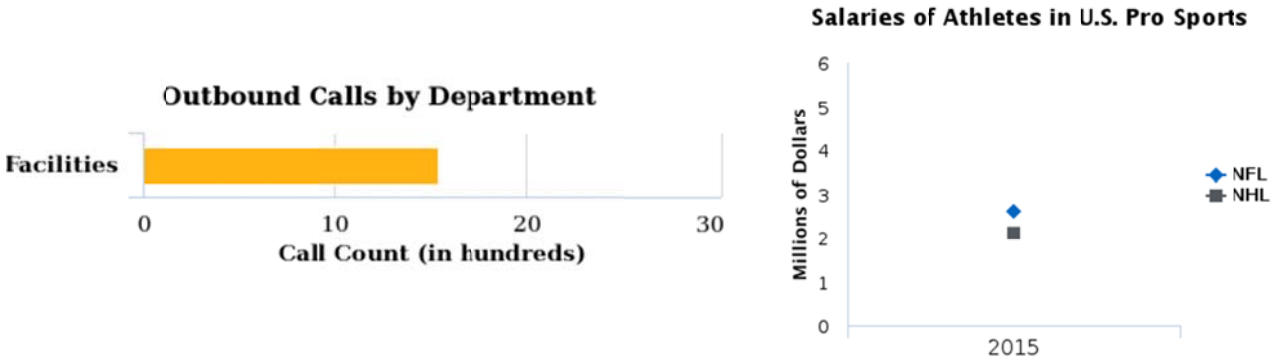

Figure 6. *Distractor* query probes for the source information graphs shown in Figure 2.

4 PILOT TEST OF QUERIES

To validate test items constructed using our method will require field testing them in a population with known graph comprehension abilities. Since one of the motivations for our work is the lack of a widely-validated test, we cannot yet implement a validation study. With recently available tests, such as VLAT [13], perhaps future work can test the consistency of tests developed under different paradigms. This section reports a pilot study using an initial set of queries generated with our method.

To build materials for a pilot test, we constructed nine source bar graphs and nine source line graphs. Some graphs showed data pared down from graphs found in media sources; two were reduced data sets from Shah and Freedman's experiment [22]. Others were constructed from a variety of ideas based on news stories or technical literature. For each graph, we wrote a JSON specification for HighCharts. We then applied the rules (Section 3) to create the four SVT query types (original, paraphrase, meaning change, distractor), still using the specification. Finally, we rendered images of all graphs using HighCharts. We wrote web pages to present the instructions, source graphs, and queries, as well as two diversionary tasks, which are described next. Of the nine graphs of each type (bar and line), one was embedded in the instructions, two were used for practice (described below), and six were used for testing.

To reduce reliance on visual memory, we added two diversionary tasks. We showed participants two images in sequence, each for three seconds. These were intended to interrupt visual pattern memory and were taken from a public database for eye tracking data [23]; they showed a variety of natural and urban imagery, with a few close-up images of common items (e.g. flowers, a sneaker). Participants also read brief, successive excerpts (about 200 words each) from an out-of-copyright novella.

For each trial, participants were asked to study a graph and a prose excerpt (as sources) and to answer corresponding queries; they were asked simply to look at the diversionary images for whatever they found interesting. The prose also gave us a baseline for comparison against the graph comprehension task. Thus, the complete sequence of a data trial was

- show a source graph (minimum time: 30 sec, maximum time: 3 min),
- show a diversion image (3 sec),
- show a blank screen (1 sec),
- show a second diversion image (3 sec),
- show a blank screen (1 sec),
- show a source prose excerpt (also 30 sec to 3 min),
- show a graph query and ask the participant whether the information in this graph query was "stated" or "not stated" in the previous source graph, and
- show a prose query and ask the participant whether the information in this prose query was "stated" or "not stated" in the previous source prose.

After giving informed consent, participants completed a pre-study questionnaire with demographic and background information. Next, they read four pages with instructions for the task: (1) examples of the SVT on prose, (2) our adaptation with a bar graph example, (3) our adaptation with a line graph example, and (4) a brief summary of the procedure. They next completed four practice trials of the above sequence. During this practice, the above sequence was followed by two screens: one for giving the correct answer for the graph query (confirming that the participant was correct or informing the participant of the correct answer), and one for giving the correct answer for the prose query (again, with confirmation or correction). After the practice, a short break was permitted and the participant was asked if he or she had any questions about the procedure. (Participants did not generally ask questions; one asked to clarify what was to be done during the display of the diversion images and was told to simply look at them for whatever may be of interest.) Then the twelve trials were conducted, grouped by graph type (bar or line). Half the participants saw the six bar graph trials as their first group; the other half saw the six line graphs first. Within each group, a Latin square ordered the graphs and another Latin square ordered

the SVT query types. After the first group of queries, another break was permitted; no participants took a break for more than a few seconds. After the second group, another break was permitted. A few participants took a break for a few minutes, but most simply continued on to the second part of the study, which is described in Section 5. After the second part, participants answered a post-study questionnaire, which asked a few more demographic questions.

Control software was implemented in web pages viewed with Google Chrome (version 49 for some data, version 54 for some data – with no effect expected of the version), under Windows 8.1. The volunteer sat at a standard desktop environment and viewed the stimuli on a 28-inch Dell U2412M running at 1920x1200@60Hz. We tracked the participants' gaze with a GazePoint GP3 eye tracker, mounted under the monitor. After consenting to participate, volunteers adjusted their seating position, keyboard location, and mouse location for comfort. Then the participant completed the nine-point calibration procedure for the eye tracker; for some volunteers, some calibration points (including all) were repeated in order for the eye tracker to report successful calibration at all nine points. Findings from the eye tracking data were reported [24] and will not be discussed here.

Twenty-four participants (20 male, 4 female) completed the study; they ranged in age from 19 to 58 (mean and median age were both 38). All self-reported having normal or corrected-to-normal visual acuity and normal color vision. All but one of our participants also reported being heavy computer users. Ten reported that they closely read bar graphs or line graphs for work or personal reasons on at least a weekly basis, and thirteen said that they create such graphs for work or personal projects. Participants came from our laboratory's research and clerical staff; fourteen held a graduate degree. For the procedure as described above, participants took an average of 54 minutes (minimum 31 min, maximum 98 min).

Overall, participants got 92.0% correct on graph queries; they got 82.6% correct on prose queries. We conducted a series of one-way analysis of variance (ANOVA) tests with Greenhouse-Geisser correction to look for statistically significant differences. We found a main effect of SVT query type on response time, for both the graph queries and the prose queries (Table 2(a)). For graph queries: $F(3,69)=7.978$, $p<0.001$, $\eta 2=0.100$ and for prose queries: $F(3,69)=5.638$, $p=0.010$, $\eta 2=0.081$. Royer et al. [7] previously noted that paraphrase and meaning change queries could be expected to be harder than original and distractor queries; this effect on response time gives some evidence of this being the case for our paraphrase queries (but not meaning change queries). Participants spent more time studying source graphs that had more data points on them, summed over all series (Table 2(b)), $F(3,69)=10.604$, $p<0.001$, $\eta 2=0.112$, so we feel confident that our participants focused on the task they were attempting to complete. However, the number of points on the source graph did not show a main effect on accuracy, $F(3,69)=1.442$, $p=0.238$, $\eta 2=0.048$. We also note that the number of cases was not counter-balanced for the number and type of graphs with each number of source data points.

While our graph sources had between three and six data values, our graph queries contained one, two, or three data values. (One query showed all three of the source data values.) We noticed a slight tendency for participants to be more accurate as queries showed more data values, $F(2,46)=2.712$, $p=0.093$, $\eta 2=0.069$ (Table 2(c)). This gives rise to a hypothesis for future studies that more context on the graph query (in the form of more of the source graph being shown) may help participants recall the information content of a graph. There was no significant main effect of sequence number on error (Pearson $r=-0.3789083$, but $t(10)=1.2948$, $p=0.2245$). So, we did not find that the length of the study session limited the performance of our participants. (Note that negative correlation would imply improvement on successive queries.) Again, we note that the number of cases was not counter-balanced for the number and type of graphs with each number of query data points.

None of the questions that we had hoped would give insight into a participant's experience with graphs (How often do you closely read graphs? Do you create graphs? How often do you read news?) showed a main effect on error. We investigated the application of the EZ-diffusion model [25] to the dependent variables of response time and error; however, in the checks for misspecification, it was shown that error responses were slower than correct responses by a statistically significant amount for both graphs and prose. Therefore, application of the EZ-diffusion model is not recommended.

Table 2: (a) SVT query type had a main effect on response time (shown in seconds) for both graph and prose queries. (b) The number of data points on a graph source had a main effect on the study time. We enforced a minimum study time of 30 sec. (c) The number of data points on a query graph showed a tendency to yield more accuracy (lower error) with more data points. RT = response time

(a)	Graph Queries		Prose Queries	
Query Type	RT (sec)	Std. Dev.	RT (sec)	Std. Dev.
Original	15.7	10.6	10.3	5.3
Paraphrase	18.4	11.3	13.8	9.5
Meaning Change	14.5	8.9	10.1	6.7
Distractor	11.7	7.2	10.4	7.5

(b) Number of Source Data Points	Number of Graphs	Study Time (sec)	Std. Dev
Three	2 bar, 2 line	36.3	0.7
Four	3 bar, 1 line	40.7	1.3
Five	0 bar, 1 line	39.9	1.6
Six	1 bar, 2 line	45.4	1.7

(c) Number of Query Data Points	Number of Graphs	Error (pct)	Std. Dev.
One	1 bar, 1 line	0.125	0.334
Two	3 bar, 4 line	0.089	0.286
Three	2 bar, 1 line	0.028	0.1065

5 PILOT SESSION REVISITING GRAPHS AND QUERIES

The second part of the study consisted of successively revisiting the twelve source graphs and their respective graph queries. Each graph and its corresponding query were shown side-by-side as a pair on a separate page. Participants were reminded of their response to the query in the first part of the study (i.e. "stated" or "not stated"). For each pair, participants then responded to these four requests (type-written text for the first two requests, radio button selection for the third and fourth requests):

- Describe the information contained in the source graph.
- Describe the differences between the source and query graphs or tell us you see no differences.
- Answer again whether the query graph gives information stated or not stated on the source graph.
- Identify which of the four query types explained earlier you believe the query graph to be.

The names and definitions of the four query types were also shown on each page. There was no time limit for this task. However, if participants typed fewer than five words for either of the descriptions, they were required to type more. If participants typed fewer than 20 words for either of the descriptions, they were asked to type more, but they could continue to the next graph without entering more words. For this second part of the study, participants needed an average of 39 additional minutes (minimum 16 min, maximum 92 min). We now present results for each of these questions, under headers that summarize the question.

Describe the information contained in the source graph: Two of the authors independently created a scoring rubric for the description of graphs; these were merged into the following rubric.

1. reference to the graph title, 1 point
2. reference to the category labels on the independent axis, 1 point (if present)
3. reference to the labels in the legend for data series, 1 point (if present)
4. reference to the data values, via individual values or comparison/trends across values, 1 point

Items 1 and 4 were present in all graphs; items 2 and 3 were present in most, but not all, graphs. However, all graphs had at least one of items 2 and 3. Thus, the possible score was either three or four points for each graph. The first author scored all graph descriptions against this rubric.

There was no significant main effect of sequence number on the completeness of the graph descriptions; readers got slightly but not significantly better as the sequence progressed (Pearson $r=0.225$, but $t(10)=0.730$, $p=0.482$). Here positive correlation indicates that more of the required elements were listed, so we have no evidence that fatigue was a limiting factor, despite concerns about the length of time some individuals needed to finish this task. There was no significant difference on this task for bar graphs versus line graphs ($F(1,23)=0.790$, $p=0.383$, $\eta 2=0.004$); responses were scored on average at 93.2% for bar graphs and 92.1% for line graphs. The number of points on the source graph had a statistically significant effect on the completeness of descriptions ($F(3,69)=4.791$, $p=0.00464$, $\eta 2=0.051$). However, the range of performance is small (92% to 97%) and not ordered by the number of points, so we merely note this as a potential issue for future research.

Describe the differences between the source graph and query graph: Lists of changes were generated from an objective text difference listing between the source graph specification and the query probe graph specification (both in JSON; see Sec. 3.1). This list was annotated for whether the change would alter the information content of the graph or not. The lead author scored each response against the list of changes.

There was no significant main effect of sequence number on the completeness of the list of differences between the source and query graphs. Again, participants got slightly but not significantly better as the sequence progressed (Pearson $r=0.340$, but $t(10)=1.142$, $p=0.280$). As with the descriptions, we have no evidence that fatigue was a limiting factor. There was no significant difference on this task for bar graphs versus line graphs ($F(1,23)=0.882$, $p=0.357$, $\eta 2=0.004$); responses were scored on average at 68.0% for bar graphs and 69.5% for line graphs.

The use of appearance changes had a main effect on the completeness of the list of information-carrying changes: $F(1,23)=11.949$, $p=0.00214$, $\eta 2=0.141$. On graphs with information changes, readers gave more complete descriptions of the information-carrying changes when there were no appearance changes than when there were appearance changes. Again, we did not see evidence of a fatigue effect, although there was less to type when there were no appearance changes. In cases of distractors, readers may have lost a point because their answer may not have conveyed to the judge (first author) that, although they mentioned the different labels, they also understood that the data itself was different. Thus, some of this difference attributed here to the use of appearance changes may be due to imprecision in our scoring rubric.

Answer again whether the query graph gives information stated or not stated in the source graph: Readers changed their answer only 23 times out of 288 trials. We saw no differences in this limited action due to chart type ($F(1,23)=0.857$, $p=0.364$, $\eta 2=0.022$) or SVT query type labeled by the authors ($F(3,69)=1.673$, $p=0.181$, $\eta 2=0.049$). Readers' revised responses were almost universally correct; only two errors were made on this query out of 288 trials.

Identify which of the four query types you believe the query graph to be: We applied confusion matrices [26] to the question of what type of query our participants thought the various queries were (Table 3). The matrix shows the number of responses in which a participant classified (the standard terminology for this is the "predicted" class) a query probe as a particular type against the correct type as we (the test authors) labeled each probe (with the term "actual"). We give two confusion matrices, one for queries in which appearance changes occur between the source graph and the query probe graph and one for queries in which they do not (cf. Section 3.2.1 regarding non-data-carrying parameters such as color or font size). Thus, the entries in each matrix sum to 144 (half the trials) and entries in each row sum to 36 (one quarter of the 144 trials in each matrix were of each SVT query type). If participants always correctly identified the query type, then the four cells (along the main diagonal from upper left to lower right of the matrix) for which the predicted type equals the actual type would contain 36 and all other cells would contain 0.

Table 3: Confusion matrices for the type of query that participants believed each of our queries to be show some confusion about the type of query, both (a) with style changes (font, colors, use of borders, et al.) and (b) without style changes.

(a) With Changes	Predicted (Participants' responses)			
Actual (Experts' label)	Original	Paraphrase	Meaning Change	Distractor
Original	21	14	1	0
Paraphrase	11	25	0	0
Meaning Change	0	0	33	3
Distractor	0	0	6	30

Accuracy = 0.757, 95% CI = (0.679, 0.825)

(b) Without Changes	Predicted (Participants' responses)			
Actual (Experts' label)	Original	Paraphrase	Meaning Change	Distractor
Original	33	3	0	0
Paraphrase	11	25	0	0
Meaning Change	0	1	33	2
Distractor	0	1	2	33

Accuracy=0.861, 95% CI = (0.794, 0.913)

Given that our participants were generally unfamiliar with the SVT prior to our pilot test, some difficulty in identifying the SVT type was expected. Both confusion matrices show that participants had trouble differentiating between original and paraphrase query types. This seems understandable, but should not have been so difficult when the definitions and both source and query probe were visible. This is an issue that requires further investigation. Most interesting to us is the difference between the matrices. Specifically, because of our desire to test for comprehension as opposed to simple visual memory, we hoped to examine differences in participants' ability to discriminate between query types on which we made style changes (to non-data carrying parameters, such as color or font size) in the query graph (relative to the source graph) and those where we did not make such changes. Although the accuracy was higher on queries for which style changes were not made, the respective 95% confidence intervals overlap, so no statistical conclusion can be made. Furthermore, we did not counterbalance this against the number of query points or the graph type. We believe that changes in these aspects of the graph could change the difficulty of recognizing the types of changes. Therefore, conclusions regarding the difficulty of recognizing the query type as a function of the use of appearance changes would be premature. This is, however, an important issue for future work, since it informs the validation of these query types for graphs.

6 DISCUSSION AND CONCLUSION

We believe that our adaptation of the SVT provides a foundation for developing reliable and robust graph comprehension tests. By combining the SVT structure, a graph specification language, and a taxonomy of graph

components, we can systematically vary graphs within the boundaries defined by the SVT. The SVT's foundation, grounded in cognitive theory, thus applies to our adaptation. The SVT query types were designed to defeat a solution of relying on rote memory. The taxonomy for graph components enables our adaptation to provide a mostly objective construction (Section 3) for a comprehension query. The specification language enables us to transform a text language rather than a graph image. Also, we believe that the combination of the taxonomy and the SVT structure will eventually enable us to compare the difficulty (level of comprehension in a given population) of varied attributes and styles of graphs.

As stated above, our primary goal in this work was to develop an algorithmic method for generating tests of graph comprehension. To that end, we adapted the methodology of the SVT, selected a graph specification that fit our purposes and our clients, and developed rules for generating queries of each type mandated by the SVT. Furthermore, we conducted a pilot study, with the goal of showing that the visual form of the SVT was functional (i.e. that participants understood the task and that queries were generally found to be reasonable). Subjectively, we found that readers generally believed that they understood the task in the resulting graph comprehension test, and they objectively demonstrated comprehension of the graphs. A far larger study will be needed to fully assess the validity of our approach, however, and this must be left for future work.

We also collected eye tracking data in the pilot study; we noted [24] that the pattern of fixations does not match the patterns that are typical for natural imagery. This leads to a hypothesis that people have distinctive patterns for reading statistical graphs; this has been noted in other work [28] and is an area for further study.

We ultimately seek to develop objective, extensible metrics by which we can measure how difficult graphs are to comprehend. As a first step, we have a reliable and algorithmic method through which we can generate tests of comprehension of statistical graphics. There are numerous obvious extensions to our first effort. We began with bar, column, and line graphs because they are frequently used by our clients, but we plan to include other types of statistical graphics (e.g. pie graphs and scatterplots). As we have previously demonstrated [18], the SVT may be adapted for more general visual representations of relational information. Eventually, we expect to include more complex graphs, interfaces composed of multiple graphs, and animated and interactive graphs in our research.

Acknowledgements

The authors wish to thank Mike Royer, Joseph Coyne, Priti Shah, Michael Svec, Autumn Toney, and the pilot study volunteers. This research was supported by the Naval Research Laboratory Base Program.

References

1. W-M. Roth, "Reading Graphs: Contributions to an Integrative Concept of Literacy," Journal of Curriculum Studies 34(1):1-24 (2002)
2. M. Galesic and R. Garcia-Retamero, "Graph Literacy: A Cross-Cultural Comparison," Medical Decision Making 31(3):444-457 (2011)
3. K. Börner, A. Maltese, R.N. Balliet, and J.Heimlich, "Investigating Aspects of Data Visualization Literacy using 20 Information Visualizations and 273 Science Museum Visitors," Information Visualization 15(3):193-213 (2016)
4. W. Kintsch, "Comprehension: A Paradigm for Cognition," Cambridge Univ. Press (1998)
5. J.M. Royer and D.J. Cunningham, "On the Theory and Measurement of Reading Comprehension," Technical Report No. 91, Univ. of Illinois at Urbana-Champaign (1978)
6. H. Wainer, "A Test of Graphicacy in Children," Applied Psychological Measurement 4(3):331-340 (1980)
7. J.M. Royer, C.N. Hastings, and C. Hook, "A Sentence Verification Technique for Measuring Reading Comprehension," Journal of Reading Behavior 11(4):355-363 (1979)
8. J. Bertin, "Sémiologie Graphique, 2nd ed.," Gauthier-Villars (1973). English translation: W.J. Berg, "Semiology of Graphics," Univ. of Wisconsin Press (1983)
9. D.L. McKenzie and M.J. Padilla, "The Construction and Validation of the Test of Graphing in Science (TOGS)," J. of Research in Science Teaching 23(7):571-579 (1986)
10. F.R. Curcio, "Comprehension of Mathematical Relationships Expressed in Graphs," Journal for Research in Mathematics Education 18(5):382-393 (1987)
11. M. Svec, "Improving Graphing Interpretation Skills and Understanding of Motion using Microcomputer Based Laboratories," Electronic Journal of Science Education 3(4) (1999)
12. K. Lai, J. Cabrera, J.M. Vitale, J. Madhok, R. Thinker, and M.C. Linn, "Measuring Graph Comprehension, Critique, and Construction in Science," Journal of Science Education and Technology 25(4):665-681 (2016)
13. S. Lee, S-H. Kim, and B.C. Kwon, "VLAT: Development of a Visualization Literacy Assessment Test," IEEE Transactions on Visualization and Computer Graphics 23(1):551-560 (2017)
14. T.A. Van Dijk and W. Kintsch, "Strategies of Discourse Comprehension," Academic Press (1983)
15. J. Boy, R.A. Rensink, E. Bertini, and J-D. Fekete, "A Principled Way of Assessing Visualization Literacy," IEEE Transactions on Visualization and Computer Graphics 20(12):1963-1972 (2014)
16. F.B. Baker, "The Basics of Item Response Theory, 2nd. ed.," ERIC Clearinghouse on Assessment and Evaluation (2001)
17. Y.Y. Yeh and E.M. McTigue, "The Frequency, Variation, and Function of Graphical Representations within Standardized State Science Tests," School Science and Mathematics 109(8):435-449 (2010)
18. M.A. Livingston, D. Brock, T. Maney, and D. Perzanowski, "Extending the Sentence Verification Technique to Tables and Node-link Diagrams," Proc. Of Applied Human Factors and Ergonomics (2018)
19. J.M. Keenan, "Measure for Measure: Challenges in Assessing Reading Comprehension," In "Measuring Up: Advances in How We Assess Reading Ability," R&L Education (2012)
20. S.M. Kosslyn, "Graph Design for the Eye and Mind," Oxford University Press (2006)
21. S.M. Kosslyn, "Understanding Charts and Graphs," Applied Cognitive Psychology 3(3):185-225 (1989)
22. P. Shah and E.G. Freedman, "Bar and Line Graph Comprehension: An Interaction of Top-Down and Bottom-Up Processes," Topics in Cognitive Science 3(3):560-578 (2011)
23. T. Judd, K. Ehinger, F. Durand, A. Torralba, "Learning to Predict Where Humans Look," IEEE International Conference on Computer Vision (pgs. 2106-2113) (2009)
24. A. Harrison, M.A. Livingston, D. Brock, J. Decker, D. Perzanowski, C. Van Dolson, J. Mathews, A. Lulushi, and A. Raglin, "The Analysis and Prediction of Eye Gaze when Viewing Statistical Graphs," Proc. of Augmented Cognition. Neurocognition and Machine Learning. Springer LNCS Vol. 10284 (pgs. 148-165) (2017)
25. E-J. Wagenmakers, H.L.J. van der Maas, and R.P.P.P. Grasman, "An EZ-Diffusion Model for Response Time and Accuracy," Psychonomic Bulletin & Review 14(1):3-22 (2007)
26. R. Kohavi and F. Provost, "Glossary of Terms for Special Issue on Applications of Machine Learning and the Knowledge Discovery Process," Journal of Machine Learning 30:271-274 (1998)
27. M. Wilson, "On Choosing a Model for Measuring," Methods of Psychological Research Online 8(3):1-22 (2003)
28. L.E. Matzen, M.J. Haass, K.M. Divis, M.C. Stites, "Patterns of Attention: How Data Visualizations Are Read," Proc. of Augmented Cognition. Neurocognition and Machine Learning. Springer LNCS Vol 10284. (pgs. 176-191) (2017)

APPENDIX: ON THE DIFFICULTY OF QUERIES

One way to visualize the trade-off between the ability level of the respondents and the difficulty of test items is through the item-person map, often known as a Wright map [27]. The Wright map consists of two histograms, which are built as follows. The percentage correct for each person is converted to a logit score, defined as the log(p / (1 – p)), where p is the percentage correct. For example, the standard to be considered better than guessing on a two-alternative forced-choice test is 75% or 0.75. This yields a logit score of log(0.75/0.25)=0.477. Similarly, the percentage correct for each test item is converted to a logit score.

Each of these two sets of logit scores (one set for all respondents, one set for all test items) are tabulated separately into histograms, typically plotted along a central axis which shows the logit score, then positioned back-to-back, so that the histogram bars extend from a center axis (although we plot this shared axis on the outer portion of the figure). Mean performance among all respondents and among all test items, standard deviations of these means, and the standard for passing the exam are also annotated on the plot. (However, we use half the standard deviation for the graph queries, so that the range required to show the standard deviation value is not larger than required to show the histogram.)

Ideally, each histogram would exhibit a Gaussian (normal) distribution, and the range of item logit scores would cover the range of respondents' logit scores. In other words, these two distributions should be centered at approximately the same logit value and be nearly mirror images of each other. Any offset of the histograms (or of individual queries or participants) indicates a difference in the difficulty or ability (respectively) of queries or participants. Because we are not prepared to compare performance on the graph queries with performance on the prose queries, we created two Wright maps, one for the prose queries and one for the graph queries.

The resulting maps (Figure 7, page 20) place the range of participant performance high on the scale and the difficulty of queries low on the scale. Because the mean of participant performance was clearly higher than the mean error rate of the queries in both Wright maps, the maps clearly show that the queries were too easy for our participants, and quite remarkably so for the graph queries. This was what we expected to see, since more than half our participants held a graduate degree, and both the graph and prose queries were designed to be simple. The prose source passages had Flesch-Kincaid grade levels ranging from 4th to 10th grade; all participants were above this education level. The graphs contained no more than six data points, and should thus have been relatively easy to comprehend. As noted above, we hope that this work will lead to a reliable and scientifically valid measure of the reading level of a graph. Until we have such a measure, the Wright map is perhaps the best indication of difficulty for graph queries, despite relying on having conducted a test with a sample population.

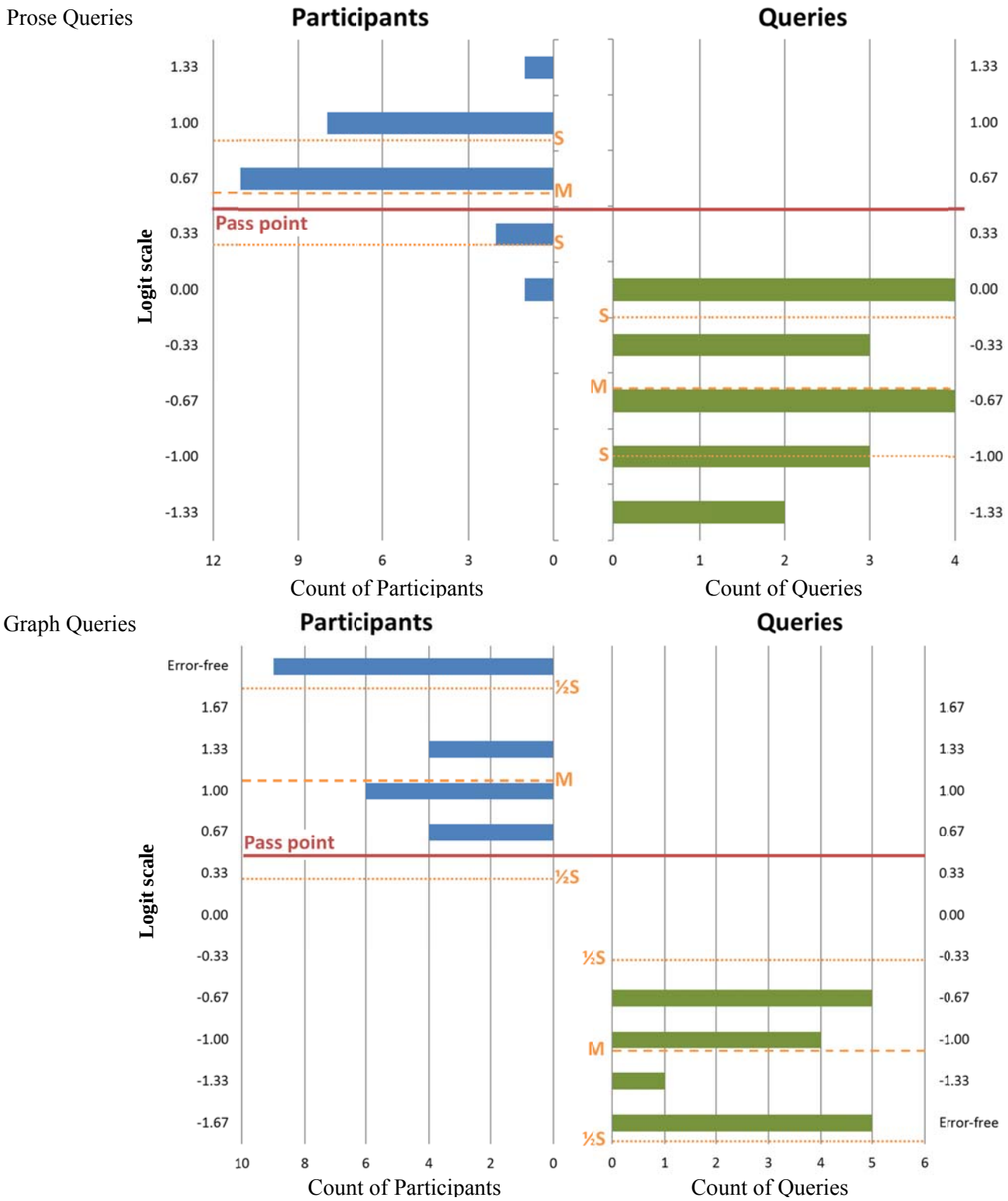

Figure 7. Wright maps for the prose queries (top) and graph queries (bottom) in our study. The graphs make clear that the queries were well beneath the respective ability levels of (nearly) all participants in our study for both prose test items and graph test items. This was expected with the low grade level of the prose source, simple graphs (no more than six data points), and our participant population (over half of whom held a graduate degree). Means and standard deviations for participants and queries are indicated on the map. For prose queries, these are at horizontal lines labeled M and S. For graph queries, we show half the standard deviation, to reduce the range required of the logit axis; the lines are indicated by M for means and ½S for half of the standard deviation. The pass point was set at a percentage correct of 0.75 (logit(0.75)=0.477),, which is the standard threshold for performance considered better than guessing on a two-alternative forced-choice test.