A Principled Way of Assessing Visualization Literacy

Jeremy Boy, Ronald A. Rensink, Enrico Bertini, and Jean-Daniel Fekete *Senior Member, IEEE*

Abstract— We describe a method for assessing the visualization literacy (VL) of a user. Assessing how well people understand visualizations has great value for research (*e. g.,* to avoid confounds), for design (*e. g.,* to best determine the capabilities of an audience), for teaching (*e. g.,* to assess the level of new students), and for recruiting (*e. g.,* to assess the level of interviewees). This paper proposes a method for assessing VL based on Item Response Theory. It describes the design and evaluation of two VL tests for line graphs, and presents the extension of the method to bar charts and scatterplots. Finally, it discusses the reimplementation of these tests for fast, effective, and scalable web-based use.

Index Terms—Literacy, Visualization literacy, Rasch Model, Item Response Theory

1 Introduction

In April 2012, Jason Oberholtzer posted an article describing two charts that portray Portuguese historical, political, and economic data [33]. While acknowledging that he is not an expert on those topics, Oberholtzer claims that thanks to the charts, he feels like he has "a well-founded opinion on the country." He attributes this to the simplicity and efficacy of the charts. He then concludes by stating: "Here's the beauty of charts. We all get it, right?"

But do we all really get it? Although the number of people familiar with visualization continues to grow, it is still difficult to estimate anyone's ability to read graphs and charts. When designing a visualization for non-specialists or when conducting an evaluation of a new visualization system, it is important to be able to pull apart the potential efficiency of the visualization and the actual ability of users to understand it.

In this paper, we address this issue by creating a set of *visualization literacy* (VL) tests for line graphs, bar charts, and scatterplots. At this point, we loosely define visualization literacy as *the ability to use well-established data visualizations (e. g., line graphs) to handle information in an effective, efficient, and confident manner.*

To generate these tests, we develop here a method based on Item Response Theory (IRT). Traditionally, IRT has been used to assess examinees' abilities via predefined tests and surveys in areas such as education [24], social sciences [14], and medicine [29]. Our method uses IRT in two ways: first, in a *design phase*, we evaluate the relevance of potential test items; and second, in an *assessment phase*, we measure users' abilities to extract information from graphical representations. Based on these measures, we then develop a series of tests for fast, effective, and scalable web-based use. The great benefit of this method is that inherits IRT's property of making ability assessments that are based not only on raw scores, but on a model that captures the standing of users on a latent trait (*e. g.,* the ability to use various graphical representations).

As such, our main contributions are as follows:

- a useful definition of visualization literacy,
- a method for: 1) assessing the relevance of visualization literacy test items, 2) assessing an examinee's level of VL, 3) creating fast and effective assessments of VL for well established visualization techniques and tasks; and
- an implementation of four online tests, based on our method.

Our immediate motivation for this work is to design a series of tests that can help Information Visualization (InfoVis) researchers detect low-ability participants when conducting online studies, in order to avoid possible confounds in their data. This requires the tests to be short, reliable, and easy to administer. However, such tests can also be applied to many other situations, such as:

- designers who want to know how capable of understanding visualizations their targeted audience is;
- teachers who want to make an assessment of the acquired knowledge of freshmen;
- practitioners who need to hire capable analysts; and
- education policy-makers who may want to set a standard for visualization literacy.

This paper is organized in the following way. It begins with a background section that defines the concept of literacy and discusses some of its best-known forms. Also introduced are the theoretical constructs of *information comprehension* and *graph comprehension*, along with the concepts behind Item Response Theory. Next, Section 3 presents the basic elements of our approach. Section 4 shows how these can be used to create and administer two VL tests using line graphs. In Section 5, our method is extended to bar charts and scatterplots. Section 6 describes how our method can be used to redesign fast, effective, and scalable web-based tests. Finally, Section 7 provides a set of "take-away" guidelines for the development of future tests.

2 Background

Very few studies investigate the ability of a user to extract information from a graphical representation such as a line graph or a bar chart. And of those that do, most make only higher-level assessments: they use such representations as a way to test mathematical skills, or the ability to handle *uncertainty* [13, 31, 32, 34, 49]. A few attempts do focus more on the interpretation of graphically-represented quantities [18, 20], but they base their assessments only on raw scores and limited test items. This makes it difficult to create a true measure of VL.

2.1 Literacy

2.1.1 Definition

The online Oxford dictionary defines *literacy* as "the ability to read and write". While historically this term has been closely tied to its textual dimension, it has grown to become a broader concept. Taylor proposes the following: "Literacy is a *gateway skill* that opens to the potential for new learning and understanding" [44].

Given this broader understanding, other forms of literacy can be distinguished. For example, *numeracy* was coined to describe the skills needed for reasoning and applying simple numerical concepts. It was

- *Jeremy Boy is with Inria, Telecom ParisTech, and EnsadLab. E-mail: myjyby@gmail.com.*
- *Ronald A. Rensink is with the University of British Columbia. E-mail: rensink@psych.ubc.ca.*
- *Enrico Bertini is with NYU Polytechnic School of Engineering. E-mail: ebertini@poly.edu.*
- *Jean-Daniel Fekete is with Inria. E-mail: jean-daniel.fekete@inria.fr.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014. Date of publication 11 Aug. 2014; date of current version 9 Nov. 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.
Digital Object Identifier 10.1109/TVCG.2014.2346984

intended to "represent the mirror image of [textual] literacy" [43, p. 269]. Like [textual] literacy, numeracy is a gateway skill.

With the advent of the Information Age, several new forms of literacy have emerged. *Computer literacy* "refers to basic keyboard skills, plus a working knowledge of how computer systems operate and of the general ways in which computers can be used" [35]. *Information literacy* is defined as the ability to "recognize when information is needed", and "the ability to locate, evaluate, and use effectively the needed information" [22]. *Media literacy* commonly relates to the "ability to access, analyze, evaluate and create media in a variety of forms" [51].

2.1.2 Higher-level Comprehension

In order to develop a meaningful measure of any form of literacy, it is necessary to understand the various components involved, starting at the higher levels. Friel et al. [16] suggest that comprehension of information in written form involves three kinds of tasks: locating, integrating, and generating information. *Locating tasks* require the reader to find a piece of information based on given cues. *Integrating tasks* require the reader to aggregate several pieces of information. *Generating tasks* not only require the reader to process given information but also require the reader to make document-based inferences or to draw on personal knowledge.

Another important aspect of information comprehension is question asking, or *question posing*. Graesser et al. [17] posit that question posing is a major factor in text comprehension. Indeed, the ability to pose low-level questions, *i. e.,* to identify a series of low-level tasks, is essential for information retrieval and for achieving higher-level, or deeper, goals.

2.1.3 Assessment

Several literacy tests are currently in common use. The two most important are the UNESCO's Literacy Assessment and Monitoring Programme (LAMP) [48], and the OECD's Programme for International Student Assessment (PISA) [34]. Other international assessments include the Adult Literacy and Lifeskills Survey (ALL) [30], the International Adult Literacy Survey (IALS) [23], and the Miller Word Identification Assessment (MWIA) [28].

Assessments are also made using more local scales like the US National Assessment of Adult Literacy (NAAL) [3], the UK's Department for Education Numeracy Skills Tests [13], or the University of Kent's Numerical Reasoning Test [49].

Most of these tests, however, take basic literacy skills for granted, and focus on higher-level assessments. For example the PISA test is designed for 15 year-olds who are finishing compulsory education. This implies that examinees should have already learned—and still remember—the basic skills required for reading and counting. It is only when examinees clearly fail these tests that certain measures are deployed to evaluate the lower-level skills.

NAAL provides a set of 2 complementary tests for examinees who fail the main textual literacy test [3]: the Fluency Addition to NAAL (FAN) and The Adult Literacy Supplemental Assessment (ALSA). These focus on adults' ability to read single words and small passages.

Meanwhile, MWIA tests whole-word dyslexia. It has 2 levels, each of which contains 2 lists of words, one Holistic and one Phonetic, that examinees are asked to read aloud. Evaluation is based on time spent reading and number of words missed. Proficient readers should find such tests extremely easy, while low ability readers should find them more challenging.

2.2 Visualization Literacy

2.2.1 Definition

The view of literacy as a gateway skill can also be applied to the extraction and manipulation of information from graphical representations. In particular, it can be the basis for what we will refer to as *visualization literacy* (VL): *the ability to confidently use a given data visualization to translate questions specified in the data domain into visual queries in the visual domain, as well as interpreting visual patterns in the visual domain as properties in the data domain.*

This definition is related to several others that have been proposed concerning visual messages. For example, a long-standing and often neglected concept is *visual literacy*. This has been defined as the "ability to understand, interpret and evaluate visual messages" [7]. Visual literacy is rooted in semiotics, *i. e.,* the study of signs and sign processes, which distinguishes it from visualization literacy. While it has probably been the most important form of literacy to date, it is nowadays frowned upon, and general literacy tests do not take it into account.

Taylor [44] has advocated for the study of *visual information literacy,* while Wainer has advocated for *graphicacy* [50]. Depending on the context, these terms refer to the ability to read charts and diagrams, or to qualify the merging of visual and information literacy teaching [2]. Because of this ambiguity, we prefer the more general term "visualization literacy."

2.2.2 Higher-level Comprehension

Bertin [4] proposed three levels on which a graph may be interpreted: elementary, intermediate, and comprehensive. The *elementary level* concerns the simple extraction of information from the data. The *intermediate level* concerns the detection of trends and relationships. The *comprehensive level* concerns the comparison of whole structures, and inferences based on both data and background knowledge. Similarly, Curcio [12] distinguishes three ways of reading from a graph: from the data, between the data, and beyond the data[1].

The higher-level cognitive processes behind the reading of graphs has been the concern of the area of *graph comprehension*. This area studies the specific expectations viewers have for different graph types [47], and has highlighted many differences in the understanding of novices and expert viewers [15, 25, 26, 45].

Several influential models of graph comprehension have been proposed. For example, Pinker [36] describes a three-way interaction between the visual features of a display, processes of perceptual organization, and what he calls the *graph schema*, which directs the search for information in the particular graph. Several other models are similar (see Trickett and Trafton [46]). All involve the following steps:

1. the user has a pre-specified goal to extract a specific piece of information
2. the user looks at the graph and the graph schema and gestalt processes are activated
3. the salient features of the graph are encoded, based on these gestalt principles
4. the user now knows which cognitive/interpretative strategies to use, because the graph is familiar
5. the user extracts the necessary goal-directed visual chunks
6. the user may compare 2 or more visual chunks
7. the user extracts the relevant information to satisfy the goal

The "visual chunking" mentioned above consists in segmenting a visual display into smaller parts, or chunks [25]. Each chunk represents a set of entities that have been grouped according to gestalt principles. Chunks can in turn be subdivided into smaller chunks.

Shah [42] identifies two cognitive processes that occur during stages 2 through 6 of this model:

1. a top-down process where the viewer's prior knowledge of semantic content influences data interpretation, and
2. a bottom-up process where the viewer shifts from perceptual processes to interpretation.

These processes are then interactively applied to various chunks, which suggests that data interpretation is a serial and incremental process. However, Carpenter & Shah [8] have shown that graph comprehension, and more specifically visual feature encoding, is more of an iterative process than a straight-forward serial one.

[1] For further reference, refer to Friel et al.'s *Taxonomy of Skills Required for Answering Questions at Each Level* [16].

Freedman & Shah [15] relate the top-down and bottom-up processes to a construction and an integration phase, respectively. During the construction phase, the viewer activates prior graphical knowledge, *i. e.,* the graph schema, and domain knowledge to construct a coherent conceptual representation of the available information. During the integration phase, disparate knowledge is activated by "reading" the graph and is combined to form a coherent representation. These two phases take place in alternating cycles. This suggests that domain knowledge can influence the interpretation of graphs. However, highly visualization-literate people should suffer less influence of both the top-down and bottom-up processes [42].

2.2.3 Assessment

Relatively little has been done on the assessment of literacy involving graphical representations. However, interesting work has been done on measuring the perceptual abilities of a user to extract information from these. For example, various studies have demonstrated that users can perceive slope, curvature, dimensionality, and continuity in line graphs (see [11]). Correll et al. [11] have also shown that users can make judgements about aggregate properties of data using these graphs.

Scatterplots have also received some attention. For example, studies have examined the ability of a user to determine Pearson correlation r [5, 9, 27, 37, 39]. Several interesting results have been obtained, such as general tendency to underestimate correlation, especially in the range $.2 < |r| < .6$, and an almost complete failure to perceive correlation when $|r| < .2$.

Concerning the outright assessment of literacy, the only relevant research work we know of is Wainer's study on the difference in graphicacy levels between third-, fourth-, and fifth-grade children [50]. He presents the design of an 8-item test using several visualizations, including line graphs and bar charts. He then describes his use of Item Response Theory [52] to score the test results, and shows the effectiveness of this method for assessing abilities. His conclusion is that children reach "adult levels of graphicacy" as soon as the fourth-grade, leaving "little room for further improvement." However, it is unclear what these "adult levels" are. If we look at textual literacy, some children are more literate than certain adults. People may also forget these skills if they do not regularly practice. Thus, while very useful, we consider Wainer's work to be limited. What is needed is a way to assess *adult levels* of visualization literacy.

2.3 Item Response Theory and the Rasch Model

Consider what we would like in an effective VL test. To begin with, it should cover a certain range of abilities, each of which could be measured by specific scores. Imagine such a test has 10 items, which are marked 1 when answered correctly, and 0 otherwise. Rob takes the test and gets a score of 2. Jenny also takes the test, and gets a score of 7. We would hope that this means that Jenny is better than Rob at reading graphs. In addition, we would expect that if Rob and Jenny were to take the test again, both would get approximately the same scores, or at least that Jenny would still get a higher score than Rob. We would also expect that whatever VL test Rob and Jenny both take, Jenny will always be better than Rob.

Now imagine that Chris takes the test and also gets a score of 2. If we based our judgement on this *raw* score, we would assume that Chris is as bad as Rob at reading graphs. However, taking a closer look at the items that Chris and Rob got right, we realize that they are different: Rob gave correct answers to the two easiest items, while Chris gave correct answers to two relatively complex items. This would of course require us to know the level of *difficulty* of each item, and would mean that while Chris gave incorrect answers to the easy items, he might still show some ability to read graphs. Thus, we would want the different scores to have "meanings" to help us determine whether Chris was simply lucky (he *guessed* the answers), or whether he is in fact *able* to get the simpler items right, even though he didn't this time.

Imagine now that Rob, Jenny, and Chris take a second VL test. Rob gets a score of 3, Chris gets 4, and Jenny gets 10. We would infer that this test is easier, since the scores are higher. However, looking at the score intervals, we see that Jenny is 7 points ahead of Rob, whereas she was only 5 points ahead in the first test. If we were to truly measure abilities, we would want these intervals to be invariant. In addition, seeing that Chris' score is once again similar to Rob's (knowing that they both got the same items right this time) would lead us to think that they do in fact have similar abilities. We could then conclude that this test provides more *information* on lower abilities than the fist one, since it is able to separate Rob and Chris' scores.

Finally, imagine that all three examinees take a third test, and all get a score of 10. While we might be tempted to conclude that this test is VL-agnostic, it may simply be that its items are too *easy*, and not sufficiently *discriminant*.

One way of fulfilling all of these requirements is by using Item Response Theory (IRT) [52]. This is a model-based approach that does not use response data directly, but transforms them into estimates of a latent trait (*e. g.,* ability), which then serves as the basis of assessment. IRT models have been applied to tests in a variety of fields such as health studies, education, psychology, marketing, economics, social sciences (see [41]), and even graphicacy [50].

The core idea of IRT is that the performance of an examinee depends on both the examinee's ability and the item's difficulty; the goal is then to separate out these two factors. An important aspect of the approach is to project them onto the same scale—that of the latent trait. *Ability*, or standing on the latent trait, is derived from a pattern of responses to a series of test items; *item difficulty* is then defined by the 0.5 probability of success of an examinee with the appropriate ability. For example, an examinee with an ability value of 0 (0 corresponding to an average achiever) will have a 50% probability of giving a correct answer to an item with a difficulty value of 0, corresponding to an average level of difficulty.

IRT offers models for data that are dichotomous (*e. g.,* true/false responses) and polytomous (*e. g.,* responses on likert-like scales). In this paper, we focus on models for dichotomous data. These define the probability of success on an item i by the function:

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}} \tag{1}$$

where θ is an examinee's standing on a latent trait (*i. e.,* his or her ability), and a_i, b_i, and c_i are the *characteristics* of the item. The central characteristic is b, the *difficulty characteristic;* if $\theta = b$, the examinee has a 0.5 probability of giving a correct answer to the item. Meanwhile, a is the *discrimination characteristic.* An item with a very high discrimination value basically sets a sharp threshold at $\theta = b$: examinees with $\theta < b$ have a probability of success of 0, and examinees with $\theta > b$ have a probability of success of 1. Conversely, an item with a low discrimination value cannot clearly separate examinees. Finally, c is the *guessing characteristic.* It sets a lower bound for the extent to which an examinee will guess an answer. We have found c to be unhelpful, so we have set it to zero (no guessing) for our development.

Note that the value of each characteristic is not absolute for a given item: it is relative to the latent trait that the test is attempting to uncover. Therefore, it cannot be expected that the characteristics of identical items be exactly the same in different tests. For example, consider a simple numeracy test with two items: $10 + 20$ (item 1) and $17 + 86$ (item 2). It should be assumed that item 1 is easier than item 2. In other words, the difficulty characteristic of item 2 should be higher than that of item 1. Now if we add another item to the test, say 51×93 (item 3), the most difficult item in the previous version of the test (item 2) will no longer seem so difficult. However, it should still be more difficult than item 1. Thus, while individual characteristics may vary, the general order of difficulty should be preserved. The same goes for ability values (or *ability scores*). If they are to be compared between different tests, the measured latent trait must be the same.

Various IRT models for dichotomous data have been proposed. One is the one-parameter logistic model (1PL), which sets a to a specific value for all items, sets c to zero, and only considers the variation of b. Another is the two-parameter logistic model (2PL), which considers the variations of a and b, and sets c to zero. A third is the three-parameter logistic model (3PL), which considers variations of a, b, and c [6]. As such, 1PL and 2PL can be regarded as special cases of

3PL, where different item characteristics are assigned specific values. A last variant is the Rasch model (RM), which is a special case of 1PL, where $a = 1$[2].

Thus, IRT offers a way to evaluate the relevance of test items during a *design phase* (*e. g.,* how difficult items are, or how discriminant they are), and a way to measure examinees' abilities during an *assessment phase*. These two phases constitute the backbone of the method we present in this paper, which is why we stress that our approach will be successful only if an IRT model fits a set of empirically collected data. Furthermore, its accuracy will depend on how closely an IRT model describes the interaction between examinees' abilities and their responses, *i. e.,* how well the model describes the latent trait. Thus, different variants of IRT models should be tested initially to find the best fit. Finally, it should be mentioned that IRT models cannot be relied upon to "fix" problematic issues in a test. Proper test design is still required.

3 Foundations

In the approach we develop here, test items generally involve a 3-part structure: 1) a stimulus, 2) a task, and 3) a question. The stimuli are the particular graphical representations used. Tasks are defined in terms of the visual operations and mental projections that an examinee should perform to answer a given question. While tasks and questions are usually linked, we emphasize this distinction because early piloting revealed that different "orientations" of a question (*e. g.,* emphasis on particular visual aspects, or data aspects) could affect performance.

To identify possible factors that may influence the difficulty of a test item, we reviewed all the literacy tests that we could find which use graphs and charts as stimuli [13, 18, 20, 31, 32, 34, 49, 50]. Note that our goal is not to investigate the effect of these factors on item difficulty; we present them here merely as elements to be considered in the design phase.

We identified 4 potential stimulus parameters: number of samples, intrinsic complexity (or variability) of the data, layout, and level of distraction. We also found 6 recurring task types: extrema (maximum and minimum), trend, intersection, average, and comparison. Finally, we distinguished 3 different question types: "perception" questions, "high-congruency" questions, and "low-congruency" questions. Each of these are described in the following subsections.

3.1 Stimulus parameters

In our survey, we first focused on identifying parameters to describe the graphical properties of a stimulus. We found four:

Number of samples This refers to the number of graphically encoded elements in the stimulus. Among other things, the value of this parameter can impact tasks that require visual chunking [25].

Complexity This refers to the local and global variability of the data. For example, a dataset of the yearly life expectancy in different countries over a 50 year time period has a low local variation (no dramatic "bounces" between two consecutive years), and low global variation (a relatively stable, linear, increasing trend). In contrast, a dataset of the daily temperature in different countries over a year shows high local variation (temperatures can vary dramatically from on day to the other) and medium global variation (temperature generally rises an decreases only once during the year).

Layout This refers to the structure of the graphical framework and its scales. Layouts can be single (*e. g.,* a 2-dimensional Euclidian space), superimposed (*e. g.,* three axes for a 2-dimensional encoding), or multiple (*e. g.,* several frameworks for a same visualization). Multiple layouts include cutout charts and broken charts [21]. Scales can be single (linear or logarithmic), bifocal, or lense-like.

[2]For a complete set of references on the Rasch model, refer to `http://rasch.org`.

Distraction This refers to the graphical elements present in the stimulus that are not necessary for the task at hand. These are considered to be *distractors.* Correll et al. [11] have shown that even small variations in attributes of distractors can impact perception. However, here we simply use distraction in a Boolean way, *i. e.,* present or not.

3.2 Tasks

Next, we focused on identifying tasks that require only visual intelligence, *i. e.,* purely visual operations or mental projections on a graphical representation. We found six: *Maximum* (T1), *Minimum* (T2), *Variation* (T3), *Intersection* (T4), *Average* (T5), and *Comparison* (T6). All are standard benchmark tasks in InfoVis. T1 and T2 consist in finding the maximum and minimum data points in the graph, respectively. T3 consists in detecting a trend, similarities, or discrepancies in the data. T4 consists in finding the point at which the graph intersected with a given value. T5 consists in estimating an average value. Finally, T6 consists in comparing different values or trends.

3.3 Congruency

Finally, we focused on identifying different types of questions. We found three: perception questions, and high- and low-congruency questions. *Perception* questions refer only to visual aspects of the display (*e. g.,* "what color are the dots?"). Conversely, *congruent* questions refer to semantic aspects of the data. The *level of congruence* is then defined by the "replaceability" of the data-related terms in the question by perceptual terms. A high-congruency question translates into a perceptual query simply by replacing data terms by perceptual terms (*e. g.,* "what is the highest value"/"what is the highest bar?"). A low-congruency question, in contrast, has no such correspondence (*e. g.,* "is A connected to B—in a matrix diagram"/"is the intersection between column A and row B highlighted?").

4 Application To Line Graphs

To illustrate our method, we first created two line graph tests—Line Graphs 1 (LG1) and Line Graphs 2 (LG2)—of slightly different designs, based on the principles described above. We then calibrated them using Amazon's Mechanical Turk (MTurk).

4.1 Design Phase

4.1.1 Line Graphs 1: General Design

For our first test (LG1), we created a set of twelve items using different stimulus parameters and tasks. We hand-tailored each item based on an expected range of difficulty. Piloting had revealed that high variation in item dimensions led to incoherent tests (*i. e.,* IRT models did not fit the response data), implying that when factors vary too much within a test, additional abilities beyond those involved in basic visualization literacy are likely at play. Thus, we kept the number of varying factors low: only distraction and tasks varied. The test used four samples for the stimuli, and a single layout with single scales. A summary is given in Table 1.

Each item was repeated five times[3]. The test was blocked by item, and all items and their repetitions were randomized to prevent carry-over effects. We added an extra condition using a standard table at the beginning of each block to give examinees the opportunity to consolidate their understanding of the new question, and to separate out the comprehension stage of the question-response process believed to occur in cognitive testing [10]. The test was thus composed of 72 trials.

In the following paragraphs, we describe other important design parameters we used in this test.

Scenario The PISA 2012 Mathematics Framework [34] emphasizes the importance of an understandable context for problem solving. The current test focuses on one's community, with problems set in a community perspective.

[3]Early piloting had revealed that examinees would stabilize their search time and confidence after a few repetitions. In addition, repeated trials usually provide more robust measures as medians can be extracted (or means in the case of Boolean values).

LG1

Item ID	Task	Distraction
LG1.1	max	0
LG1.2	min	0
LG1.3	variation	0
LG1.4	intersection	0
LG1.5	average	0
LG1.6	comparison	0
LG1.7	max	1
LG1.8	min	1
LG1.9	variation	1
LG1.10	intersection	1
LG1.11	average	1
LG1.12	comparison	1

LG2

Item ID	Task	Congruency
LG2.1	max	high
LG2.2	min	high
LG2.3	variation	high
LG2.4	intersection	high
LG2.5	average	high
LG2.6	comparison	high
LG2.7	max	low
LG2.8	min	low
LG2.9	variation	low
LG2.10	intersection	low
LG2.11	average	low
LG2.12	comparison	low

BC

Item ID	Task	Samples
BC.1	max	10
BC.2	min	10
BC.3	variation	10
BC.4	intersection	10
BC.5	average	10
BC.6	comparison	10
BC.7	max	20
BC.8	min	20
BC.9	variation	20
BC.10	intersection	20
BC.11	average	20
BC.12	comparison	20

SP

Item ID	Task	Distraction
SP.1	max	0
SP.2	min	0
SP.3	variation	0
SP.4	intersection	0
SP.5	average	0
SP.6	comparison	0
SP.7	max	1
SP.8	min	1
SP.9	variation	1
SP.10	intersection	1
SP.11	average	1
SP.12	comparison	1

Table 1: Designs of Line Graphs 1 (LG1), Line Graphs 2 (LG2), Bar Chart (BC), and Scatterplot (SP). Only varying dimensions are shown. Each item is repeated 6 times, beginning with a table condition (repetitions are not shown). Pink cells in the Item ID column indicate duplicate items in LG1 and LG2. Tasks with the same color coding are the same. Gray cells in the Distraction, Congruency, and Samples columns indicate difference with white cells. The Distraction column uses a binary encoding: 0 = no distractors, 1 = presence of distractors.

To avoid the potential bias of *a priori* domain knowledge, the test was set within the following science-fiction scenario: *The year is 2813. The Earth is a desolate place. Most of mankind has migrated throughout the universe. The last handful of humans remaining on earth are now actively seeking another planet to settle on. Please help these people determine what the most hospitable planet is by answering the following series of questions as quickly and accurately as possible.*

Data The dataset we used had a low-local and medium-global level of variability. It presented the monthly evolution of unemployment in different countries between the years 2000 and 2008. Country names were changed to fictitious planet names listed in Wikipedia, and years were modified to fit the scenario.

Priming and Pacing Before each new block of repetitions, examinees were primed with the upcoming graph type, so that the concepts and operations necessary for information extraction could be set up [38]. To separate out the time required to read questions, a specific pacing was given to each block. First, the question was displayed, along with a button labeled "Proceed to graph framework"; this led participants to the graphical framework with the appropriate title and labels. At the bottom of this was another button labeled "Display data," which displayed the full stimulus.

As mentioned, to give examinees the opportunity to fully comprehended each question, every block began with a "question comprehension" condition in which the data were shown in table form. This was intended to remove potential effects caused by the setup of high-level operations for solving a particular kind of problem.

Finally, to make sure ability (and not capacity) was being tested, an 11s timeout was set for each repetition. This was based on the mean time required to answer the items in our pilot studies.

Response format To respond, examinees were required to click on one of several possible answers, displayed in the form of buttons below the stimulus. In some cases, correct answers were not directly displayed. For example, certain values were not explicitly shown with labeled ticks on the graph's axes. This was done to test examinees' ability to make confident estimations (*i. e.,* to handle *uncertainty* [34]). In addition, although the stimuli used color coding to show different planets, the response buttons did not. This forced examinees to translate the answer found in the visual domain back into the data domain.

4.1.2 Setup

To calibrate our test, we administered it on MTurk. While the validity of using this platform may be debated, due to lack of control over particular experimental conditions [19], we considered it best to perform our calibration using the results of a wide variety of people.

Participants To our knowledge, no particular number of samples is recommended for IRT modeling. We recruited 40 participants who were required to have a 98% acceptance rate and a total of 1000 or more HITS approved.

Coding Six Turkers spent less than 1.5s on average reading and answering questions; they were considered as random clickers, and their results were removed from further analysis. All retained Turkers were native English speakers.

The remaining data were sorted according to item and repetition ID (assigned before randomization). Responses for the table conditions were removed. A score dataset (LG1*s*) was then created in accord with the requirements of IRT modeling: correct answers were scored 1 and incorrect answers 0. Scores for each set of item repetitions were then *compressed* by computing the rounded mean values. This resulted in a set of twelve dichotomous item scores for each examinee.

4.1.3 Results

The purpose of this calibration is to remove items that are unhelpful for distinguishing between low and high levels of VL. To do so, we need to: 1) check that the simplest variant of IRT models (*i. e.,* the Rasch model) fits the data, 2) find the best variant of the model to get the most accurate characteristic values for each item, and 3) assess the usefulness of each item.

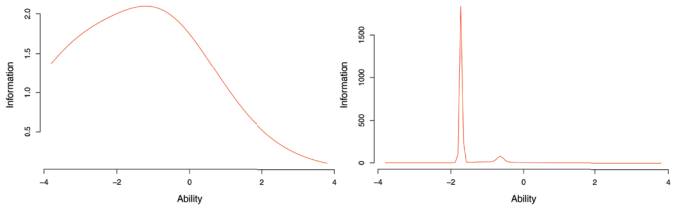
Checking the Rasch model The Rasch model (RM) was first fitted to the score dataset. A 200 sample parametric Bootstrap goodness-of-fit test using Pearson's χ^2 statistic revealed a non-significant p-value for LG1*s* ($p > 0.54$), suggesting an acceptable fit[4]. The Test Information Curve (TIC) is shown in Fig. 1a. It reveals a near-normal distribution of test information across different ability levels, with a slight bump around -2, and a peak around -1. This means that the test provides more information about examinees with relatively low abilities (0 being the ability level of an average achiever) than about examinees with high abilities.

Finding the right model variant Different IRT models, implemented in the **ltm** R package [40], were then fitted to LG1*s*. A series of pairwise likelihood ratio tests showed that the two-parameter logistic model (2PL) was most suitable. The new TIC is shown in Fig. 1b.

Assessing the usefulness of test items The big spike in the TIC (Fig. 1b) suggests that several items with difficulty characteristics just above -2 have high discrimination values. This is confirmed by the very steep Item Characteristic Curves (ICCs) (Fig. 3a) for items LG1.1, LG1.4, and LG1.9 ($\alpha > 51$), and can explain the slight distortion in Fig. 1a.

The probability estimates revealed that examinees with average abilities have a 100% probability of giving a correct answer to the easiest items (LG1.1, LG1.4, and LG1.9), and a 41% probability of giving a correct answer to the hardest item (LG1.11). However, the fact that LG1.11 has a relatively low discrimination value ($\alpha < 0.7$) suggests that it is not very effective for separating ability levels.

[4]For more information about this statistic, refer to [40].



(a) TIC of LG1*s* under RM (b) TIC of LG1*s* under 2PL

Fig. 1: Test Information Curves (TICs) of the score dataset of the first line graph test under the original constrained Rasch model (RM) (a) and the two-parameter logistic model (2PL) (b). The ability scale shows the θ-values. The slight bump in the normal distribution of (a) can be explained by the presence of several highly discriminating items, as shown by the big spike in (b).

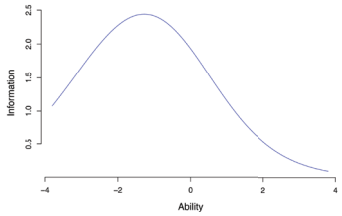
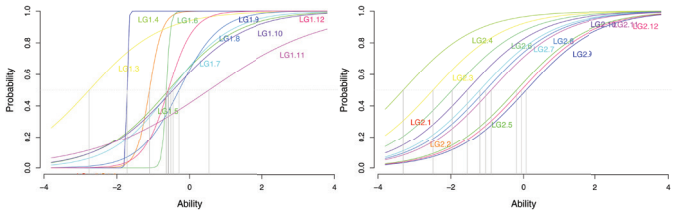


Fig. 2: Test Information Curve of the score dataset of the second line graph test under the original constrained Rasch model. The test information is normally distributed.



(a) ICCs of LG1*s* under 2PL (b) ICCs of LG2*s* under RM

Fig. 3: Item Characteristic Curves (ICCs) of the score datasets of the first line graph test (LG1*s*) under the original the two-parameter logistic model (a), and of the second line graph test (LG2*s*) under the constrained Rasch model (b). The different curve steepnesses in (a) are due to the fact that 2PL computes individual discrimination values for each item, while RM sets all discrimination values to 1.

4.1.4 Discussion

IRT modeling appears to be a solid approach for calibrating our test design. Our results (Fig. 1) show that LG1 is useful for differentiating between examinees with relatively low abilities, but not so much for ones with high abilities.

The slight bump in the distribution of the TIC (Fig. 1a) suggests that several test items are quite effective for separating ability levels around -2. This is confirmed by the spike in Fig. 1b, which indicates the presence of highly discriminating items. Overall, both Test Information Curves reveal that the test is best suited for examinees with relatively low abilities, since most of the information it provides concerns ability levels below zero.

In addition, Fig. 3a reveals that several items in the test have identical difficulty and discrimination characteristics. Some of these could be considered for removal, as they provide only redundant information. Similarly, item LG1.11, which has a low discrimination characteristic, could be dropped, as it is less effective than others.

4.1.5 Line Graphs 2: General Design

For our second line graph test (LG2), we also created twelve items, with varying factors restricted to question congruency and tasks. The test used four samples for the stimuli, and a single layout with single scales. The same scenario, dataset, pacing, and response format as for LG1 were kept, as well as the five repetitions, the question comprehension condition, and the 11s timeout. As such, six items in this test were identical to items in LG1 (see pink cells in Table 1). This was done to ensure that the order of item difficulty would remain consistent across the different tests.

The calibration was again conducted on MTurk. 40 participants were recruited; the work of three Turkers was rejected, for the same reason as before.

4.1.6 Results and Discussion

Our analysis was driven by the requirements listed above. Data were sorted and encoded in the same way as before, and a score dataset for LG2 was obtained (LG2*s*).

RM was fitted to the score dataset, and the goodness-of-fit test revealed an acceptable fit ($p > 0.3$). The pairwise likelihood ratio test showed that RM was the best of all possible IRT models. The Test Information Curve (Fig. 2) is normally distributed, with a peak around -1. This indicates that like our first line graph test, LG2 is best suited for examinees with relatively low abilities.

The Item Characteristic Curves of both tests were then compared. While it cannot be expected that identical items have the exact same characteristics, their difficulty order should remain consistent (see Sect. 2.3). Fig. 3 shows some slight discrepancies for items 1, 3, and 6 between the two tests. However, the fact that item LG1.3 is further to the left in Fig. 3a is misleading. It is due to the extremely high a-values of items LG1.1 and LG1.4. Thus, while their b-values are slightly higher than that of LG1.3, the probability of success of an average achiever is higher for these items than it is for LG1.3 ($1 > 0.94$). Furthermore, the difficulty characteristics of LG1.3 and LG2.3 are very similar ($0.94 \simeq 0.92$). Therefore, the only exception in the ordering of item difficulties is item 6, which is estimated to be more difficult than item 2 in LG1, and not in LG2.

This suggests that LG1 and LG2 cover the same latent trait, *i. e.,* ability to read line graphs. To examine this, we equated the test scores using a *common item equation* approach. RM was fitted to the resulting dataset, the goodness-of-fit test showed an acceptable fit, and 2PL provided best fit. Individual item characteristics were generally preserved, with the exception of item 6, which, interestingly, ended up with characteristics very similar to those of item 2. This confirms that the two tests cover the same latent trait. Thus, although individual characteristics are slightly altered by the equation (*e. g.,* item 6), items in LG1 can safely be transposed to LG2, without hindering the overall coherence of the test, and *vice-versa*.

4.2 Assessment Phase

Having shown that our test items have a sound basis in theory, we now turn to the assessment of visualization literacy. While a standard method would simply sum up the correct responses, our method considers each response individually, with regard to the difficulty of the item it was given for. To make this assessment, we inspected the *ability scores* derived from the fitted IRT models. These scores represent examinees' standings (θ) on the latent trait, and correspond to a unique response pattern. They have great predictive power as they can determine an examinee's probability of success on items that s/he has not completed, provided that these items follow the same latent variable scale as other test items. As such, ability scores are perfect indicators for assessing VL.

LG1 revealed 27 different ability scores, ranging from -1.85 to 1. The distribution of these scores was near-normal, with a slight bump around -1.75. 40.7% of participants were above average (*i. e.,* $\theta > 0$), and the mean was -0.27.

LG2 revealed 33 different ability scores, ranging from -1.83 to 1.19. The distribution was also near-normal, with a bump around -1. 39.4% of participants were above average, and the mean was -0.17.

These results show that the means are close to zero, and the distributions near-normal. This suggests that most Turkers, while somewhat below average in visualization literacy for line graphs, have fairly standard abilities.

While it should be interesting to develop broader ranges of item complexities for the line graph stimulus (by using the common item equation approach), thus extending the psychometric quality of the tests, we consider LG1 and LG2 to be sufficient for our current line of research. Furthermore, we believe that these low levels of difficulty reflect the general simplicity of, and massive exposure to, line graphs.

5 Extensions

To see whether our method also applies to other types of visualizations, we created two additional tests: one for bar charts (BC) and one for scatterplots (SP).

5.1 Design Phase

5.1.1 Bar Charts: General Design

Like LG1 and LG2, the design of our bar chart test (BC) was based on the principles described in Section 3. We created twelve items, with varying factors restricted to number of samples and tasks (see Table 1). The same scenario, pacing, response format, repetitions, question comprehension condition, and 11s timeout were kept. The dataset presented life expectancy in various countries, with country names again changed to fictitious planet names.

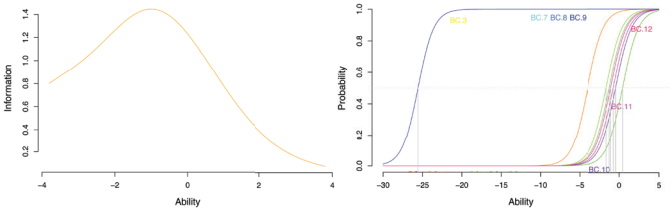
The only difference with the factors used earlier (apart from the stimulus) involved the variation task, which is essentially a trend detection task. Bar charts are sub-optimal for determining trends, so this task was replaced by a "global similarity detection" task, as done in [20] (*e. g.,* "Do all the bars have the same value?").

The calibration was again conducted on MTurk. 40 participants were recruited; the work of six Turkers was rejected, for the same reason as before.

5.1.2 Results and Discussion

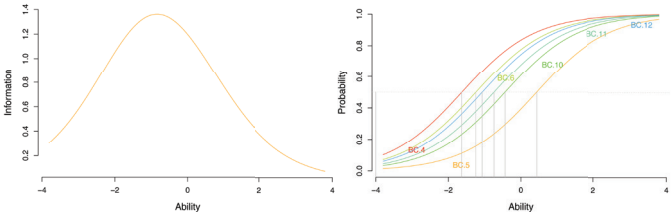
Our analysis was driven by the same requirements as for the line graph tests. Data were sorted and encoded in the same way, resulting in a score dataset for BC (BC*s*).

RM was first fitted to BC*s*; the goodness-of-fit test revealed an acceptable fit ($p > 0.37$), and the likelihood test proved that it fit best. However, the Test Information Curve (Fig. 4a) is not normally distributed. This is due to the presence of several extremely low difficulty (*i. e., easy*) items (BC.3, BC.7, BC.8, and BC.9; $b = -25.6$), as shown in Fig. 4b. Inspecting the raw scores for these items revealed a 100% success rate. Thus, they were considered too easy, and were removed. Similarly, items BC.1 and BC.2 (for both, $b < -4$) were also removed.



(a) TIC of BC*s* under RM (b) ICCs of BC*s* under RM

Fig. 4: Test Information Curve (a) and Item Characteristic Curves (b) of the score dataset of the bar chart test under the constrained Rasch model. The TIC in (a) is not normally distributed because of several very low difficulty items, as shown by the curves to the far left of (b).



(a) TIC of the subset of BC*s* under RM (b) ICCs of the subset of BC*s* under RM

Fig. 5: Test Information Curve (a) and Item Characteristic Curves (b) of the subset of the score dataset of the bar chart test under the constrained Rasch model. The subset was obtained by removing the very low difficulty items shown in Fig. 4b.

To check the coherence of the resulting subset of items, RM was fitted again to the remaining set of scores. Goodness-of-fit was maintained ($p > 0.33$), and RM still fitted best. The new TIC (Fig. 5a) is normally distributed, with a peak around -1. This indicates that like our line graph tests, this subset of BC is best suited for examinees with relatively low abilities.

5.1.3 Scatterplots: General Design

For our scatterplot test (SP), we once again created twelve items, with varying factors restricted to distraction and tasks (see Table 1). The same scenario, pacing, response format, repetitions, and question comprehension condition were kept. The dataset presented levels of adult literacy by expenditure per student in primary school in different countries, with country names again changed to fictitious planet names.

Slight changes were required for some of the tasks, since scatterplots use two spatial dimensions (as opposed to bar charts and line graphs). For example, stimuli with distractors in LG1 only required examinees to focus on one of several samples; here, stimuli with distractors could either require examinees to focus on a single datapoint or on a single dimension.

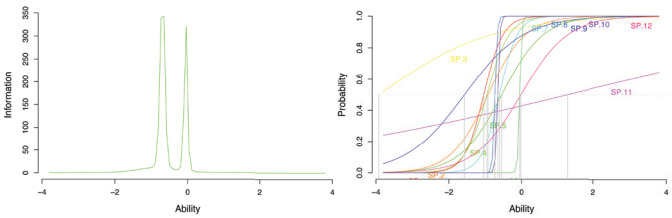
We had initially expected that SP would be more difficult, and items would require more time to complete. However, a pilot study showed that the average response time per item was again roughly 11s. Therefore, the 11s timeout condition was kept.

The calibration was again conducted on MTurk. 40 participants were recruited; the work of one Turker was not kept because of technical (logging) issues.

5.1.4 Results and Discussion

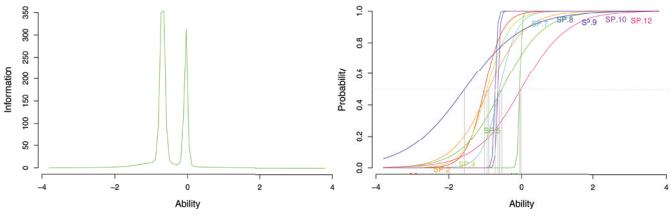
Our analysis was once again driven by the same requirements as before. The same sorting and coding was applied to the data, resulting in the score dataset SP*s*. The fitting procedure was then applied, revealing a good fit for RM ($p = 0.6$), and a best fit for 2PL.

The Test Information Curve (Fig. 6a) shows the presence of several highly discriminating items around $b \simeq -1$ and $b \simeq 0$. The Item Characteristic Curves (Fig. 6b) confirm that there are three (SP.6, SP.8, and SP.10; $a > 31$). However, they also show that two items (SP.3, and



(a) TIC of SP*s* under 2PL (b) ICCs of SP*s* under 2PL

Fig. 6: Test Information Curve (a) and Item Characteristic Curves (b) of the score dataset of the scatterplot test under the two-parameter logistic model. The TIC (a) shows that there are several highly discriminating items, which is confirmed by the very steep curves in (b). In addition, (b) shows that there are also a two poorly discriminating items, represented by the very gradual slopes of items SP.3 and SP.11.



(a) TIC of the subset of SP*s* under 2PL (b) ICCs of the subset of SP*s* under 2PL

Fig. 7: Test Information Curve (a) and Item Characteristic Curves (b) of the subset of the score dataset of the scatterplot test under the two-parameter logistic model. The subset was obtained by removing the poorly discriminating items shown in Fig. 6b.

SP.11) have quite low discrimination values ($a < 0.6$). Here, we set a threshold for $a > 0.8$. Thus, items SP.3 and SP.11 were removed. The resulting subset of 10 items' scores was fitted once again. RM fitted well ($p = 0.69$), and 2PL fitted best. The different curves of the subset are plotted in Fig. 7. They show a good amount of information for abilities that are slightly below average (Fig. 7a), which indicates that the subset of SP is once again best suited for examinees with relatively low abilities.

5.2 Assessment phase

Here again, we inspected the Turkers' ability scores. Only the items retained at the end of the design phase were used.

BC revealed 21 different ability scores, ranging from -1.75 to 0.99. The distribution of these scores was near-normal, with a slight bump around -1.5, and the mean was -0.39. However, only 14.3% of participants were above average.

SP revealed 23 different ability scores, ranging from -1.72 to 0.72. However, the distribution here was not normal. 43.5% of participants were above average, and the median was -0.14.

These results show that the majority of recruited Turkers had somewhat below average levels of visualization literacy for bar charts and scatterplots. The very low percentage of Turkers above average in BC led us to reconsider the removal of items BC.1 and BC.2, as they were not truly problematic. After reintegrating them in the test scores, 21 ability scores were observed, ranging from -1.67 to 0.99, and 42.8% of participants were above average. This seemed more convincing. However, this important difference illustrates the relativity of these values, and shows how important it is to properly calibrate the tests during the design phase.

Finally, we did not attempt to equate these tests, since—unlike LG1 and LG2—we ran them independently without any overlapping items. To have a fully comprehensive test, *i. e.,* a generic test for visualization literacy, intermediate tests are required where the stimulus itself is a varying factor. If such tests prove to be coherent (*i. e.,* if IRT models fit the results), then it should be possible to assert that VL is a general trait that allows one to understand any kind of graphical representation. Although we believe that this ability varies with exposure and habit of use, a study to confirm it is outside of the scope of this paper.

6 Fast, Effective Testing

If these tests are to be used as practical ways of assessing VL, the process must be sped up, both in the administration of the tests and in the analysis of the results. While IRT provides useful information on the quality of tests and on the ability of those who take them, it is quite costly, both in time and in computation. This must be changed.

In this section, we present a way in which the tests we have developed in the previous sections can be optimized to be faster, while still maintaining their effectiveness.

6.1 Test Administration Time

As we have seen, several items can be removed from the tests, while keeping good psychometric quality. However, this should be done carefully, as some of these items may provide useful information (like in the case of BC.1 and BC.2, Sect. 5.2).

We first removed LG1.11 from LG1, as its discrimination value was < 0.8 (see Sect. 5.1.4). We then inspected items with identical difficulty and discrimination characteristics, represented by overlapping ICCs' (see Fig. 3). These were prime candidates for removal, since they provide only redundant information. There was one group of overlapping items in LG1 ([LG1.1, LG1.4, LG1.9]), and two in LG2 ([LG2.1, LG2.3], [LG2.2, LG2.7]). For each group, we kept only one item. Thus LG1.1, LG1.4, LG2.3, and LG2.7 were dropped.

We reintegrated items BC.1 and BC.2 to BC, as they proved to have a big impact on ability scores (Sect. 5.2). The subset of SP created at the end of the design phase was kept, and no extra items were removed.

RM was fitted to the newly created subsets of LG1, LG2, and BC; the goodness-of-fit test showed acceptable fits for all ($p > 0.69$ for LG1, and $p > 0.3$ for both LG2 and BC). 2PL fitted best for LG1, and RM fitted best for both LG2 and BC.

We conducted *post-hoc* analysis to see whether the number of item repetitions could be reduced (first to three, then to one). Results showed that RM fitted all score datasets using three repetitions. However, several examinees had lower scores. In addition, while BC and SP showed similar amounts of information for the same ability levels, the three very easy items in BC (*i. e.,* BC.3, BC.7, BC.8, and BC.9) were no longer problematic. This suggests that several participants did not get a score of 1 for these items, and confirms that, for some examinees, more repetitions are needed. Results for one-repetition-tests showed that RM no longer fitted the scores of BC, suggesting that unique repetitions are noisy. Therefore, we decided to keep the five repetitions.

In the end, the redesign of LG1 contained 9 items (with a $\simeq 10$ min completion time), the redesigns of LG2 and SP contained 10 items (11 min), and the redesign of BC contained 8 items ($\simeq 9$ min).

6.2 Analysis Time and Computational Power

To speed up the analysis, we first considered setting up the procedure we had used in R on a server. However, this solution would have required a lot of computational power, so we dropped it.

Instead, we chose to tabulate all possible ability scores for each test. An interesting feature of IRT modeling is that it can derive ability scores from unobserved response patterns (*i. e.,* patterns that do not exist in the empirical data), as well as from partial response patterns (*i. e.,* patterns with missing values). Consequently, we generated all the $2^{n_i} - 1$ possible patterns for each test, where n_i is the number of items in a test. This resulted in 511 patterns for LG1, 1023 for both LG2 and SP, and 255 for BC. We then derived the different ability scores that could be obtained in each test.

To ensure that removing certain test items did not greatly affect the ability scores, we computed all the scores for the full LG1 and LG2 tests, and compared them to the ones previously obtained. We found some small differences in the upper and lower bounds of ability, but these were considered negligible, since our tests were not designed for fine distinction between very low abilities or high abilities. We

also tested the impact of refitting the IRT models after item removal. For this, we repeated the procedure using partial response patterns for LG1 and LG2, *i. e.,* we replaced the dichotomous response values for the items considered for removal by *not available* (NA) values. The scores were exactly the same as the ones obtained with our already shortened and refitted tests, which proves they can be trusted.

Finally, we integrated all ability scores and their corresponding response patterns into the web-based, shortened versions of the tests, to make them readily available. This way, by administering our online tests, researchers can have direct access to participants' levels of visualization literacy. Informed decisions can then be made as whether to keep these people for further studies or not. All four tests are accessible at `http://peopleviz.gforge.inria.fr/trunk/vLiteracy/home/`

7 Methodology Guidelines

As the preceding sections have shown, we have developed and validated a fast and effective method for assessing visualization literacy. This section summarizes the major steps, written in the form of easy "take-away" guidelines.

7.1 Initial Design

1. **Pay careful attention to the design of all 3 components of a test item**, *i. e.,* stimulus, task, and question. Each can influence item difficulty, and too much variation may fail to produce a coherent test—as was seen in our pilot studies.
2. **Repeat each item several times**. We did 5 repetitions + 1 "question comprehension" condition for each item. This is important as repeated trials provide more robust measures. Ultimately, it may be feasible to reduce the number of repetitions to 3[5], although our results show that this can be problematic (Sect. 6.1).
3. **Use a different—and ideally, non-graphical—representation for question comprehension**. We chose a table condition. While our present study did not focus on proving its essentialness, we believe that this attribute is important.
4. **Randomize the order of items and of repetitions**. This is common practice in experiment design, having the benefit of preventing carryover effects.
5. Once the results are in, **sort the data according to item and repetition ID, remove the data for the question comprehension condition, and encode examinees' scores in a dichotomous way**, *i. e.,* 1 for correct answers and 0 for incorrect answers.
6. **Calculate the mean score for all repetitions of an item and round the result**. This will give a finer estimate of the examinee's ability since it erases one-time errors which may be due to lack of attention or to clicking on the wrong answer by mistake.
7. **Begin model fitting with the Rasch model.** RM is the simplest variant of IRT models. If it does not fit the data, other variants will not either. Then **Check the fit of the model**. Here we used a 200 sample parametric Bootstrap goodness-of-fit test using Pearsons χ^2 statistic. To reveal an acceptable fit, the returned p-value should not be statistically significant ($p > 0.05$). In some cases (like in our pilot studies), the model may not fit. Options here are to inspect the χ^2p-values for pairwise associations, or the two- and three-way χ^2 residuals, to find problematic items[6].
8. **Determine which IRT model variant best fits the data.** A series of pairwise likelihood ratio tests can be used for this. If several models fit, it is usually good to go with the model that fits best. Our experience showed that such models were most often RM and 2PL.
9. **Identify potentially useless items.** In our examples of LG1 and SP, certain items had low discrimination characteristics. These are not very effective for separating ability levels, and can be removed. In cases like the one for BC, items may also simply be too easy. Before removing them permanently, however, it is advised to check their impact on ability scores. Finally, it is important the model be refitted at this stage (reproducing steps 7 and 8), as removing these items may affect examinee and item characteristics.

7.2 Final Design

10. **Identify overlapping items and remove them.** If the goal is to design a short test, such items can safely be removed, as they provide only redundant information (see Sect. 6.1).
11. **Generate all $2^{n_i} - 1$ possible score patterns**, where n_i is the number of retained items in the test. These patterns represent series of dichotomous response values for each test item.
12. **Derive the ability scores from the model**, using the patterns of responses generated in step 11. These scores represent the range of visualization literacy levels that the test can assess.
13. **Integrate the ability scores into the test** to make fast, effective, and scalable estimates of people's visualization literacy.

8 Conclusions and Future Work

In this paper, we have developed a method for assessing visualization literacy, based on a principled set of considerations. In particular, we used Item Response Theory to allow a separation of the effects of item difficulty and examinee ability. Our motivation was to make a series of fast, effective, and reliable tests which researchers could use to detect participants with low VL abilities before conducting online studies. We have shown how these tests can be tailored to get immediate estimates of examinee's levels of VL.

We intend to continue developing this approach, as well as examine the suitability of this method for other kinds of representation (*e. g.,* parallel coordinates, node link diagrams, starplots, etc.), and possibly for other purposes. For example, in contexts like a classroom evaluation, the tests could be longer, and broader assessments of visualization literacy could be made. This would imply further exploration of the design parameters proposed in Section 3. Evaluating the impact of these parameters on item difficulty should also be interesting.

Finally, we acknowledge that this work is but a small step into the realm of visualization literacy. As such, we have made our tests available on GitHub for versioning [1]. Ultimately, we hope that this will serve as a foundation for further research into VL.

9 Acknowledgements

This work was funded by a Google Research Award, granted for a project called "Data Visualization for the People".

References

[1] `https://github.com/INRIA/Visualization-Literacy-101`.
[2] D. Abilock. Visual information literacy: Reading a documentary photograph. *Knowledge Quest*, 36(3), January–February 2008.
[3] J. Baer, M. Kutner, and J. Sabatini. Basic reading skills and the literacy of the america's least literate adults: Results from the 2003 national assessment of adult literacy (naal) supplemental studies. Technical report, National Center for Education Statistics (NCES), February 2009.
[4] J. Bertin and M. Barbut. *Semiologie Graphique*. Mouton, 1973.
[5] P. BOBKO and R. KARREN. The perception of pearson product moment correlations from bivariate scatterplots. *Personnel Psychology*, 32(2):313–325, 1979.
[6] M. T. Brannick. Item response theory. `http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm`.
[7] V. J. Bristor and S. V. Drake. Linking the language arts and content areas through visual technology. *T.H.E. Journal*, 22(2):74–77, 1994.
[8] P. Carpenter and P. Shah. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, 4(2):75–100, 1998.
[9] W. S. Cleveland, P. Diaconis, and R. McGill. Variables on scatterplots look more highly correlated when the scales are increased. *Science*, 216(4550):1138–1141, 1982.
[10] Cognitive testing interview guide. `http://www.cdc.gov/nchs/data/washington_group/meeting5/WG5_Appendix4.pdf`.

[5]The number of repetitions should be odd, so as to not end up with a mean score of 0.5 for an item.

[6]For more information, refer to [40].

[11] M. Correll, D. Albers, S. Franconeri, and M. Gleicher. Comparing averages in time series data. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 1095–1104. ACM, 2012.

[12] F. R. Curcio. Comprehension of mathematical relationships expressed in graphs. *Journal for research in mathematics education*, pages 382–393, 1987.

[13] Department for Education. Numeracy skills tests: Bar charts. http://www.education.gov.uk/schools/careers/traininganddevelopment/professional/b00211213/numeracy/areas/barcharts, 2012.

[14] R. C. Fraley, N. G. Waller, and K. A. Brennan. An item response theory analysis of self-report measures of adult attachment. *Journal of personality and social psychology*, 78(2):350, 2000.

[15] E. Freedman and P. Shah. Toward a model of knowledge-based graph comprehension. In M. Hegarty, B. Meyer, and N. Narayanan, editors, *Diagrammatic Representation and Inference*, volume 2317 of *Lecture Notes in Computer Science*, pages 18–30. Springer Berlin Heidelberg, 2002.

[16] 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.

[17] A. C. Graesser, S. S. Swamer, W. B. Baggett, and M. A. Sell. New models of deep comprehension. *Models of understanding text*, pages 1–32, 1996.

[18] Graph design i.q. test. http://perceptualedge.com/files/GraphDesignIQ.html.

[19] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 203–212, New York, NY, USA, 2010. ACM.

[20] How to read a bar chart. http://www.quizrevolution.com/act101820/mini/go/.

[21] P. Isenberg, A. Bezerianos, P. Dragicevic, and J. Fekete. A study on dual-scale data charts. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2469–2478, Dec 2011.

[22] Join, ACRL and Council, Chapters. Presidential committee on information literacy: Final report. Online publication, 1989.

[23] I. Kirsch. The international adult literacy survey (ials): Understanding what was measured. Technical report, Educational Testing Service, December 2001.

[24] N. Knutson, K. S. Akers, and K. D. Bradley. Applying the Rasch model to measure first-year students perceptions of college academic readiness. In *Paper presented at the annual meeting of the MWERA Annual Meeting*, 2010.

[25] R. Lowe. "Reading" scientific diagrams: Characterising components of skilled performance. *Research in Science Education*, 18(1):112–122, 1988.

[26] R. Lowe. *Scientific Diagrams: How Well Can Students Read Them? [microform]: What Research Says to the Science and Mathematics Teacher. Number 3 / Richard K. Lowe*. Distributed by ERIC Clearinghouse [Washington, D.C.], 1989.

[27] J. Meyer, M. Taieb, and I. Flascher. Correlation estimates as perceptual judgments. *Journal of Experimental Psychology: Applied*, 3(1):3, 1997.

[28] E. Miller. The Miller word-identification assessment. http://www.donpotter.net/pdf/mwia.pdf, 1991.

[29] National Cancer Institute. Item response theory modeling. http://appliedresearch.cancer.gov/areas/cognitive/item.html.

[30] National Center for Education Statistics. Adult literacy and lifeskills survey. http://nces.ed.gov/surveys/all/.

[31] Numerical reasoning - table/graph. http://www.jobtestprep.co.uk/jobtestprep/testPlayer.aspx?localeSetting=en-GB&sLanguage=en-GB&&test_id=free_num_advanced&isSequence=0&testIndex=0&mode=timed&skinID=9&template=FreeTestTemplate&endURL=signout.aspx%3fReturnURL%3d..%2fnumerical-reasoning-test&endURLName=Back+to+JobTestPrep&cancelURLName=Exit+Test&cancelURL=signout.aspx%3fReturnURL%3d..%2fnumerical-reasoning-test&testTitle=Numerical+Reasoning+-+Table%2fGraph&layoutype=1&adPageID=&itemCode=FreeNum_GMAT.

[32] Numerical reasoning online. http://numericalreasoningtest.org/.

[33] J. Oberholtzer. Why two charts make me feel like an expert on Portugal . http://tiny.cc/ujgjdx, 2012.

[34] OECD. Pisa 2012 assessment and analytical framework. Technical report, OECD, 2012.

[35] OTA. *Computerized Manufacturing Automation: Employment, Education, and the Workplace*. United States Office of technology Assessment, 1984.

[36] S. Pinker. *A theory of graph comprehension*, pages 73–126. Lawrence Erlbaum Associates, Hillsdale, NJ, 1990.

[37] I. Pollack. Identification of visual correlational scatterplots. *Journal of experimental psychology*, 59(6):351, 1960.

[38] R. Ratwani and J. Gregory Trafton. Shedding light on the graph schema: Perceptual features versus invariant structure. *Psychonomic Bulletin and Review*, 15(4):757–762, 2008.

[39] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Comput. Graph. Forum*, 29(3):1203–1210, 2010.

[40] D. Rizopoulos. ltm: An R package for latent variable modeling and Item Response Analysis. *Journal of Statistical Software*, 17(5):1–25, 11 2006.

[41] RUMMLaboratory. Rasch analysis. http://www.rasch-analysis.com/rasch-models.htm.

[42] P. Shah. A model of the cognitive and perceptual processes in graphical display comprehension. *Reasoning with diagrammatic representations*, pages 94–101, 1997.

[43] G. Sir Crowther. *The Crowther Report*, volume 1. Her Majesty's Stationery Office, 1959.

[44] C. Taylor. New kinds of literacy, and the world of visual information. *Literacy*, 2003.

[45] J. G. Trafton, S. P. Marshall, F. Mintz, and S. B. Trickett. Extracting explicit and implicit information from complex visualizations. *Diagrams*, pages 206–220, 2002.

[46] S. Trickett and J. Trafton. Toward a comprehensive model of graph comprehension: Making the case for spatial cognition. In D. Barker-Plummer, R. Cox, and N. Swoboda, editors, *Diagrammatic Representation and Inference*, volume 4045 of *Lecture Notes in Computer Science*, pages 286–300. Springer Berlin Heidelberg, 2006.

[47] B. Tversky. *Semantics, Syntax, and Pragmatics of graphics.*, pages 141–158. Lund University Press, Lund, 2004.

[48] UNESCO. Literacy assessment and monitoring programme. http://www.uis.unesco.org/Literacy/Pages/lamp-literacy-assessment.aspx.

[49] University of Kent. Numerical reasoning test. http://www.kent.ac.uk/careers/tests/mathstest.htm.

[50] H. Wainer. A test of graphicacy in children. *Applied Psychological Measurement*, 4(3):331–340, 1980.

[51] What is media literacy? a definition... and more. http://www.medialit.org/reading-room/what-media-literacy-definitionand-more.

[52] M. Wu, R. Adams, and E. M. Solutions. *Applying the Rasch model to psycho-social measurement [electronic resource] : A practical approach / Margaret Wu and Ray Adams* . Educational Measurement Solutions Melbourne, Vic, 2007.