VLAT: Development of a Visualization Literacy Assessment Test

Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon, *Member, IEEE*

Abstract— The Information Visualization community has begun to pay attention to visualization literacy; however, researchers still lack instruments for measuring the visualization literacy of users. In order to address this gap, we systematically developed a visualization literacy assessment test (VLAT), especially for non-expert users in data visualization, by following the established procedure of test development in Psychological and Educational Measurement: (1) Test Blueprint Construction, (2) Test Item Generation, (3) Content Validity Evaluation, (4) Test Tryout and Item Analysis, (5) Test Item Selection, and (6) Reliability Evaluation. The VLAT consists of 12 data visualizations and 53 multiple-choice test items that cover eight data visualization tasks. The test items in the VLAT were evaluated with respect to their essentialness by five domain experts in Information Visualization and Visual Analytics (average content validity ratio = 0.66). The VLAT was also tried out on a sample of 191 test takers and showed high reliability (reliability coefficient omega = 0.76). In addition, we demonstrated the relationship between users' visualization literacy and aptitude for learning an unfamiliar visualization and showed that they had a fairly high positive relationship (correlation coefficient = 0.64). Finally, we discuss evidence for the validity of the VLAT and potential research areas that are related to the instrument.

Index Terms—Visualization Literacy; Assessment Test; Instrument; Measurement; Aptitude; Education

1 Introduction

Data visualizations have become more popular and important as the amount of available data increases, called data democratization. People want to see and explore data, extract useful information from the data, grasp the meaning of the information, and visually represent findings. For that reason, a lot of data visualization dashboards, applications, and software are developed and launched to satisfy these needs. Moreover, people encounter numerous data visualizations in everyday life, especially through web browsers. The government, institutions, and news outlets actively apply various data visualization techniques to deliver information and stories. The visualizations allow them to represent complex underlying data concisely, and people also believe that the visualizations are persuasive and attractive interfaces [40]. In these circumstances, an individual's ability to read, comprehend, and interpret data visualizations can strongly influence his/her tasks and communication. In other words, *visualization literacy* is becoming as important as the ability to read and comprehend text [8, 25, 34].

In recent years, researchers in the Information Visualization community have endeavored to investigate users' visualization literacy. Throughout a few studies, the researchers proposed a method to assess visualization literacy [10]; tried to understand the current status of data visualization comprehension of the general public [8]; and suggested various ways of learning unfamiliar visualizations to improve users' visualization literacy [31, 42]. Some researchers qualitatively investigated users' cognitive activities when they made the effort to understand data visualizations [33]. The community has also recognized the importance of this topic and revealed it through two consecutive workshops: *EuroVis 2014 Workshop: Towards Visualization Literacy* followed by *IEEE VIS 2014 Workshop: Towards an Open Visualization Literacy Testing Platform*. A panel discussion in IEEE VIS 2015, *Vis, The Next Generation: Teaching Across the Researcher-Practitioner Gap*, also showed urgent needs to address this topic. However, the researchers still lack a validated and reliable instrument for measuring visualization literacy of users that can be widely used [54]. Addressing this gap is important because measurement results from the instrument can help researchers and designers make better decisions in the process of designing and developing visualization applications and software. Furthermore, accurate measures of visualization literacy can help them to educate potential users in a systematic way.

Thus, the goal of this study is to develop a test for measuring the visualization literacy of users, especially non-expert users in data visualization. In order to systematically develop the test, we went through the established procedure of test development in Psychological and Educational Measurement [15, 47]. While we rigorously followed the procedure, we designed the test that is applicable to the Information Visualization community. In order to choose the contents of the test (i.e., data visualizations), we surveyed three different sources: K-12 curriculums, data visualization authoring tools, and news articles, by considering the target test taker population. We also reviewed task taxonomies in Information Visualization [4, 11, 14] and a classification of typical dataset types of visualizations [37], and generated test items based on them. We also invited five domain experts in Information Visualization and Visual Analytics to evaluate the essentialness of each test item. Then, the test was tried out on a sample of the potential test takers and the result of the tryout was analyzed in terms of difficulty and discrimination. Finally, we collected evidence for the validity of the test in the process of the development and showed the relationship between visualization literacy test scores and aptitude for learning an unfamiliar visualization. Our endeavor toward the research goal was based on a psychological assumption that human traits and skills can be quantified and measured [15].

The main contributions of this study are as follows:

- We show the methodical procedure to construct a visualization literacy assessment test.
- We present a systematically developed visualization literacy assessment test that consists of 12 data visualizations and 53 multiple-choice test items.
- We provide the test content-related evidence and the test reliability-related evidence for the validity of the test.
- We demonstrate the relationship between user's visualization literacy and aptitude for learning an unfamiliar visualization.

The primary idea of this study has originated from a paper [54] and *IEEE VIS 2014 Workshop: Towards an Open Visualization Literacy Testing Platform*, which was co-organized by two of the authors with knowledgeable colleagues. In particular, we had active discussions about how to test ordinary users' visualization literacy and how to generate test items in the workshop with the panelists and participants. The direction of this paper was inspired by the discussions.

- *Sukwon Lee is with the School of Industrial Engineering at Purdue University in West Lafayette, IN, USA. E-mail: sukwon@purdue.edu*
- *Sung-Hee Kim, the corresponding author, is with Samsung Electronics Co., Ltd. in Seoul, South Korea. E-mail: sungheekim02@gmail.com*
- *Bum Chul Kwon is with IBM T.J. Watson Research Center in Yorktown Heights, NY, USA. E-mail: bumchul.kwon@us.ibm.com*

Manuscript received 31 Mar. 2016; accepted 1 Aug. 2016. Date of publication 15 Aug. 2016; date of current version 23 Oct. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2016.2598920

2 Background

2.1 Visualization Literacy

The topic of visualization literacy is gaining recognition within the Information Visualization community, and a number of researchers have begun to explore this topic [8, 10, 33]. In this section, we summarize the recent studies and review the definition of visualization literacy.

2.1.1 Recent Work on Visualization Literacy

Boy et al. [10] conducted one of the pioneering studies on visualization literacy in the community. They defined the term of visualization literacy (see Section 2.1.2) and proposed a method to assess an individual's level of visualization literacy with a set of test items. The proposed method was based on item response theory (IRT), which is a well-known modern psychometric test theory, especially for a test item analysis [15, 18, 47]. The core ideas of IRT are (1) to mathematically calculate the probability that a person will answer a test item correctly based on the person's ability and the item's difficulty, and (2) to mathematically estimate the person's ability [18]. They conducted the item response modeling with two line charts and applied the model to a bar chart and a scatterplot. The considered visualization tasks for the modeling were finding maximum, finding minimum, finding variation, finding intersection, calculating average, and comparing. Even though they provided an inspiring idea for analyzing visualization literacy test items and showed the applicability of IRT, there were potential areas that could be expanded to construct a comprehensive visualization literacy assessment test: containing various data visualizations, covering inclusive data visualization tasks, and following the whole procedure of test development.

In the following year, Börner et al. [8] tried to understand the general public's current level of visualization literacy in an indirect way. They showed 20 different printed data visualizations to the participants and asked five questions, for example, "Does this type of data presentation look at all familiar?" "Where might you have seen images like this?" and "How do you think you read this type of data presentation?" The questions were more related to familiarity with the visualizations. From the participants' responses, they found the current status of the general public's visualization literacy. For instance, the participants considered colors, lines, and text as important features of data visualizations to understand them. In addition, those participants who had low visualization literacy had no ability to read network visualizations.

More interestingly, some researchers carefully described how non-expert users made sense of data visualizations and what their cognitive activities were [33]. This type of exploratory study would be a preceding work for investigating visualization literacy; however, they did not directly tackle the issues of how to measure users' visualization literacy and how to evaluate whether users can correctly read and interpret data visualizations or not.

2.1.2 Definition of Visualization Literacy

Since the primary goal of this study is to develop an assessment test of a skill, visualization literacy in this paper, arriving at a clear definition of the skill is important. Because tasks that reflect the skill vary depending on the definition, and test items that would be included in the test are generated according to the tasks [47].

According to the Merriam-Webster dictionary, the term of literacy is simply defined as "the ability to read and write" [1]. In the dictionary definition, the term of literacy has two different aspects of ability: consumption aspect and production aspect. However, literacy is used as the ability to understand and use something by emphasizing the consumption aspect as the term is combined with other subjects (e.g., information literacy, health literacy, and energy literacy). Furthermore, Educational Testing Service (ETS), which is one of the biggest commercial test companies in the world, defines literacy as "how well adults can use printed and written information to function in society, to achieve their goals, and to develop their knowledge and potential" [2] and also emphasizes the consumption nature of literacy.

Boy et al. [10] and Börner et al. [8] defined visualization literacy from this standpoint. However, they have reached no agreement on the definition. The definition by Boy et al. is "the ability to use well established data visualization (e.g., line graphs) to handle information in an effective, efficient, and confident manner" [10] (p. 1963); while the definition by Börner et al. is " the ability to make meaning from and interpret patterns, trends, and correlations in visual representations of data" [8] (p. 3). Even though we acknowledge their initial thoughtful effort to define visualization literacy, we still find areas for improvement in the definitions that are: (1) ambiguity in the use of words, for example, well established, effective, efficient, and confident manner, and (2) narrowly defined task types, such as interpreting patterns, trends, and correlations.

Thus, we devise the definition of visualization literacy that is referred in this study. We tried to refine and embrace their definitions in order to clarify the target skill to be measured by a test and expand task types in data visualization to be covered by the test.

> *Visualization literacy is the ability and skill to read and interpret visually represented data in and to extract information from data visualizations.*

2.1.3 Measuring Graph Comprehension

Several researchers conducted fundamental studies in the field of graph comprehension. They defined graph comprehension as reading and interpreting a graph [24], and they proposed the three levels of graph comprehension framework [7, 13, 49]: (1) the first level is the elementary level in which a graph reader can read a specific value in a graph (e.g., a graph reader is able to read the height of a bar in a bar chart); (2) the second level is the intermediate level in which a graph reader can read relationships or trends in a graph (e.g., a graph reader is able to compare the heights of two bars in a bar chart or to discern the slope between two point in a line chart); and (3) the third level is the advanced level in which a graph reader can read beyond what is presented in a graph (e.g., a graph reader is able to predict a future trend in a line chart). In particular, Curcio [16] called the three levels *read the data*, *read between the data*, and *read beyond the data*.

Based on the three-level framework, researchers generated question items about primitive graphs in order to understand graph readers' comprehension level [16, 17, 25, 49]. Most of the items were multiple-choice items and short-answer items (i.e., fill-in the blank). They reached common findings: graph readers made more errors with the second-level items than with the first-level items, and graph readers had difficulty in answering the third-level items. However, diverse data visualization tasks were not covered by the questions that were based on the three-level framework. On the contrary, Shah and Freedman [45] tried to understand the graph comprehension level using an open-ended item (i.e., "What is the most important information in the graph?"). Open-ended items provided more detailed descriptions about individual's graph comprehension and they might be useful to examine individual's insights from data visualizations [51, 52]. However, this type of items was difficult to evaluate reliably because it required judgmental evaluations and qualitative analyses.

2.2 Procedure of Test Development

As mentioned before, we followed the established procedure in Psychological and Educational Measurement to develop a visualization literacy assessment test. In this section, we briefly describe the general procedure and some points to be considered by test developers.

All measurement tests are not created equally. However, a good test is the product of established principles and the procedure of test construction [5, 15]. First of all, the test construction begins with having a clear definition of a human trait or skill, visualization literacy in this paper, to be measured by a test because major contents in a test and cognitive tasks within the contents vary depending on the definition. Once test developers have the definition, they follow the procedure of test development that generally occurs in six phases: (1) Test Blueprint Construction, (2) Test Item Generation, (3) Content Validity Evaluation, (4) Test Tryout and Item Analysis, (5) Test Item Selection, and (6) Reliability Evaluation.

Test Blueprint Construction A test blueprint (also called a test specification table) is a table that identifies two main components of a test: (1) major contents to be covered by the test and (2) associated cognitive tasks within the contents. Constructing a test blueprint is crucial because the test blueprint is an explicit plan that guides the subsequent test development processes, and it could be used as one of materials for evaluating the quality of a test in terms of content validity [47]. Thus, the test developers should consider the characteristics of a potential test taker population when they decide the two main components.

Test Item Generation Based on the constructed test blueprint, the test developers generate test items that would compose the test. Items should be written clearly to express the tasks. In this phase, the test developers encounter the following questions and should answer the questions while writing test items: (1) What types of items would be employed, selected-response items or constructed-response items? and (2) How many items would be generated? In particular, the test developers need to consider test takers' level of vocabularies and phrases related to tasks to avoid poor performance due to lexical issues.

Content Validity Evaluation After having a set of test items, it should be reviewed by multiple independent domain experts to ensure the test contains appropriate contents and requires appropriate tasks within the contents. One widely used method to evaluate content validity is a quantitative approach by calculating the content validity ratio (*CVR*), which was devised by Lawshe [32]. *CVR* ranges from -1.0 to 1.0 and it indicates the experts' agreement on how a particular item is essential to measure the trait or skill. The result of this evaluation would be the first evidence for validity that test scores have the meaning that is intended when the test was developed [5, 36].

Test Tryout, Item Analysis, and Item Selection The reviewed items are tried out on a group of sample test takers. Even though the sample test takers should not necessarily be representative of the target population, the sample should include individuals who have similar average abilities as the target population. The collected answers to the items are analyzed in order to get empirical evidence of the item quality, referred to as item analysis [15, 47]. Based on the results of the item analysis, inappropriate items are discarded or modified. The qualified remaining items are included in the final set of test items.

One of the widely employed approaches to conduct the item analysis is classical test theory (CTT) analysis. In the analysis, the test developers should answer the following two questions for each item [47]: (1) How difficult is the item? and (2) Does the item distinguish between the upper and lower scoring groups? In order to answer these questions, straightforward tools are used, for example, mean, item difficulty index, and item discrimination index. An alternative approach to CTT analysis is item response theory (IRT) analysis (see the Boy et al.'s paper [10] for the application of IRT in visualization literacy). A merit of this approach is to obtain the probability that a test taker will make a correct response to a certain item depending on his/her trait or skill level through the IRT models. However, the weakness of IRT analysis is that large samples of test takers are required to accurately estimate the parameters in the IRT models, for example, the recommended minimum sample sizes are 100 for *b*-parameter, 500 for *a*-parameter, and 2,000 for *c*-parameter [18]. A surprising fact is that the results derived from CTT analysis and IRT analysis are highly comparable in the practical situation of test development [19, 22]. In this study, we adopted the CTT analysis approach.

Reliability Evaluation Lastly, the test containing the final set of items is evaluated in terms of reliability. The reliability is the property of observed test scores and the attribute of consistency in a test. There are several methods to estimate reliability of a test: test-retest reliability, parallel test forms reliability, single administration reliability, and Cronbach's coefficient alpha. However, these traditional methods have an assumption, unbiased estimation of items, that are rarely met with real data [44, 46]. Thus, an alternative method, reliability coefficient omega [35], is often used in a practical situation [21, 35]. The results of the reliability evaluation serve as another evidence for validity that visualization literacy is measured precisely and consistently [5].

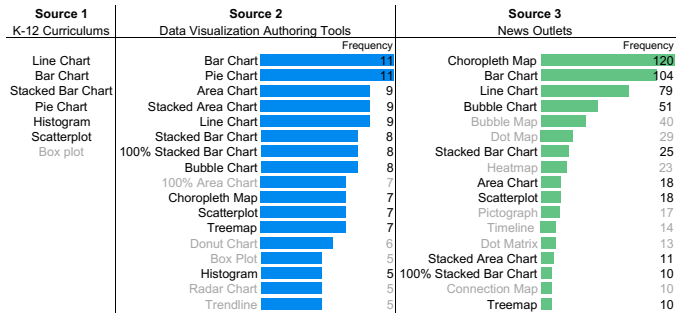


Fig. 1. Surveyed data visualization types from three sources: K-12 curriculums, data visualization authoring tools, and news outlets. In particular, the second and third columns show the most frequently occurring data visualization types from the tools and the news outlets respectively.

3 Development of the VLAT

By following the procedure of test development (see Section 2.2), we systematically developed a visualization literacy assessment test. We will refer to this test as the VLAT. In this section, we describe the six phases that we went through to construct and evaluate the VLAT.

3.1 Phase I: Test Blueprint Construction

As we described before, the two main components of a test blueprint were major contents to be covered by a test and associated cognitive tasks within the contents. In case of the VLAT, they were data visualizations and associated data visualization tasks respectively.

3.1.1 Data Visualizations

In order to cover the appropriate major contents to be contained in the VLAT, we surveyed data visualization types while taking the characteristics of the potential test taker population into account. We expected the potential test takers of the VLAT to be: (1) non-expert users in data visualization over 18 years old; (2) those who learned required skills to read and interpret data visualizations within K-12 curriculums (most of them were called graphs/charts/plots in the curriculums); and (3) those who created, used, and/or encountered various types of data visualizations in everyday life. Thus, we surveyed data visualization types from three sources: K-12 curriculums (Source 1), data visualization authoring tools (Source 2), and news outlets (Source 3).

First, we surveyed data visualization types that were covered by K-12 curriculums. We reviewed a number of core state standards for mathematics [12, 28, 38, 50] and found that people learned seven visualization types within the curriculums: Line Chart, Bar Chart, Stacked Bar Chart, Pie Chart, Histogram, Scatterplot, and Box Plot [24].

Second, we surveyed data visualization types that could be produced by visualization authoring tools. For the tools, we considered five developer tools that required computer programming skill, also known as programming tools [53], (i.e., Google Chart Tools, D3.js, Chart.js, JavaScript InfoVis Toolkit, and Dimple) and six non-developer tools that do not required computer programming skill, also known as out-of-the-box tools [53], (i.e., Tableau, Microsoft Excel, IBM Watson Analytics, Many Eyes, Plotly, and Datawrapper). In total, 65 different types of data visualizations could be created using the tools. This indirectly shows that people could create and/or encounter various types of data visualizations depending on their purpose.

Third, we surveyed data visualization types from news outlets. We collected a total of 494 news articles that included data visualizations from *The New York Times* (from 2003 to 2015), *The Guardian* (from 2008 to 2015), and *The Washington Post* (from 2014 to 2015), and then we surveyed the types in the 494 articles. In total, 44 different types of data visualizations were shown in the news articles.

After surveying all data visualization types from the three sources, we sorted out the visualizations based on the frequency. We picked the most frequently occurring visualization types out from Source 2 and Source 3, which were included in the upper quartile. Figure 1 summarizes the results (the frequency table of all data visualization types used in Sources 2 and 3 is included in the supplemental material). It shows the data visualization types covered by Source 1 and

Table 1. The test blueprint of the VLAT. It shows major contents (12 data visualization types) and associated cognitive tasks covered by the test. Please note that the empty spaces in the blueprint do not mean that users cannot do the tasks with the visualizations (for some cases, of course, users cannot perform the tasks). We mark the tasks that are more suitable to perform with the data visualizations and the typical dataset types.

Visualization	Dataset Type	Visualization Task									Note of X†
		Retrieve Value	Find Extremum	Determine Range	Characterize Distribution	Find Anomalies	Find Clusters	Find Correlations/ Trends	Make Comparisons	ETC	
Line Chart	Table: One quantitative value attribute, One ordered key attribute	X	X	X				X	X		
Bar Chart	Table: One quantitative value attribute, One categorical key attribute	X	X	X					X		
Stacked Bar Chart	Multidimensional Table: One quantitative value attribute, Two categorical key attributes	X†	X	X					X†		† Both Absolute Value and Relative Value
100% Stacked Bar Chart	Multidimensional Table: One quantitative value attribute, Two categorical key attributes	X†	X†						X†		† Only Relative Value
Pie Chart	Table: One quantitative attribute, One categorical attribute	X†	X†						X†		† Only Relative Value
Histogram	Table: One quantitative value attribute	X†	X†		X				X†	Identify the Characteristic of Bins	† Only Derived Value
Scatterplot	Table: Two quantitative value attributes	X	X	X	X	X	X	X	X		
Area Chart	Table: One quantitative value attribute, One ordered key attribute	X	X	X				X	X		
Stacked Area Chart	Multidimensional Table: One quantitative value attribute, One categorical key attribute, One ordered key attribute	X†	X	X				X	X†		† Both Absolute Value and Relative Value
Bubble Chart	Multidimensional Table: Three quantitative value attributes	X	X	X	X	X	X	X	X		
Choropleth Map	Geographic geometry data and Table: One quantitative attribute per region	X†	X†						X†		† Only Approximate Value
Treemap	Tree and Table: One quantitative value attribute per node	X†	X†						X†	Identify the Hierarchical Structure of Dataset	† Only Relative Value

the most frequently occurring data visualization types from Source 2 and Source 3. Most of the visualizations covered Source 1 were both well supported by Source 2 and Source 3. Interestingly, Choropleth Map was the most frequently shown in Source 3 even though it was not covered by Source 1. On the contrary, Pie Chart and Histogram were included in Source 1 and well supported by Source 2 but they were not in the upper rank of the visualization types in Source 3.

Lastly, in order to decide the final set of data visualization types to be contained in the VLAT, we selected the visualization types that were listed under at least two sources in Figure 1 among the three sources. Thus, we had 12 data visualization types as the final set of contents of the test: Line Chart, Bar Chart, Stacked Bar Chart, 100% Stacked Bar Chart, Pie Chart, Histogram, Scatterplot, Bubble Chart, Area Chart, Stacked Area Chart, Choropleth Map, and Treemap[1]. In Table 1, they compose the rows of the test blueprint of the VLAT.

3.1.2 Tasks

For the second main component of the test blueprint, we determined associated tasks of each of the 12 data visualizations. In order to do so, we considered (1) task taxonomies in Information Visualization [4, 11, 14] and (2) the associated dataset types of the 12 visualizations [37].

First, we identified possible tasks of the data visualizations based on task taxonomies in Information Visualization [4, 11, 14]. We merged the low-level taxonomy, which was based on users' analysis questions [4], and the fact taxonomy, which was based on a literature survey, user study, and domain expert review [14]. Then, we ruled out some tasks that were included in the *how* and *why: produce* categories in the tasks typology [11]. Because these tasks were more related to manipulating and generating new elements from a visualization (e.g., Compute Derived Value and Sort) [11] rather than reading and interpreting visually represented data (see Section 2.1.2). In the result, we had eight possible data visualization tasks: Retrieve Value, Find Extremum, Determine Range, Characterize Distribution, Find Anomalies, Find Clusters, Find Correlations/Trends, and Make Comparisons. In Table 1, they compose the columns of the test blueprint.

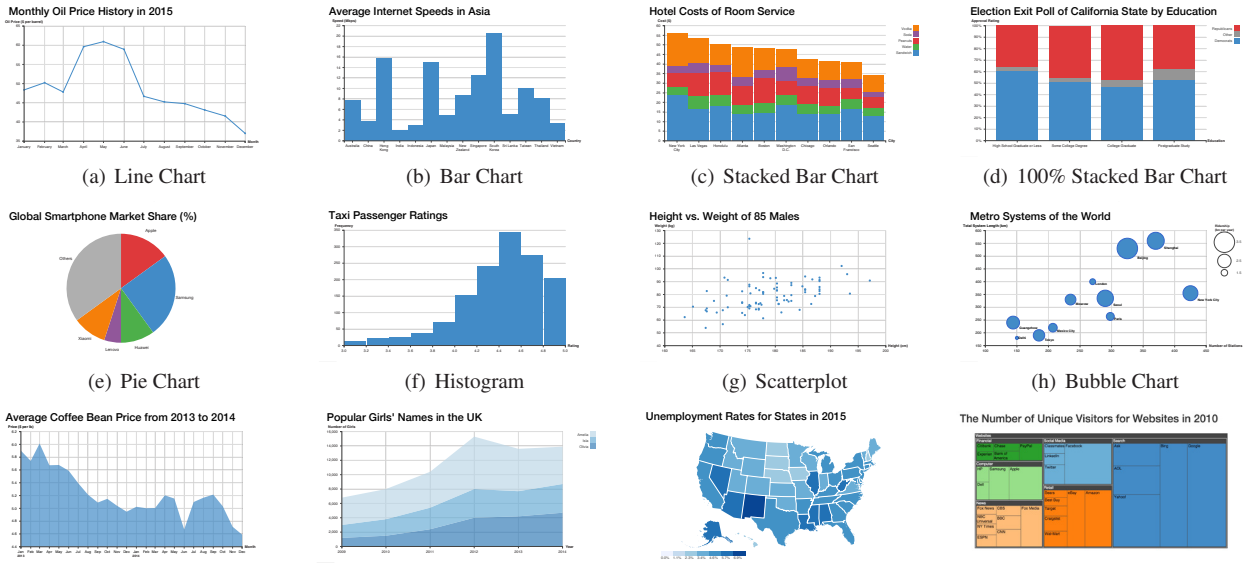
Then, we identified the typical dataset type of each visualization according to the classification by Munzner [37] because the associated tasks of a data visualization were decided not only according to the visual encoding schemes of the visualization but also according to the underlying dataset type. The typical dataset type of each visualization is also shown in Table 1. Among the four basic dataset types (i.e., table type, network type, field type, and geometry type) [37], most of the visualizations represented table types. Choropleth Map and Treemap represented a geometry type and a network type with a table type respectively. These dataset types informed us about the typical dataset structure of each visualization, furthermore, they guided us to the solid determination of associated tasks for each visualization.

Finally, we determined associated tasks of the 12 data visualizations based on the task taxonomies and the dataset types. The result is shown in the final test blueprint (Table 1). For each visualization, performable cognitive tasks are marked with an $\times$ at the intersections. In most cases, when a user performs the tasks with a visualization, the user utilizes visual objects that directly represent the absolute values of the underlying dataset. Examples include Line Chart, Bar Chart, Scatterplot, and Bubble Chart. However, there are some exceptions. First, depending on the visualization, a user sometimes utilizes visual objects that only represent the transformed values of absolute values to perform the tasks. For instance, only the proportion/percentage of each value to the total of a group is represented on 100% Stacked Bar Chart, Pie Chart, and Treemap; and only the approximately segmented range of each value is represented on Choropleth Map. Second, interestingly, even though the absolute values of a dataset are represented on Stacked Bar Chart and Stacked Area Chart, a user is allowed to perform some tasks with the proportion/percentage of each value to the total of a group because of the structural feature of the visualizations. Third, with Histogram, a user utilizes visual objects that represent derived values (i.e., the frequency/count of each value) that vary according to the choice of bin size by the visualization designer. These cases are marked with an $\times^{\dagger}$. Note that we included an additional task for both Histogram and Treemap since we believed that identifying the characteristic of bins and identifying the hierarchical structure of the dataset were critical underlying tasks to read and interpret visually represented data from the visualizations respectively [29, 37, 53].

3.1.3 Datasets, Contexts, and Interaction Techniques

The final step of Phase I was to create 12 specific data visualizations that would be used in the VLAT. To create specific visualizations, we selected datasets. The datasets should be close to real datasets instead of synthetic datasets. Thus, we collected datasets that were actually used in the news articles from Source 3. We especially considered the dataset types in the test blueprint. Furthermore, when we decided the final datasets, we carefully considered dataset contexts because the familiarity of the contexts was one of the main factors influencing visualization comprehension [16, 33, 45]. Since the primary purpose of the

[1]Box Plot was selected as a candidate visualization for the VLAT but we decided to exclude Box Plot because it requires specific statistical knowledge (e.g., percentile, quartile, interquartile range) to understand the visualization.



(a) Line Chart (b) Bar Chart (c) Stacked Bar Chart (d) 100% Stacked Bar Chart

(e) Pie Chart (f) Histogram (g) Scatterplot (h) Bubble Chart

(i) Area Chart (j) Stacked Area Chart (k) Choropleth Map (l) Treemap

Fig. 2. The 12 data visualizations that compose the VLAT. Note that Choropleth Map includes the abbreviations of the state names in the instrument.

VLAT was to assess an individual's visualization literacy, we avoided the potential bias of the familiarity of the contexts. In particular, we reviewed not only the contexts of datasets but also their attributes because people might be familiar with a context but not its attributes. For example, one might be familiar with the context of "car" but not the specific attributes of "displacement" or "0-60 mph." Eventually, we decided on 12 datasets with general contexts that did not require specific expertise (e.g., monthly oil price, height vs. weight, popular girls' names, and unemployment rates).

We did not include any interaction techniques in the visualizations. This was because people had a limited ability to detect interaction techniques embedded in a visualization initially [9], and interaction techniques may have invited potential test takers to do additional complex tasks. Furthermore, it would become more like reading text rather than reading and interpreting visually represented data if we provide *Elaborate* interaction techniques [55]. Instead, we included grids in the visualizations that had a Cartesian coordinate system in order to help the potential test takers read values on axes. Finally, we came up with the final set of 12 data visualizations of the VLAT in Figure 2.

3.2 Phase II: Test Item Generation

The second phase was to generate test items that would be contained in the VLAT based on the blueprint. Before generating the test items, we conducted a pilot work to understand the usage of vocabularies and phrases when people read and interpreted the data visualizations. Understanding and ensuring the lexical level of the potential test taker population are important because vocabularies and phrases that compose test items may affect test takers' performance [5].

3.2.1 Pilot Work

A total of 69 participants were initially recruited through Amazon Mechanical Turk (MTurk). We recruited crowdsourced workers who had a total of 1,000 or more approved HITs and a 95% or greater HIT approval rate only from the United States. The workers well represent the US population [6, 27]. In particular, we intentionally recruited workers who were native English speakers because using a non-native language might have been an obstacle to thought processes and lead to poor performance [41, 51]. So, we ruled out one participant who self-reported that he/she was not a native English speaker. We also ruled out four participants who self-reported that they were color blind since the 12 data visualizations did not use color-blind safe colors. As a result, a total of 64 participants remained. The remaining participants were 35 females and 29 males with the self-reported age range of 21 to 65 ($M = 37$). Everybody had an education level of high school or above, 42% of the participants had a bachelor's degree, and 19% of the participants had a master's or a doctoral degree.

We released a survey to the participants. In the survey, we randomly provided the 12 visualizations and the associated tasks with the descriptions of the tasks based on the test blueprint. Each time we explained a task, we asked the participants the following question: "While doing the task, `[taskname]`, you may gain a piece of information from the data visualization. Please state the information in a sentence using your own words." The participants were asked to type the information they grasped using their own words.

From the survey, we collected a total of 3,352 statements. The collected statements, especially vocabularies and phrases, were reviewed by the test developers (i.e., all of the authors) and they were used as reference materials when we wrote test items in order to keep the appropriate vocabulary level of the items (see Section 3.2.2). Note that we did not use the participants' statements for any other purposes.

3.2.2 Writing Items

After reviewing the statements collected in the pilot work, we wrote test items based on the test blueprint (Table 1). We intentionally wrote selected-response items (e.g., multiple-choice items and true-false items) rather than constructed-response items (e.g., short answer items and open-ended items) to generate test items that can be objective and efficient in terms of test constructing, taking, and scoring [47]. Some researchers adopted open-ended items to examine extracted insights and generated explanations from people [51, 52]; however, our primary purpose was to measure the test takers' ability and skill.

We wrote one test item for each of the associated tasks. However, we wrote two test items in case a certain task could be performed with both absolute values and relative values with Stacked Bar Chart and Stacked Area Chart. Once we wrote all test items, we reviewed the items several times to ensure each item clearly reflected the associated task. In addition, we reviewed the collected statements from the pilot work again and modified the items to ensure the lexical level.

As a consequence, we had a total of 61 potential test items with 39 four-option multiple-choice items, 3 three-option multiple-choice items, and 19 true-false items. Each test item consisted of a stem, which presented a problem, two to four options, and only one correct or best answer. The form of the stem was either a question or an incomplete statement for the multiple-choice items and a complete statement for the true-false items.

3.3 Phase III: Content Validity Evaluation

In the next phase, we conducted a content validity evaluation with five domain experts, data visualization researchers in various locations, to

ensure the generated items in Phase II were appropriate. In order to evaluate the content validity, we calculated the content validity ratio (*CVR*) for each item according to the Lawshe's suggestion [32].

Participants We invited five domain experts in Information Visualization and Visual Analytics. Their average age was 36 years old. Everybody held a doctorate, and they had 7 to 15 years ($M = 10$) of professional experience in the domain. Three experts were in academia and two experts were in industry.

Procedure We took the following procedure to conduct the evaluation with the experts. First, we informed them about the purpose of this evaluation, the definition of visualization literacy (see Section 2.1.2), the selection processes for the 12 data visualizations and the associated tasks in order to provide the context. Then, we presented the generated 61 test items with the visualizations and the tasks one by one, and asked the following typical question to get *CVR* for each item: "Is the task measured by this item [`essential`, `useful but not essential`, or `not necessary`] to the visualization literacy?" [32]. Based on the collected responses from the experts, we counted the number of experts indicating "essential" and calculated *CVR* for each item using the Lawshe's *CVR* formula [32].

Content Validity Ratio and Filtering Out Items A positive value of *CVR* is interpreted as more than half of the experts rate this item as "essential" to what it is intended to measure. Thus, we retained 54 items with $CVR > 0$ and filtered out seven items with $CVR \leq 0$. Specifically, the items for Determine Range with Stacked Bar Chart (Item 13), Identify the Characteristic of Bins with Histogram (Item 26), Characteristic Distribution with Scatterplot (Item 30), Make Comparisons with Area Chart (Item 39), Determine Range with Stacked Area Chart (Item 43), Characteristic Distribution with Bubble Chart (Item 50), and Retrieve Value with Treemap (Item 58) were discarded. The retained 54 items were composed of 34 four-option multiple-choice items, 3 three-option multiple-choice items, and 17 true-false items. The *CVR*s of the 54 items are presented in Table 2. The average *CVR* of the 54 items was 0.65.

3.4 Phase IV: Test Tryout and Item Analysis

Once the test items were thoroughly reviewed by the domain experts, the retained 54 items were tried out on a sample of test takers. The results of the tryout were analyzed to examine the quality of the items.

3.4.1 Test Tryout

Participants A total of 200 participants were initially recruited through MTurk for the tryout. We applied the same requirements as the participants in Phase II (see Section 3.2.1). So, we ruled out one self-reported non-native English speaker and two self-reported color blind. Furthermore, in order to remove random clickers, we ruled out six participants who completed an item under five seconds for more than 12 items. We also confirmed that they were random clickers based on their responses to a filtering question, "In the test that you took, can you recall what the data visualizations were about (e.g., hotel, bicycle, or airfare)? Please type the context as your memory serves." Eventually, a total of 191 participants remained. They consisted of 105 females and 86 males with the self-reported age range of 19 to 72 ($M = 37$). Everybody had an education level of high school or above, 42% of the participants had a bachelor's degree, and 15% of the participants had a master's or a doctoral degree.

Procedure We administered the test that consisted of the retained 54 potential test items. First, we provided test instructions to the participants. In the instructions, the participants were informed about the purpose of the test and asked to select the best answer to each item within a time limit. Even though the test was not a speed test, we were not able to allow the participants unlimited time to take the test in order for the test to be a standardized assessment test. At the same time, we should provide enough time for the participants to attempt to answer the items. Thus, we allowed the participants a maximum of 25 seconds to answer to an item (approximately a maximum of 23 minutes in total to complete the test) based on our pilot study.

After the participants read the instructions, the test items were randomly presented one by one. In particular, an "Omit" option was provided to the participants for every item in order to address the issue of guessing in multiple-choice items. The participants were also clearly instructed to select the "Omit" option when their answer was based on a guess and that the score would be corrected for guessing [47]. This setting would influence the participants' test taking strategies and reduce the test error caused by their guessing, which was a weakness of multiple-choice items [20, 23].

Scoring In order to address the issue of guessing in multiple-choice items, we applied the correction-for-guessing to score [20, 23]. The raw score of each participant on the test was adjusted with the correction-for-guessing formula:

$$CS = R - \frac{W}{C-1} \tag{1}$$

where CS was the corrected score, R was the number of items answered correctly (i.e., the raw score), W was the number of items answered incorrectly, and C was the number of choices for an item [47].

3.4.2 Item Analysis

To better understand the performance of the test items of the test, we conducted an item analysis based on classical test theory (CTT): basic statistics, item difficulty index, and item discrimination index.

Basic Statistics We observed the raw scores of the test takers and the corrected scores, which were adjusted according to Equation (1). The total possible score points on the test were 54. The raw scores of the test takers ranged from 14 to 50 ($M = 34.72$, $SD = 7.05$). The corrected scores ranged from 3.22 to 49.34 ($M = 27.51$, $SD = 8.78$). After adjusting, the test takers' scores dropped an average of 7.21 points.

We also conducted the Shapiro-Wilk test in order to check the normality of the corrected scores. Because the normal distribution of scores on the test would facilitate the interpretation of individual scores. The result showed that the corrected scores on the test were normally distributed ($W = .99$, $p = .43$) and the distribution had own $M = 27.51$, $SD = 8.78$.

In addition, we observed the test completion time. The average test completion time of the test takers was 14 minutes 50 seconds ($SD = 3$ minutes 6 seconds), ranging from 8 minutes 31 seconds to 22 minutes 42 seconds. It indicated that the assigned time limit (25 seconds per item) was fair on the test takers to complete the whole test items.

Item Difficulty Index The item difficulty index is a portion of the test takers who answered the item correctly. The value of the index ranges from 0 to 1.0. It is computed using the following formula:

$$P_i = \frac{N_c}{N} \tag{2}$$

where P_i is the item difficulty index of item i, N_c is the number of test takers who responded item i correctly, and N is the total number of test takers [47].

We calculated the item difficult indexes of the 54 items of the test using Equation 2. The item difficulty indexes of the items are presented in Table 2 and each item is classified as an easy item if the value is above 0.85; a moderate item if the value is between 0.5 and 0.85; and a hard item if the value is below 0.5 based on the classification by the Office of Educational Assessment at University of Washington [3]. The item difficulty indexes ranged from 0.15 to 1.0 with an average of 0.64. Among the 54 items, there were 17 easy items, 19 moderate items, and 18 hard items. The items were almost evenly spread over the difficulties.

Item Discrimination Index The item discrimination index indicates that how well an item distinguishes between high scored test takers and low scored test takers. It can be obtained by a portion of the difference between test takers who answered the item correctly among the upper group and test takers who answered the item correctly among the lower group. The value of the index ranges from -1.0 to 1.0. It is computed using the following formula:

$$D_i = \frac{N_U - N_L}{N} \tag{3}$$

where D_i is the item discrimination index of item i, N_U is the number of test takers who responded item i correctly in the upper group, N_L is the number of test takers who responded item i correctly in the lower group, and N is the total number of test takers [47].

We calculated the item discrimination indexes of the 54 items of the test using Equation 3. The item discrimination indexes of the items are presented in Table 2 and each item is classified as a high discriminating item if the value is above 0.3; a medium discriminating item if the value is between 0.1 and 0.3; and a low discriminating item if the value is below 0.1 [3]. The indexes ranged from -0.04 to 0.66 with a average of 0.28. Among the 54 items, there were 25 high discriminating items, 16 medium discriminating items, and 13 low discriminating items.

3.5 Phase V: Test Item Selection

Based on the results in Phase IV, we carefully reviewed all the items in order to filter out inappropriate items and select the final items for the VLAT. When selecting test items for the final set, both the difficulty and the discrimination of each item should be considered. This process should be carefully done because each item was directly related to a data visualization and an associated task, thus it might provide useful information in future decision-making [10].

As shown in Table 2, easy items had low discrimination and hard items had high discrimination in general. However, as exceptional cases, three hard items: Item 24, Item 53, and Item 55, had low discrimination. In particular, Item 24 had a negative discrimination value (D = -0.04). Since items with negative discrimination values were not desirable in the test, we decided to drop Item 24 to improve the quality of the test. Another two hard items, Item 53 and Item 55, had quite low discrimination values. Even though they were hard and low discriminating items, we decided to retain the items in order to keep the quality of the test in terms of reliability (see Section 3.6). Note that there is no clear criteria for selecting the final items. Test developers need to determine whether an item is included in the final set or not [47]. As we did in here, some test developers may select test items liberally to collect more information from the measurement results. On the contrary, some test developers may select a few test items (e.g., only high discriminating items) in order to design an efficient test.

In consequence, we finalized the test items of the VLAT with the 53 items. We initially generated the 61 items based on the test blueprint, but dropped the seven items after the content validity evaluation and the one item after the item analysis. The final set of test items consisted of 34 four-option multiple-choice items, 3 three-option multiple-choice items, and 16 true-false items. The final set of items had an average of 0.66 content validity ratio, an average of 0.65 item difficulty index, and an average of 0.29 item discrimination index (all the 53 test items of the VLAT: associated data visualizations, stems, and options, are included in the supplemental material).

3.6 Phase VI: Reliability Evaluation

For the last phase, we evaluated the quality of the VLAT with the final 53 test items in terms of reliability. The reliability of the VLAT was estimated based on the reliability coefficient omega [35]. As we discussed in Section 2.2, this modern method to estimate the reliability was more appropriate than traditional methods (e.g., test-retest reliability and Cronbach's coefficient alpha) in a practical test development situation [21, 35]. The result showed that the VLAT had acceptably good reliability ($\omega = .76$), and it indicated that scores on the test were consistent and were not unduly influenced by random error [39]. In order to make sure the quality of the VLAT, we also observed the coefficient omega without Item 53 and Item 55. The value slightly decreased but still showed the good level of reliability ($\omega = .75$).

4 Visualization Literacy and Aptitude for Learning An Unfamiliar Visualization

In the previous section, we systematically developed a visualization literacy assessment test (VLAT). In order to examine if the meaning of scores on the VLAT was able to expand, we tested a potential relationship between the current level of visualization literacy and the aptitude for learning an unfamiliar visualization with some forms of education.

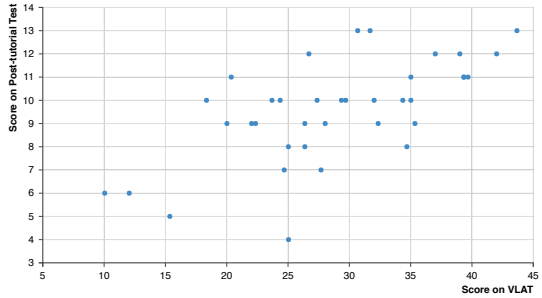


Fig. 3. A scatterplot that represents the scores on the VLAT and the post-tutorial test ($r = .64$, $n = 37$, $p < .001$).

In here, aptitude refers to the ability to learn if appropriate education is provided [47]. For users' visualization literacy, we administered the VLAT. For users' aptitude for leaning an unfamiliar visualization, we adopted an online learning tutorial and test items about Parallel Coordinates Plot (PCP) developed by Kwon and Lee [31].

Participants A total of 46 participants were originally recruited through MTurk. We applied the same requirements as the participants in Phase III (see Section 3.4.1). So, we ruled out only two random clickers but there was no self-reported non-native English speaker and self-reported color blind. We also ruled out seven participants who self-reported that they were aware of PCP in advance in order to see the pure aptitude for learning PCP of the participants. As a consequence, a total of 37 participants remained. The remaining participants consisted of 14 females and 23 males with the self-reported age range of 22 to 58 ($M = 36$). Everybody had an education level of high school or above, 34% of the participants had a bachelor's degree, and 8% of the participants had a master's or a doctoral degree.

Procedure The experiment consisted of two sections: (1) measuring visualization literacy and (2) measuring aptitude for learning an unfamiliar visualization. For the first section, we used the VLAT with the final 53 test items. The procedure was same as the test tryout (see Section 3.4.1). After the participants finished the first section, they were redirected to an online learning tutorial about PCP. We provided six static tutorial pages to describe how PCP was constructed with a sample dataset table, how to interpret PCP, and how to utilize embedded interaction techniques: sorting, brushing, and axes reordering, for the participants. After the tutorial material, we asked the participants to answer 13 test items that were associated with PCP.

Results We observed that the corrected scores of the test takers on the VLAT ranged from 10.05 to 43.67 out of 53 ($M = 28.82$, $SD = 8.16$) and the scores were normally distributed according to the result of Shapiro-Wilk test ($W = .98$, $p = .79$). The scores on the post-tutorial test ranged from 4 to 13 out of 13 ($M = 9.57$, $SD = 2.18$) and the scores were normally distributed ($W = .95$, $p = .10$) as well.

In order to assess the relationship between the visualization literacy and the aptitude for learning, we calculated a Pearson's product-moment correlation coefficient between the VLAT scores and the post-tutorial scores. Figure 3 shows the result that there was a positive correlation between the two scores ($r = .64$, $n = 37$, $p < .001$). This Pearson's correlation coefficient was reasonably high in Psychological Measurement [47]. This result indicates that a users' visualization literacy has a fairly high and positive correlation with the users' aptitude for learning an unfamiliar visualization.

5 Discussion

5.1 Validity Evidence of the VLAT

In order for scores on the VLAT to have the meaning that is intended when the instrument was developed and in order for the interpretation of the scores to be supported, evidence for the validity of the VLAT should be gathered [5, 36]. Furthermore, the evidence is important for other researchers and practitioners to administrate the VLAT for their own purposes. While developing the instrument, we provided major

Table 2. The final set of test items in the VLAT and the content validity ratio (*CVR*), item difficulty index (*P*), and item discrimination index (*D*) of each item. Item 24 is excluded from the final set based on the results of item analysis. Each item is classified as an easy, moderated, or hard item according to the value of *P*, and also classified as a high, medium, or low discriminating item according to the value of *D*.

Item ID	Visualization	Task	Stem	*CVR*	*P*	*D*
Item 1	Line Chart	Retrieve Value	What was the price of a barrel of oil in February 2015?	1	0.95	0.07
Item 2		Find Extremum	In which month was the price of a barrel of oil the lowest in 2015?	1	0.97	0.06
Item 3		Determine Range	What was the price range of a barrel of oil in 2015?	1	0.56	0.66
Item 4		Find Correlations/Trends	Over the course of the second half of 2015, the price of a barrel of oil was _____.	1	0.98	0.03
Item 5		Make Comparisons	About how much did the price of a barrel of oil fall from April to September in 2015?	0.2	0.77	0.44
Item 6	Bar Chart	Retrieve Value	What is the average internet speed in Japan?	1	0.88	0.21
Item 7		Find Extremum	In which country is the average internet speed the fastest in Asia?	1	0.98	0.05
Item 8		Determine Range	What is the range of the average internet speed in Asia?	0.6	0.54	0.61
Item 9		Make Comparisons	How many countries in Asia is the average internet speed slower than Thailand?	0.2	0.4	0.23
Item 10	Stacked Bar Chart	Retrieve Value (Absolute Value)	What is the cost of peanuts in Las Vegas?	0.2	0.38	0.66
Item 11		Retrieve Value (Relative Value)	About what is the ratio of the cost of a sandwich to the total cost of room service in Seattle?	0.2	0.36	0.48
Item 12		Find Extremum	In which city is the cost of soda the highest?	0.2	0.69	0.45
Item 14		Make Comparisons (Absolute Value)	The cost of vodka in Atlanta is higher than that of Honolulu.	0.2	0.59	0.52
Item 15		Make Comparisons (Relative Value)	The ratio of the cost of Soda to the cost of Water in Orlando is higher than that of Washington D.C.	0.2	0.47	0.32
Item 16	100% Stacked Bar Chart	Retrieve Value (Relative Value)	What is the approval rating of Republicans among the people who have the education level of Postgraduate Study?	1	0.49	0.57
Item 17		Find Extremum (Relative Value)	What is the education level of people in which the Democrats have the lowest approval rating?	0.6	0.9	0.21
Item 18		Make Comparisons (Relative Value)	The approval rating of Republicans for the people who have the education level of Some College Degree is lower than that for the people who have the education level of Postgraduate Study.	1	0.54	0.54
Item 19	Pie Chart	Retrieve Value (Relative Value)	About what is the global smartphone market share of Samsung?	0.6	0.72	0.34
Item 20		Find Extremum (Relative Value)	In which company is the global smartphone market share the smallest?	0.6	0.98	0.03
Item 21		Make Comparisons (Relative Value)	The global smartphone market share of Apple is larger than that of Huawei.	0.6	1	0
Item 22	Histogram	Retrieve value (Derived Value)	How many people have rated the taxi between 4.0 and 4.2?	1	0.84	0.26
Item 23		Find Extremum (Derived Value)	What is the rating that the people have rated the taxi the most?	1	0.94	0.06
Item 24		Characterize Distribution	The distribution of the taxi passenger rating is generally skewed to the left.	0.2	0.16	-0.04
Item 25		Make Comparisons (Derived Value)	More people have rated the taxi between 4.6 and 4.8 than between 4.2 and 4.4.	1	0.86	0.18
Item 27	Scatterplot	Retrieve Value	What is the weight for the person who is 165.1 cm tall?	1	0.85	0.27
Item 28		Find Extremum	What is the height for the tallest person among the 85 males?	1	0.76	0.39
Item 29		Determine Range	What is the range in weight for the 85 males?	0.6	0.53	0.49
Item 31		Find Anomalies	What is the height for a person who lies outside the others the most?	0.6	0.42	0.29
Item 32		Find Clusters	A group of males are gathered around the height of 176 cm and the weight of 70 kg.	0.2	0.9	0.23
Item 33		Find Correlations/Trends	There is a negative linear relationship between the height and the weight of the 85 males.	1	0.52	0.66
Item 34		Make Comparisons	The weights for males with the height of 188 cm are all the same.	0.6	0.79	0.2
Item 35	Area Chart	Retrieve Value	What was the average price of a pound of coffee beans in September 2013?	1	0.75	0.34
Item 36		Find Extremum	When was the average price of a pound of coffee beans at minimum?	1	0.44	0.33
Item 37		Determine Range	What was the range of the average price of a pound of coffee beans between January 2013 and December 2014?	0.6	0.38	0.31
Item 38		Find Correlations/Trends	Over the course of 2013, the average price of a pound of coffee beans was _____.	1	0.94	0.14
Item 40	Stacked Area Chart	Retrieve Value (Absolute Value)	What was the number of girls named 'Amelia' in 2010 in the UK?	0.2	0.15	0.29
Item 41		Retrieve Value (Relative Value)	About what was the ratio of the number of girls named 'Olivia' to those named 'Isla' in 2014 in the UK?	0.6	0.25	0.29
Item 42		Find Extremum	Over the course of years between 2009 and 2014, when was the number of girls named 'Amelia' at the maximum?	0.6	0.97	0.04
Item 44		Find Correlations/Trends	The number of girls named 'Isla' was _____ from 2009 to 2012.	1	0.96	0.09
Item 45		Make Comparisons (Absolute Value)	In the UK, the number of girls named 'Amelia' in 2014 was more than it was in 2013,	0.6	0.2	0.17
Item 46		Make Comparisons (Relative Value)	Over the course of years between 2009 and 2014, the number of girls named 'Isla' was always more than 'Olivia'.	0.2	0.24	0.2
Item 47	Bubble Chart	Retrieve Value	What is the total length of the metro system in Beijing?	1	0.41	0.46
Item 48		Find Extremum	Which city's metro system has the largest number of stations?	1	0.69	0.41
Item 49		Determine Range	What is the range of the total length of the metro systems?	0.6	0.29	0.46
Item 51		Find Anomalies	Which city's metro system does lie outside the relationship between the total system length and the number of stations most?	0.2	0.53	0.32
Item 52		Find Clusters	A group of the metro systems of the world has approximately 300 stations and around a 200 km system length.	0.2	0.59	0.5
Item 53		Find Correlations/Trends	In general, the ridership of the metro system increases as the number of stations increases.	0.2	0.26	0.09
Item 54		Make Comparisons	The metro system in Shanghai has more ridership than the metro system in Beijing.	1	0.8	0.33
Item 55	Choropleth Map	Retrieve Value (Approximate Value)	What was the unemployment rate for Indiana (IN) in 2015?	0.6	0.24	0.01
Item 56		Find Extremum (Approximate Value)	In which state was the unemployment rate the highest in 2015?	0.6	0.97	0.06
Item 57		Make Comparisons (Approximate Value)	In 2015, the unemployment rate for Washington (WA) was higher than that of Wisconsin (WI).	0.6	0.92	0.15
Item 59	Treemap	Find Extremum (Relative Value)	For which website was the number of unique visitors the largest in 2010?	0.6	0.68	0.37
Item 60		Make Comparisons (Relative Value)	The number of unique visitors for Amazon was more than that of Yahoo in 2010.	0.2	0.42	0.38
Item 61		Identify the Hierarchical Structure	Samsung is nested in the Financial category.	1	0.92	0.13

evidence for the validity of the VLAT: the test content-related evidence and the test reliability-related evidence. In terms of the content-related evidence, every test item with an associated visualization was reviewed by the domain experts, and the result showed that the final set of items in the VLAT had an average of 0.66 content validity ratio. It would be interpreted that the test items reflect required knowledge for visualization literacy, sample cognitive tasks underlying visualization literacy well, and measure what the test is intended to measure [36].

In addition to the content-related evidence, we provided the reliability-related evidence for the validity of the VLAT. At the end of the test construction procedure, the reliability coefficient omega was calculated based on the test scores from the tryout, and the result

showed that the VLAT had acceptably high reliability ($\omega = .76$). This indicates that the most variation in scores among test takers is due to the variation in the skill, not random measurement error [36]. This is to say that scores on the VLAT are consistent and reproducible. Note that both types of evidence are commonly collected evidence in the test development procedure. Besides the evidence, test developers are able to consider further evidence through other methods, for example, factor analysis and differential item functioning analysis.

On top of the validity evidence, the normally distributed test takers' scores would simplify the interpretation of individual scores on the VLAT. A normal distribution of scores is what most test developers intend to pursue and it provides additional information to the test developers [15]. For example, if an individual score is 45.07 ($M + 2SD$), approximately 97.6% of all the scores fall below the score.

5.2 Reaching out to the Users

The developed test for measuring visualization literacy can be used to understand the everyday users of data visualizations. A variety of follow-up studies can help us get a deeper understanding of our target user population. Here we provide a couple of potential areas, but note that there are other possible studies to pursue.

Tailored Education After one takes the VLAT, the score on the test will promote the self-understanding of his/her level of visualization literacy. The test taker will not only know an absolute score but also know a relative standing in the whole test takers through the percentile rank of the score. Furthermore, the test taker will know strengths and weaknesses in reading and interpreting visually represented data in terms of data visualization types and tasks. For instance, the test taker will see misconceptions about specific visualization types and errors in specific visualization tasks from the test results. On top of that, visualization researchers and practitioners need to find a way to improve users' visualization literacy in order to promote the usage of data visualizations and enable users to do critical thinking through the visualizations. Individual strengths and weaknesses in reading and interpreting data visualizations are directly related to visualization literacy education and training. Not only scores on the VLAT but also specific test results will be critical considerations for designing education and training systems to improve visualization literacy.

Large-scale Test on Data Visualization Users As visualization researchers, we often wonder what percentage of people correctly comprehend certain types of data visualizations. We can conduct large-scale tests regularly using standardized instruments to estimate the user population's level of visualization literacy. In addition, the areas of visualizations that need more education and training will be enlightened based on the large-scale test results. This can lead us to answering a more interesting question of how the level of visualization literacy changes from a point to another. Furthermore, by combining with other demographic information, the researchers will discover relationships between visualization literacy and other personal traits, social status, and educational backgrounds, especially in mathematics and statistics. Eventually, the VLAT and further instruments can serve as tools to capture the areas of improvements that the Information Visualization community need to work on.

5.3 Toward the Investigation of Data Comprehension

We demonstrated that there was a positive correlation between one's visualization literacy level and the aptitude for learning and using an unfamiliar visualization. However, this does not mean that a user's score on the test can be directly translated to the user's ability to comprehend data and execute analytic tasks with visualization tools. To accurately measure the ability, we should derive skills and knowledge to process data analytic tasks. We need to investigate users' abilities to understand an analytical goal, to strategically plan the sequence of analytic tasks, and to proceed the tasks using given visualization tools. Furthermore, we need to understand perceptual skills for pattern detection and cognitive skills to process the series of tasks efficiently for analyzing data, gaining insights, and generating knowledge [43]. We also need to understand whether users fully understand given visualization tools to utilize for problems they attempt to solve because many users encounter roadblocks in understanding and choosing proper visual representations for the problems [30]. In sum, data comprehension and visual analytic tasks require many different levels of abilities and further work needs to investigate a holistic framework that encompasses the entire pipeline of visual analytic processes.

5.4 Limitations and Future Work

Even though we rigorously followed the test development procedure and provided the validity evidence of the test, there are several limitations that readers need to keep in mind.

Since the test tryout was conducted on a crowdsourcing platform, we were not able to control the participants as in controlled lab studies. We noted below-chance performances to filter out random clickers but we did not investigate further. Even though the crowdsourcing platform is commonly used for conducting studies, the participants' level of understanding of the task or level of engagement in testing might have influenced their performance. Thus, systematic ways to gauge the participants' attention are necessary [26].

In addition, a shade of meaning in a phrase of an item might have influenced the participants' reasoning even though we tried to maintain the lexical level consistent through the pilot work. Likewise, a slight difference in the quality of the data visualizations, which were the test contents, might also affect the test takers' performance. Thus, writing items and authoring data visualizations should be done and reviewed very carefully.

Lastly, the current VLAT might be limited to comprehensively measure visualization literacy because the test covers the 12 data visualization types and the eight visualization tasks. Furthermore, it could be argued that visualization literacy is too complex to assign a single number that represents the ability because there are many possible tasks with data visualizations and each task consists of a variety of visual routines [48]. However, we believe that having a validated and reliable test is an important first step to move forward and lead to further investigation. By analyzing the answers of each item, we could get more information about visualization literacy (e.g., common misconceptions and errors of users). Investigating the relationships between visualization literacy and other cognitive traits, abilities, or demographic characteristics would also be interesting. Finally, exploring the dimensionality of the VLAT using factor analysis models should be a critical next step to expand and improve the test.

6 Conclusions

In this study, we systematically developed a visualization literacy assessment test (VLAT) by following the test construction procedure in Psychological and Educational Measurement. The VLAT consisted of 12 data visualizations and 53 test items that covered eight major tasks. While developing the VLAT, we showed the major evidence for the validity of the test in terms of the test content and the test reliability. Furthermore, we demonstrated the relationship between users' scores on the VLAT and the aptitude for learning an unfamiliar visualization.

This study makes the following contributions to the Information Visualization community. We believe that we provide one of the earliest validated and reliable instruments measuring visualization literacy of users. The procedure of test development presented in this paper would be an exemplary way to construct further instruments in the Information Visualization community. With the VLAT, the community will gain a better understanding of users' visualization literacy and the understanding will have a positive effect on making sound decisions in designing, developing, and evaluating. Moreover, it can be a clue about how to improve overall visualization literacy through educational experiences and training.

Acknowledgments

We acknowledge Ji Soo Yi, Jeremy Boy, and Niklas Elmqvist for initializing and developing the idea of this work. We thank Anne Traynor for constructive and practical advice on the procedure of test development. We also thank visualization experts, Heidi Lam, Zhicheng Liu, Youn-ah Kang, Jaegul Choo, and Sungahn Ko, who participated in the content validity evaluation of the VLAT. Finally, the Google Research Award (Winter, 2013) has made this study possible.

REFERENCES

[1] http://www.merriam-webster.com/dictionary/literacy.

[2] https://www.ets.org/literacy/research/what/.

[3] http://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/.

[4] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization*, pages 111–117, 2005.

[5] American Educational Research Association, National Council on Measurement in Education, and American Psychological Association. *Standards for Educational and Psychological Tests and Manuals*. American Psychological Association, 2014.

[6] A. J. Berinsky, G. A. Huber, and G. S. Lenz. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 20(3):351–368, 2012.

[7] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. Esri Press, 2010.

[8] 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*, Advance Online Publication:1–16, 2015.

[9] J. Boy, L. Eveillard, F. Detienne, and J.-D. Fekete. Suggested interactivity: Seeking perceived affordances for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):639–648, 2016.

[10] J. Boy, R. 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.

[11] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[12] California Department of Education. *The California Common Core State Standards: Mathematics*. 2013.

[13] C. M. Carswell. Choosing specifiers: an evaluation of the basic tasks model of graphical perception. *Human Factors*, 34(5):535–554, 1992.

[14] Y. Chen, J. Yang, and W. Ribarsky. Toward effective insight management in visual analytics systems. In *IEEE Pacific Visualization Symposium*, pages 49–56, 2009.

[15] R. J. Cohen and M. E. Swerdlik. *Psychological Testing and Assessment: An Introduction to Tests and Measurement*. McGraw-Hill, 2002.

[16] F. R. Curcio. Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education*, 18(5):382–393, 1987.

[17] R. delMas, J. Garfield, and A. Ooms. Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In *Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy*, 2005.

[18] C. DeMars. *Item Response Theory*. Oxford University Press, 2010.

[19] R. F. DeVellis. *Scale Development: Theory and Applications*. Sage publications, 2012.

[20] J. Diamond and W. Evans. The correction for guessing. *Review of Educational Research*, 43(2):181–191, 1973.

[21] T. J. Dunn, T. Baguley, and V. Brunsden. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3):399–412, 2014.

[22] X. Fan. Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3):357–381, 1998.

[23] R. B. Frary. Formula scoring of multiple-choice tests (Correction for guessing). *Educational Measurement: Issues and Practice*, 7(2):33–38, 1988.

[24] S. N. Friel, F. R. Curcio, and G. W. Bright. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, 32(2):124–158, 2001.

[25] M. Galesic and R. Garcia-Retamero. Graph literacy a cross-cultural comparison. *Medical Decision Making*, 31(3):444–457, 2011.

[26] J. K. Goodman, C. E. Cryder, and A. Cheema. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2013.

[27] C. Huff and D. Tingley. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, 2(3):1–12, 2015.

[28] Indiana Department of Education. *The 2014 Indiana Academic Standards for Mathematics*. 2014.

[29] B. Johnson and B. Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *IEEE Conference on Visualization*, pages 284–291, 1991.

[30] B. C. Kwon, B. Fisher, and J. S. Yi. Visual analytic roadblocks for novice investigators. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 3–11, 2011.

[31] B. C. Kwon and B. Lee. A comparative evaluation on online learning approaches using parallel coordinate visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 993–997, 2016.

[32] C. H. Lawshe. A quantitative approach to content validity. *Personnel Psychology*, 28(4):563–575, 1975.

[33] S. Lee, S.-H. Kim, Y.-H. Hung, H. Lam, Y.-a. Kang, and J. S. Yi. How do people make sense of Unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):499–508, 2016.

[34] A. V. Maltese, J. A. Harsh, and D. Svetina. Data visualization literacy: Investigating data interpretation along the novice-expert continuum. *Journal of College Science Teaching*, 45(1):84–90, 2015.

[35] R. P. McDonald. *Test Theory: A Unified Treatment*. Lawrence Erlbaum Associates, 1999.

[36] R. J. Mislevy and G. D. Haertel. Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(4):6–20, 2006.

[37] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[38] New York State Education Department. *New York State P-12 Common Core Learning Standards for Mathematics*. 2013.

[39] J. C. Nunally and I. H. Bernstein. *Psychometric Theory*. McGraw-Hill, 1978.

[40] A. Pandey, A. Manivannan, O. Nov, M. Satterthwaite, and E. Bertini. The persuasive power of data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2211–2220, 2014.

[41] D. Qi. An inquiry into language-switching in second language composing processes. *Canadian Modern Language Review*, 54(3):413–435, 1998.

[42] P. Ruchikachorn and K. Mueller. Learning visualizations by analogy: Promoting visual literacy through visualization morphing. *IEEE Transactions on Visualization and Computer Graphics*, 21(9):1028–1044, 2015.

[43] A. Sacha, Dominik aflawshend Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, 2014.

[44] N. Schmitt. Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4):350–353, 1996.

[45] 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.

[46] K. Sijtsma. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1):107–120, 2008.

[47] R. L. Thorndike and E. Hagen. *Measurement and Evaluation in Psychology and Education*. Pearson, 2010.

[48] S. Ullman. Visual routines. *Cognition*, 18(1–3):97–159, 1984.

[49] H. Wainer. Understanding graphs and tables. *Educational Researcher*, 21(1):14–23, 1992.

[50] Washington State Office of Superintendent of Public Instruction. *Common Core State Standards for Mathematics*. 2013.

[51] W. Willett, J. Heer, and M. Agrawala. Strategies for Crowdsourcing Social Data Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 227–236, 2012.

[52] H. Yang, Y. Li, and M. X. Zhou. Understand users' comprehension and preferences for composing information visualizations. *ACM Transactions on Computer-Human Interaction*, 21(1):6:1–6:30, 2014.

[53] N. Yau. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. Wiley, 2011.

[54] J. S. Yi. Implications of individual differences on evaluating information visualization techniques. In *Proceedings of the 3rd BELIV Workshop*, 2010.

[55] J. S. Yi, Y.-a. Kang, J. T. Stasko, and J. A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007.