

A Query Generation Technique for Measuring Comprehension of Statistical Graphics

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

Naval Research Laboratory, 4555 Overlook Ave SW, Washington, DC 20735,
USA

{mark.livingston,derek.brock,jonathan.decker,
dennis.perzanowski,joseph.mathews}@nrl.navy.mil,
christopher.vandolso@navy.mil, alulushi657@gmail.com

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 therefore our 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.

Keywords: Graph comprehension · Sentence Verification Technique (SVT) · Statistical graphics · Quantitative evaluation

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 perhaps design statistical graphics important for visual or even general literacy [1–3]. An essential component of literacy is an individual’s ability to comprehend information; to know whether a person has achieved comprehension (or literacy), 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 structured, progressed to

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2020

W. Karwowski et al. (Eds.): AHFE 2019, AISC 963, pp. 3–14, 2020.

https://doi.org/10.1007/978-3-030-20135-7_1

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. Sect. 2). Standard practice is subjective development of test items 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 communication was noted as a concern long ago [5]. We overcome this challenge with a more algorithmic (but not automated) approach.

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 of users (via testing with representative readers). But a battery of tests (requiring an even larger corpus of queries) could determine the parameters of 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 and 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 as well.

Our primary goal is to develop an algorithmic method of generating queries to measure comprehension of statistical graphics. The technique for generating queries described in this paper 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. For brevity, applicable features of the SVT are described below (Sect. 3) as they become relevant to our presentation.

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 tasks. 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.

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. Multiple tests (including [11]) reuse TOGS questions rather than develop new items, decreasing the independence of tests and offering some evidence of the difficulty of writing questions.

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 controls for general knowledge, which the SVT does by verifying agreement of query probes with source material.

Boy et al. [15] employed a test development method based on evaluation of manually-constructed test items through item response theory [16]. They found that a 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. Half the questions on a bar graph test were either too easy or too hard. To us, this argues for building a graph comprehension test that controls for general knowledge. 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.

The Visualization Literacy Assessment Test [13] was developed according to the established procedure of test creation in psychological and educational measurement. The authors developed several types of graphs and maps and a series of three to seven questions for each graph. Of 61 questions developed, only 54 were deemed by a panel of five experts to measure the ability to read and interpret visually represented data. One further item was dropped due to low discriminability found after administration of the test to 191 volunteers. While VLAT is likely to be a useful tool, we note that the authors reported taking a month to develop these 61 test items from twelve source graphs, which were only then given to the expert panel for review and subsequently tested with volunteers. Our examination of their test materials leads us to believe that some questions also may have been answered from general knowledge the SVT framework mitigates this challenge through a four-fold structure for query probes (Sect. 3.2).

We believe these contributions and results with them show several challenges for writing tests of graph comprehension. It requires many queries to adequately test many aspects of graph comprehension, emphasizing the need for a better way to generate test questions. Thus, the process becomes quite labor-intensive. Even experts, writing subjective questions, may not realize the difficulty of a query and it may have to be removed from the test. 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 considerable with tools then in existence. We noted this difficulty in developing comprehension questions related to a node-link diagram using subjective development techniques [17]. 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 paraphrase, meaning change, and distractor query probes that are central to the SVT.

3.1 Graph Specification

Our clients make information dashboards for their customers. They use, and thus we adopted, HighCharts <<http://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 given in JavaScript Object Notation (JSON). This forms a hierarchical set of keys and values (Table 1), which lends itself to our need to manipulate graphical elements (Fig. 1) systematically.

Table 1. A JavaScript Object Notation (JSON) specification for a graph in HighCharts. See Fig. 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, },
}

```

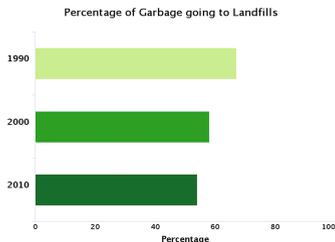


Fig. 1. The bar graph corresponding to the specification in Table 1.

3.2 Graph Query Definitions and Rules

Kosslyn [18] decomposed graphs into five components, of which three are important for our work (our *background* component thus far is a solid white field, and we do not use *captions*). The *framework* for most statistical graphs consists of the axes. The *content* is the representation of the data: points, lines, or bars. *Labels* name the variables, give titles to the graph or axes, or create a legend.

The SVT defines transformations of prose sentences into four types of query probes. Readers are asked to identify whether a probe gives information that was “stated” or “not stated” in the source prose. 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 the meaningful informational statements or assertions that are coordinated, collectively, by the graph’s components. A lone bar, divorced from a graph, is not an informational statement, but it becomes one 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 need not 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 opt to use one data point or multiple data points, to reflect the various information statements that are shown in a graph.

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. We now use two source graphs (Fig. 2) for examples of applying (some) rules for transformations from source to query graphs.

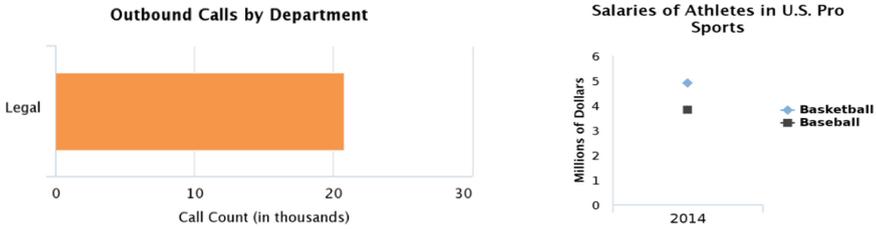


Fig. 2. An example bar graph and line graph used in the tutorial instructions for the study and used here to demonstrate the query variations as we adapted them from sentences to graphs.

3.3 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 style features in a graph that do not alter the meaning of the underlying data are not fundamental to the graph. *Content* may have different colors, fill, shapes, et al. *Labels* may be drawn in different font family, size, or style and be centered differently. 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* queries.

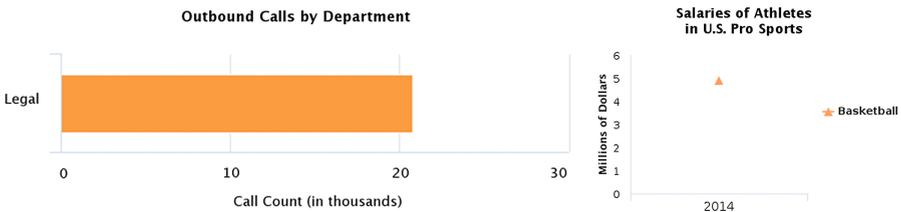


Fig. 3. *Original* query probes for the source information graphs shown in Fig. 2.

3.4 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 (Fig. 4); we argued above that these changes would not change the meaning, so they fit both definitions. Thus, style changes to *content* are the same as for original queries. We also deem rounding 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, this 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

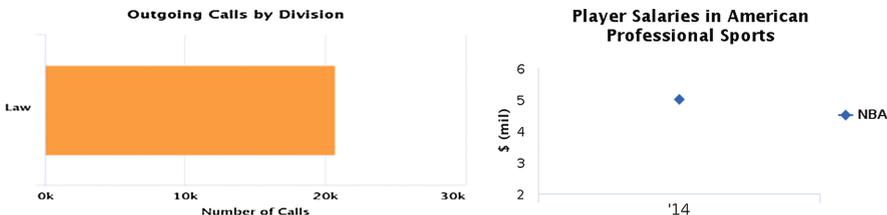


Fig. 4. *Paraphrase* query probes for the source information graphs shown in Fig. 2.

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 for *original* query types, we choose not to transpose axes, change the range of an axis or convert to logarithmic. 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 such framework changes could be permitted.

3.5 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 one constituent in a way that alters the meaning, and that

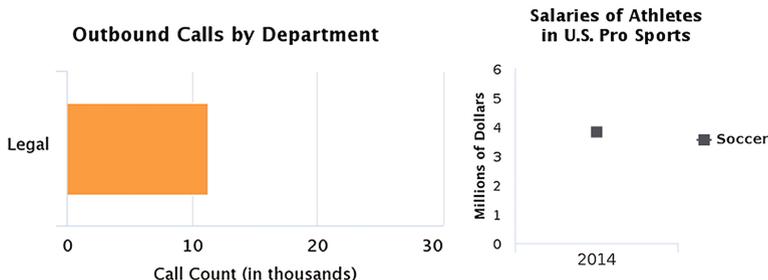


Fig. 5. *Meaning change* query probes for the source information graphs shown in Fig. 2. Note that the datum on the graph at right corresponds to the bottom point in the source graph, rather than the top point, which was used in the original and paraphrase query probes.

no further changes are permitted. But style changes to these three components are still permitted. Noticeable change to a datum (*content*) is perhaps the most obvious approach (Fig. 5, left), though changes to *labels* (Fig. 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. These changes cannot include the introduction of unrelated categories or series of data, since the introduction of new material belongs to the *distractor* query type.

3.6 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 (Fig. 6). However, we must limit ourselves to changes that stay within the topic of the source graph.

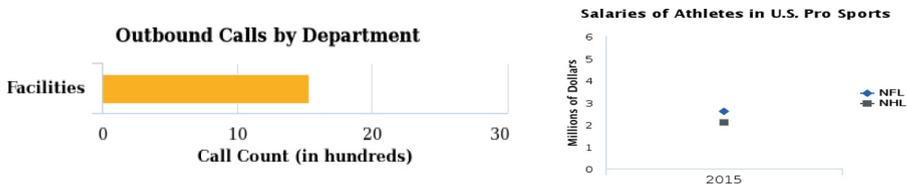


Fig. 6. *Distractor* query probes for the source information graphs shown in Fig. 2.

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 undertake this test. With recently available tests, such as VLAT [13], perhaps future work can test the consistency of tests developed under different paradigms.

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 [19]. 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 (Sect. 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, 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 [20]; they showed a variety of natural and urban imagery, with a few close-up images of common items. Participants also read brief, successive excerpts (about 200 words) from a 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 s, maximum time: 3 min),
- show a diversion image (3 s),
- show a blank screen (1 s),
- show a second diversion image (3 s),
- show a blank screen (1 s),
- show a source prose excerpt (also 30 s 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.

All material and layout from the study may be requested from the contact author. Participants completed a pre-study questionnaire with demographic and background information. They next 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 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.

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 1920 × 1200@60 Hz.

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. Thirteen said that they create such graphs for work or personal projects. Our participants came from the research and clerical staff at our laboratory; fourteen held a graduate degree. For the procedure as described above, participants took an average of 54 min (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, $F(3,69) = 1.442$, $p = 0.238$, $\eta^2 = 0.048$.

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.)

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 s. (c) The number of data points on a query showed a tendency to yield more accuracy with more data points. RT = response time

(a)		Graph Queries		Prose Queries	
Query Type	RT (sec)	Std. Dev.	RT (sec)	Std. Dev.	
<i>Original</i>	15.7	10.6	10.3	5.3	
<i>Paraphrase</i>	18.4	11.3	13.8	9.5	
<i>Meaning Change</i>	14.5	8.9	10.1	6.7	
<i>Distractor</i>	11.7	7.2	10.4	7.5	
(b) Number of Source Data	Study Time (sec)	Std. Dev	(c) Number of Query Data	Error (pct)	Std. Dev.
Three	36.3	0.7	One	0.125	0.334
Four	40.7	1.3	Two	0.089	0.286
Five	39.9	1.6	Three	0.028	0.107
Six	45.4	1.7			

5 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 with graph specification languages 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 (Sect. 3) for a comprehension query. The specification language enables us to transform a text language rather than a graph image. We believe that the combination of the taxonomy and the SVT structure also 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 (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 [21] 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 [22] and is an area for further study.

We seek ultimately 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 [17], 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, and the pilot study volunteers. This research was supported by the Naval Research Laboratory Base Program.

References

1. Roth, W.-M.: Reading graphs: contributions to an integrative concept of literacy. *J. Curric. Stud.* **34**(1), 1–24 (2002)
2. Galesic, M., Garcia-Retamero, R.: Graph literacy: a cross-cultural comparison. *Med. Decis. Mak.* **31**(3), 444–457 (2011)
3. Börner, K., Maltese, A., Balliet, R.N., Heimlich, J.: Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Inf. Vis.* **15**(3), 193–213 (2016)
4. Kintsch, W.: *Comprehension: A Paradigm for Cognition*. Cambridge University Press (1998)
5. Royer, J.M., Cunningham, D.J.: On the theory and measurement of reading comprehension. Technical report no. 91, University of Illinois at Urbana-Champaign (1978)
6. Wainer, H.: A test of graphicacy in children. *Appl. Psychol. Meas.* **4**(3), 331–340 (1980)
7. Royer, J.M., Hastings, C.N., Hook, C.: A sentence verification technique for measuring reading comprehension. *J. Read. Behav.* **11**(4), 355–363 (1979)
8. Bertin, J.: *Sémiologie Graphique*, 2nd edn., Gauthier-Villars (1973). English translation: Berg, W.J.: *Semiology of Graphics*. University of Wisconsin Press (1983)
9. McKenzie, D.L., Padilla, M.J.: The construction and validation of the Test of Graphing in Science (TOGS). *J. Res. Sci. Teach.* **23**(7), 571–579 (1986)
10. Curcio, F.R.: Comprehension of mathematical relationships expressed in graphs. *J. Res. Math. Educ.* **18**(5), 382–393 (1987)
11. Svec, M.: Improving graphing interpretation skills and understanding of motion using microcomputer based laboratories. *Electron. J. Sci. Educ.* **3**(4) (1999)
12. Lai, K., Cabrera, J., Vitale, J.M., Madhok, J., Thinker, R., Linn, M.C.: Measuring graph comprehension, critique, and construction in science. *J. Sci. Educ. Technol.* **25**(4), 665–681 (2016)
13. Lee, S., Kim, S.-H., Kwon, B.C.: VLAT: development of a visualization literacy assessment test. *IEEE Trans. Vis. Comput. Graph.* **23**(1), 551–560 (2017)
14. Van Dijk, T.A., Kintsch, W.: *Strategies of Discourse Comprehension*. Academic Press (1983)
15. Boy, J., Rensink, R.A., Bertini, E., Fekete, J.-D.: A principled way of assessing visualization literacy. *IEEE Trans. Vis. Comput. Graph.* **20**(12), 1963–1972 (2014)
16. Baker, F.B.: *The Basics of Item Response Theory*, 2nd ed. ERIC Clearinghouse on Assessment and Evaluation (2001)
17. Livingston, M.A., Brock, D., Maney, T., Perzanowski, D.: Extending the sentence verification technique to tables and node-link diagrams. In: *Proceedings of Applied Human Factors and Ergonomics* (2018)
18. Kosslyn, S.M.: *Graph Design for the Eye and Mind*. Oxford University Press (2006)
19. Shah, P., Freedman, E.G.: Bar and line graph comprehension: an interaction of top-down and bottom-up processes. *Top. Cogn. Sci.* **3**(3), 560–578 (2011)
20. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: *IEEE International Conference on Computer Vision*, pp. 2106–2113 (2009)
21. Harrison, A., Livingston, M.A., Brock, D., Decker, J., Perzanowski, D., Van Dolson, C., Mathews, J. Lulushi, A., Raglin, A.: The analysis and prediction of eye gaze when viewing statistical graphs. In: *Proceedings of Augmented Cognition. Neurocognition and Machine Learning*. LNCS, vol. 10284, pp. 148–165. Springer (2017)
22. Matzen, L.E., Haass, M.J., Divis, K.M., Stites, M.C.: Patterns of attention: how data visualizations are read. In: *Proceedings of Augmented Cognition. Neurocognition and Machine Learning*. LNCS, vol. 10284, pp. 176–191. Springer (2017)