Measuring and predicting variation in the difficulty of questions about data visualizations

Arnav Verma (arnavv@stanford.edu)
Department of Psychology, Stanford University
Stanford, CA, United States

Judith E. Fan (jefan@stanford.edu)
Department of Psychology, Stanford University
Stanford, CA, United States

Abstract

Understanding what is communicated by data visualizations is a critical component of scientific literacy in the modern era. However, it remains unclear why some tasks involving data visualizations are more difficult than others. Here we administered a composite test composed of five widely used tests of data visualization literacy to a large sample of U.S. adults ($N = 503$ participants). We found that items in the composite test spanned the full range of possible difficulty levels, and that our estimates of item-level difficulty were highly reliable. However, the type of data visualization shown and the type of task involved only explained a modest amount of variation in performance across items, relative to the reliability of the estimates we obtained. These results highlight the need for finer-grained ways of characterizing these items that predict the reliable variation in difficulty measured in this study, and that generalize to other tests of data visualization understanding.

Keywords: data visualization literacy; graph comprehension; statistical literacy; quantitative reasoning; STEM education; psychometric evaluation

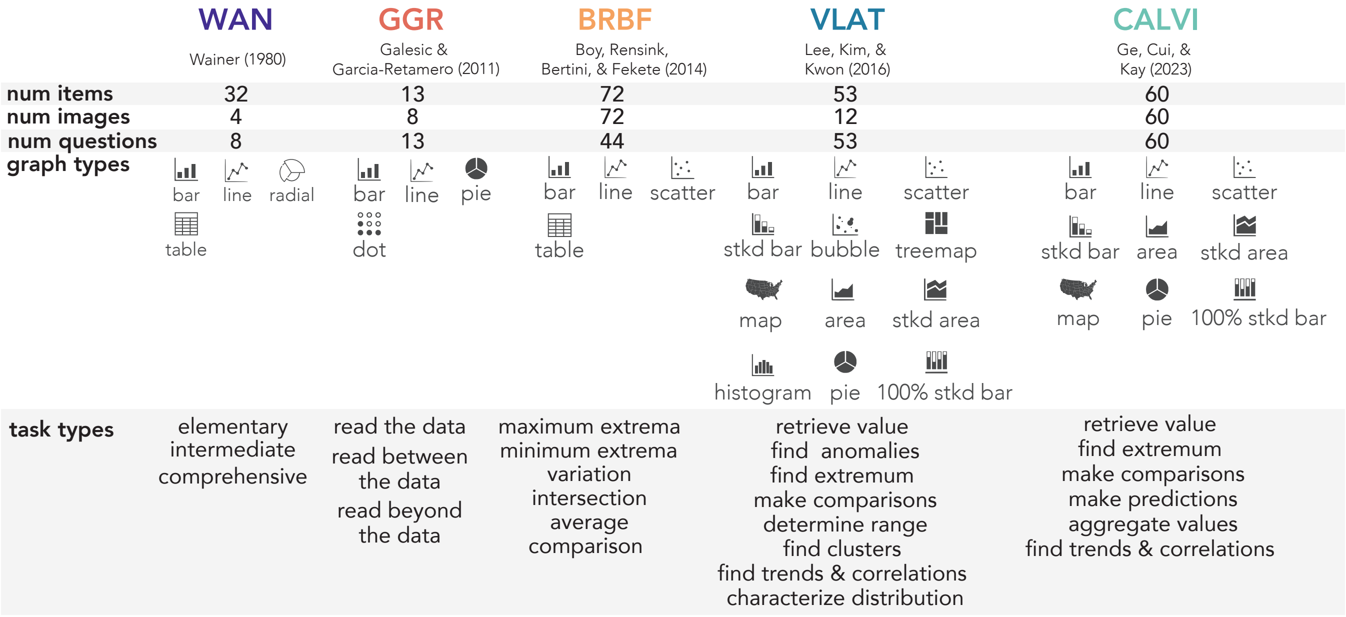
Introduction

Data visualizations are powerful tools invented by humans for making sense of a complex world. Although they have only existed for a few centuries, they are practically indispensable in modern scientific workflows (Munzner, 2014) and are pervasive in social media and the news (C. Lee et al., 2021). Data visualizations (or equivalently, *graphs*, *charts*, or *plots*) enable people to reason about quantitative information through visual encodings of data (Keim et al., 2008; Munzner, 2014). A single visualization can even serve multiple purposes. For example, a scatter plot can help with finding outliers in the data while also assisting in deriving broader insights about complex trends (Boy et al., 2014; S. Lee et al., 2016; Börner et al., 2019; Lundgard & Satyanarayan, 2021).

Performing these tasks relies on the coordination of many cognitive processes, including rapid visual computations (Cleveland & McGill, 1984; Ciccione et al., 2023; Cui & Liu, 2021), invocation of the appropriate graph schema (Pinker, 2014), mathematical operations (Gillan & Lewis, 1994), control of finite attentional and working memory resources (Padilla et al., 2018), and other reasoning processes to derive general insights informed by prior knowledge (Carpenter & Shah, 1998; Shah & Freedman, 2011). However, people do not automatically acquire the ability to reason about data visualizations; rather, this ability is acquired gradually, and usually in formal educational contexts (Alper et al., 2017; Mix & Cheng, 2012; Wainer, 1980). However, the effectiveness of educational interventions for helping people develop core data visualization literacy skills remains unclear.

This lack of clarity reflects, in part, the lack of a coherent suite of reliable and valid tools for measuring data visualization literacy. Several test-based measures currently exist, each of which generally consist of a series of items, with each consisting of a question paired with a data visualization (Delmas et al., 2005; Galesic & Garcia-Retamero, 2011; Boy et al., 2014; S. Lee et al., 2016; Pandey & Ottley, 2023; Ge et al., 2023; Maltese et al., 2015). However, because they have not been directly compared, the extent to which they reliably measure the same underlying construct and whether they imply a consistent decomposition of data visualization literacy into distinct components remains unknown (Brockbank et al., 2025; Börner et al., 2019; Brehmer & Munzner, 2013; Friel et al., 2001). Some tests contain items meant to measure a compact hierarchy of abstract abilities — e.g., progressing from “reading the data” to “reading beyond the data” (Galesic & Garcia-Retamero, 2011; Wainer, 1980) — while other tests are designed to assess performance on a broader suite of more concrete tasks, such as finding extreme values or detecting correlations (S. Lee et al., 2016; Ge et al., 2023; Pandey & Ottley, 2023; Boy et al., 2014). Additionally, there are tests which also focus on measuring the ability to overcome misleadingly constructed data visualizations, such as ones using inappropriate axis limits (Ge et al., 2023). Here we leverage the diversity of the tasks and data visualizations represented across several existing tests to develop consistent procedures for measuring and comparing the difficulty of tasks involving data visualizations.

In this study, we aggregated 230 items from five tests of data visualization literacy: a 32-item assessment developed by Wainer (1980), which we refer to as **WAN**; a 13-item assessment developed by Galesic & Garcia-Retamero (2011), which we refer to as **GGR**; a 72-item assessment developed by Boy et al. (2014), which we refer to as **BRBF**; a 53-item Visualization Literacy Assessment Test (**VLAT**; S. Lee et al. (2016)); and a 60-item assessment known as **CALVI** (Ge et al., 2023). Together, these assessments represent some of the most widely used and influential tools for measuring data visualization literacy in several research communities, including computer science, education, and psychology.



	WAN Wainer (1980)	GGR Galesic & Garcia-Retamero (2011)	BRBF Boy, Rensink, Bertini, & Fekete (2014)	VLAT Lee, Kim, & Kwon (2016)	CALVI Ge, Cui, & Kay (2023)
num items	32	13	72	53	60
num images	4	8	72	12	60
num questions	8	13	44	53	60
graph types	bar line radial table	bar line pie dot	bar line scatter table	bar line scatter stkd bar bubble treemap map area stkd area histogram pie 100% stkd bar	bar line scatter stkd bar area stkd area map pie 100% stkd bar
task types	elementary intermediate comprehensive	read the data read between the data read beyond the data	maximum extrema minimum extrema variation intersection average comparison	retrieve value find anomalies find extremum make comparisons determine range find clusters find trends & correlations characterize distribution	retrieve value find extremum make comparisons make predictions aggregate values find trends & correlations

Figure 1: We used 230 items from five popular tests of data visualization literacy, which vary in *graph type* and *task type*.

Method

Participants

We used Prolific to recruit 503 U.S.-based participants who spoke English as their primary language and have maintained an approval rate of at least 95%. We maximized the number of participants recruited with the resources available, allowing us to obtain at least 80 responses for each item in the stimuli set.

All participants were given up to three opportunities to complete the warmup trials that included items from the National Assessment of Educational Progress intended to assess middle-school level quantitative literacy skills. Participants who did not pass after three attempts on these tasks did not proceed to the main experiment.

In total, there were 37 participants who failed to complete the warm up trials and additionally another 40 participants who quit the study before completing at least 50% of all trials (23 items). These 77 participants were all omitted from our analysis, leading to a total of 426 participants being used in our analysis and a range of 80 to 92 responses per item.

Materials

We aggregated 230 multiple-choice items from five widely used assessments of data visualization understanding (Figure 1): WAN (Wainer, 1980), GGR (Galesic & Garcia-Retamero, 2011), BRBF (Boy et al., 2014), VLAT (S. Lee et al., 2016), CALVI (Ge et al., 2023).

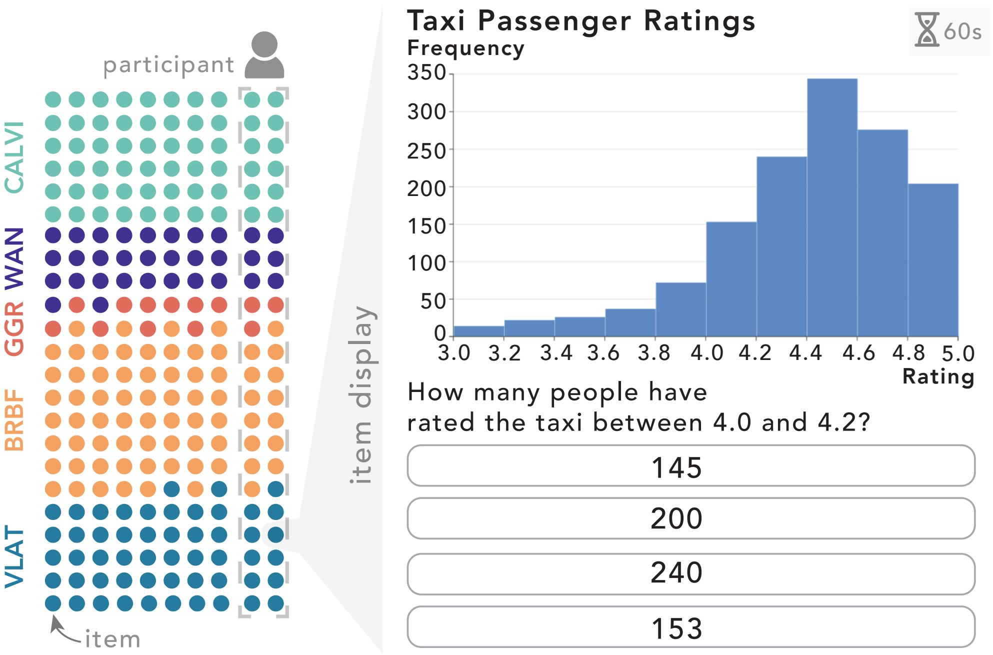
All assessments categorized items using at least two common features: by *task type*, which refers to the reasoning steps a participant performs to answer a question, and by *graph type*, which describes how the image encodes data into visual features.

Since different assessments sometimes use different labels for similar tasks, we additionally defined a simpler common set of *task types* that could apply to all assessments: `value identification`, where participants need to retrieve an individual value appearing in a plot (e.g., finding the maximum value); `arithmetic computation`, where participants are expected to perform arithmetic operations over multiple values displayed in the plot (e.g., finding the average of two values); and `statistical inference`, where participants are required to estimate latent parameters based on the values shown (e.g., judge the strength of trends or presence of clusters).

To explore the potential impact of presenting information in a data visualization as opposed to a table, we also included a small number of table-based items that were otherwise equivalent to the visualization-based ones.

WAN The test by Wainer (1980) was developed to evaluate children at the third- to fifth-grade level in the United States and includes 32 items. It uses six questions which are paired across one table and three images with different *graph types*: `line chart`, `bar chart`, and `radial plot`.

GGR The test developed by Galesic & Garcia-Retamero (2011) is a widely used 13-item assessment comprising of three `bar plots`, three `line plots`, an `icon array`, and a `pie chart`. It was initially designed to explore a compact hierarchy of abstract abilities, progressing from “reading the data” to “reading between the data” and finally, “reading beyond the data.” Originally, nine of the test items required a numerical response, and four were multiple choice. However, to maintain consistency with other test items, we mapped items requiring a numerical response into a multiple-choice



How many people have rated the taxi between 4.0 and 4.2?

145

200

240

153

Figure 2: Our item sampling procedure selected 46 items from the total set of 230 items for each participant (left). Everyone was presented with multiple-choice items from all six tests, with a 60 second time limit to answer each question (right).

format by selecting the top four most frequent responses based on a prior study (Verma et al., 2024).

BRBF The test by Boy et al. (2014) measures the influence of different data and visual properties used across three different *graph types*. It originally consisted of 60 bar charts, 60 scatter plots, 120 line graphs, and 48 tables. All items were initially categorized by six *task types*: maximum extrema, minimum extrema, intersection, variation, average, and comparison. Each combination of *task type* and *graph type* included at least two unique questions, five images of charts, and one image of a table. We create a subset of 72 items representing each unique question, *task type*, and *graph type* combination used in the assessment to reduce the total number of items while maintaining item diversity across different categories.

VLAT The Visualization Literacy Assessment Test by S. Lee et al. (2016) is an influential 53-item assessment containing 12 chart images generated from real-world data sources with a unique image of a line chart, a bar chart, a stacked bar chart, a normalized stacked bar chart, a pie chart, a histogram, a scatter plot, a bubble chart, an area chart, a stacked area chart, a choropleth map, and a tree map. To maintain consistency with other assessments items, we re-classify the bubble chart to a scatter plot. The test originally grouped questions into eight *task types*: retrieving values from a graph, finding correlations & trends, finding anomalies, finding extrema, making comparisons between values, characterizing distributions, determining the range of values in a graph, and finding clusters of common values.

CALVI The Critical Thinking Assessment for Literacy in Visualizations by Ge et al. (2023) is a 60-item test that contains 45 items intended to mislead users with unconventional graphs and questions, alongside 15 standard items, following the design guidelines of VLAT. This includes a subset of *graph types*: line chart, bar chart, stacked bar chart, normalized stacked bar chart, pie chart, scatter plot, area chart, stacked area chart, and choropleth map; and a subset of original *task types*: retrieve value, find trends & correlations, find extremum, make comparisons, make predictions, aggregate values.

Procedure

We evaluated all participants on a representative subset of items across all five tests (Figure 2). Specifically, each participant completed 46 items (20% of the total stimulus set) containing items sampled evenly from all five tests, *graph types*, and *task types*. To ensure that all participants already possessed basic quantitative literacy skills, each session began with five questions taken from the version of the National Assessment of Educational Progress (NAEP) assessment administered to fourth-grade students. Items from the same assessment were presented within the same block. Participants were given a maximum of 60 seconds to answer each question, and provided with immediate feedback indicating whether each response was correct.

Results

To what degree do items vary in difficulty? To determine if there was reliable variation in average performance

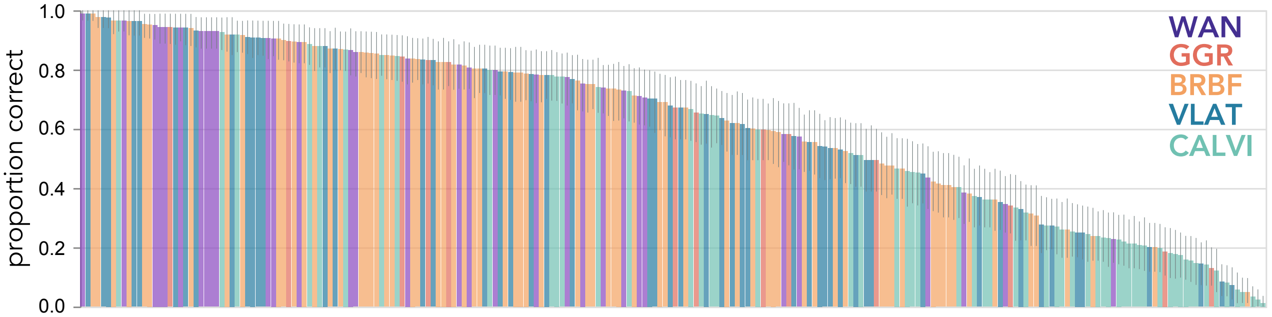


Figure 3: Average performance across all items. Items belonging to the same *test* share the same color. Error bars represent bootstrapped 95% confidence intervals.

within our selected test items, we assessed the degree to which the full set of items, pooled across assessments, spanned a wide range of difficulty levels (Figure 3). Our results suggest that this suite of assessments cover nearly the entire range of possible levels of difficulty, ranging from items that nearly all participants succeeded on (max. proportion correct: 0.99; 95% CI = [0.96, 1.00]) to items that nearly all participants responded to incorrectly (min. proportion correct: 0.012; 95% CI = [0.0, 0.036]). Moreover, the average gap in performance *between* items reliably exceeded the degree of uncertainty in estimates of performance on individual items (lowest-precision estimate: 0.46; 95% CI = [0.34, 0.57]), indicating that the observed variation in performance across items is reliable. We also found that participants performed below chance across 36 items out of the 230 total items (proportion correct furthest from chance: 0.012; 95% CI = [0.0, 0.036]).

How does performance vary across *tests*? While some of these tests were designed to measure data visualization literacy in the general adult population, others, such as **WAN**, aimed to assess data visualization comprehension skills among elementary-school-aged children, and **CALVI** focuses on the skills needed to detect misleading graphs. However, these tests have never before been directly compared to each other under consistent testing conditions.

Here we compared differences in performance across all five tests (Figure 4A) by fitting a logistic mixed-effects regression model predicting success on each trial, with *test* as a fixed-effects predictor and random intercepts for different items. We found that performance differs significantly between tests ($\chi^2(4) = 39.764$, $p < .001$), and on average, the suite of tests together covers a range of difficulties. This includes **WAN**, where participants' performance is closest to the ceiling (0.78; 95% CI = [0.71, 0.84]), and **CALVI** (0.41; 95% CI = [0.35, 0.48]), where participants performed closest to chance (**CALVI** average chance: 0.28), with other tests varying between these (**GGR**: 0.62; 95% CI = [0.48, 0.75]; **BRBF**: 0.69; 95% CI = [0.63, 0.73]; **VLAT**: 0.64; 95% CI = [0.58, 0.71]).

Taken together, these results are consistent with the notion that some of these tests are reliably harder than others, perhaps because they probe more advanced skills.

How does performance vary across *task types*? Perhaps one of the most salient ways different items can differ is the type of task they require people to perform, with some tasks being relatively simple (e.g., retrieving a single value from the visualization) and other tasks requiring additional computation (e.g., inferring the correlation between two variables). Here we sought to evaluate the degree to which there was a reliable difference in performance across items belonging to the three different categories of tasks: `value identification`, `arithmetic computation`, `statistical inference` (Figure 4B). As before, we fit a logistic mixed-effects regression model predicting success on each trial from *task* as a fixed-effects predictor and random intercepts for different items. This analysis revealed reliably different levels of performance across *task types* ($\chi^2(2) = 13.847$, $p < .001$; `value identification`: 0.71; 95% CI = [0.65, 0.77]; `statistical inference`: 0.60; 95% CI = [0.54, 0.65]; `arithmetic computation`: 0.51; 95% CI = [0.45, 0.58]). These findings are compatible with multiple possibilities, including the notion that some tasks are inherently harder than others — for instance, that `value identification` is easier than `arithmetic computation` with `statistical inference` in between. However, they also remain consistent with the possibility that these differences are a product of the way that the specific questions were posed (or the response options generated) in these tests, such that it is possible, in principle, for `value identification` to be arbitrarily difficult or to design easier `arithmetic computation` items that reduce the gaps between them.

How does performance vary across different *graph types*? Another prominent feature by which these items differ is the visual encoding used to present data. For example, a `bar chart` maps numerical values to the height of bars, while a `stacked bar chart` not only uses bar height to convey quantitative values, but can additionally convey the

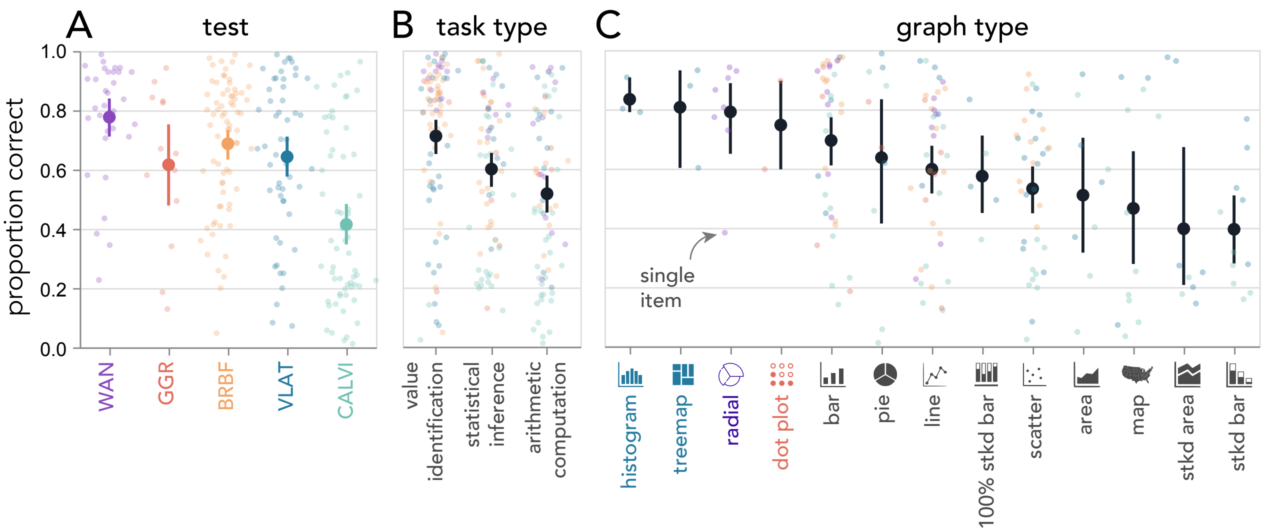


Figure 4: Performance across different *tests* (A), *task types* (B), and *graph types* (C), measured by the mean proportion of correct responses. Opaque dots indicate the mean proportion of correct responses for individual items. Error bars represent bootstrapped 95% confidence intervals.

distribution of values of a categorical variable. As such, it might be the case that some types of visualizations are simply more complex than others, and thus more difficult to comprehend. Here, we evaluate the degree to which there was a reliable difference in performance across the 13 different types of data visualizations (Figure 4C). As in the last section, we fit a logistic mixed-effects regression model predicting success on each trial from *graph type*, with random intercepts for different items, excluding those with tables.

We found that items using different *graph types* reliably differed in performance ($\chi^2(12) = 29.106$, $p = .004$), with the most difficult being `stacked bar charts` (0.40; 95% CI = [0.28, 0.51]) and the easiest being `histograms` (0.83; 95% CI = [0.79, 0.91]). However, some *graph types*, like `tree maps`, only appeared only in a single test, so performance on them is difficult to interpret, as it may rely on the difficulty of the test it came from rather than the type of graph. These findings provide some support for the notion that some types of graphs are inherently easier to interpret than others, though further work that uses more controlled manipulations of *graph type* (avoiding confounds with test) would provide a stronger test of this possibility.

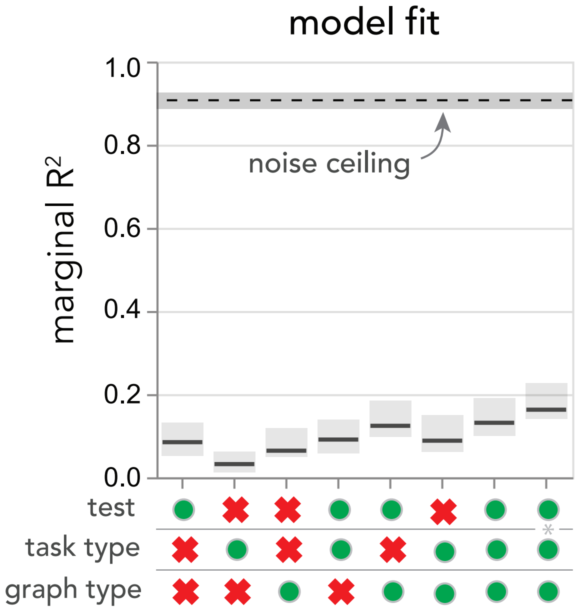


Figure 5: Comparison of model fit across mixed-effects logistic regression models, measured using marginal R^2. Green circles indicate fixed effects included in the model, with ‘*’ indicating an interaction between fixed effects. The noise ceiling is estimated by computing the squared Pearson correlation between split halves over participants’ data. Gray bands represent the expected variation R^2 due to sampling variability across samples of participants, estimated by bootstrap resampling.

How well do all of these features explain variation in performance? So far we have found that all three features explain some amount of variation in item-level difficulty. To what degree do they account for unique or shared variance? To answer that question, we examined to what degree every combination of these features further improved fit to our performance data. We fit every combination of these three features as fixed effects to different mixed-effects logistic regression models to participant errors (7 total; Figure 5). We additionally fit a model that includes an interaction term between *test* and *task type*, reflecting the possibility that the difficulty of an `arithmetic computation` item might depend on which test it was sourced from. Since not all *graph types* were paired with all three *task types*, we omitted

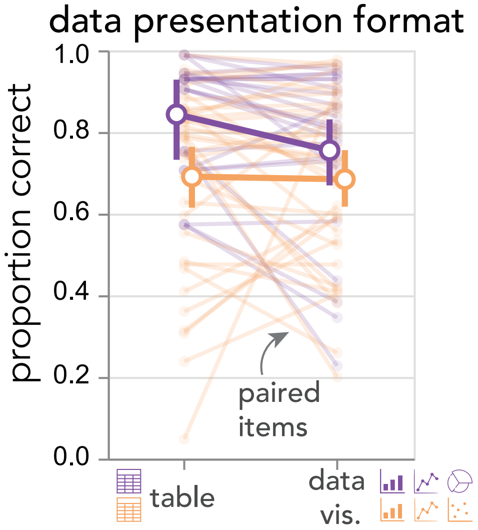


Figure 6: Performance across different *data presentation formats* (i.e. table or data visualization). Each line segment represents a pair of items in either **WAN** and **BRBF** which differ only in *data presentation format*. Error bars represent bootstrapped 95% confidence intervals.

the interaction between *graph type* and *task type*. Variation across items was modeled as a random effect.

We found that while the model with only *task type* as a fixed effect predicted the least amount of variance in the data (marginal R^2: 0.03; 95% CI = [0.01, 0.06]), model fit was improved significantly with the addition of *test* as a fixed effect ($\chi^2(4) = 29.233$, $p < .001$), and moderately with *graph type* as a fixed effect ($\chi^2(12) = 27.387$, $p = .006$).

We found that including all three factors provided the best fit (marginal R^2: 0.13; 95% CI = [0.09, 0.19]), with a reliable interaction between test and task (marginal R^2: 0.16; 95% CI = [0.14, 0.23]). Nevertheless, we found that all models still fell short of the noise ceiling (split-half R^2: 0.91; 95% CI = [0.89, 0.93]), a measure of reliability in our item-level estimates of performance. The relative size of this gap between even the best performing three-factor model and the noise ceiling suggests that the majority of the variance to be explained requires a model that can capture more subtle characteristics of each item that cause some to be more challenging than others.

How does performance vary between different *data presentation formats*? One potential barrier to answering certain questions about data visualizations might be the level of precision provided by these displays by comparison with other formats for displaying quantitative information, such as tables. For example, a `value identification` question might even be easier to answer precisely with a `table` than with a `scatter plot`. To explore potential differences in difficulty attributable to *data presentation format*, we included several pairs of items from two of the tests (Wainer, 1980; Boy et al., 2014) that were otherwise identical, except that one of them included a `table` and the other a data visualization.

We did not find evidence for an overall difference in performance on items with tables and those with data visualizations ($t(59) = -1.42$, $p = .160$), neither in **BRBF** (Δtable − data visualization: 0.01; 95% CI = [-0.10, 0.10]) nor in **WAN** (Δ table − data visualization: 0.09; 95% CI = [-0.04, 0.22]; Figure 6). These null results are compatible with several possibilities, including that data presentation formats do not strongly impact performance and the suite of items included in this evaluation were insufficient to resolve a global difference between formats. Given the substantial variability we observed across pairs of items, however, it seems more plausible that the degree to which data presentation format impacts performance depends more strongly on other factors (e.g., the question being asked, other characteristics of the data).

Discussion

Here we administered five assessments of data visualization literacy to a large sample of U.S. adults to obtain precise estimates of the difficulty of these items under consistent test conditions. We found that various features of these items (i.e., *graph type*, *task type*) could explain some item-level variation in performance, but there was substantial and reliable variation left unexplained. Thus, other features, or additional features, are needed in order to predict why some of them are more difficult than others.

In recent years, there has been broadening agreement on a general conceptual framework for data visualization literacy (Börner et al., 2019; Hedayati et al., 2024). However, there is not yet a consensus on a concrete set of instruments for measuring these literacy skills in a comprehensive manner, organized around components that are predictive of detailed patterns of performance. Our findings suggest major opportunities to develop unified measures of data visualization literacy which reliably evaluate the same skill set across individuals.

Having more unified measures is especially valuable for contributing to development and evaluation of computational models that can make explicit predictions on an item-by-item basis, as well as be used to test more specific hypotheses concerning the underlying mechanisms that support the understanding of complex visual inputs. In the future, such models might have the potential to explain in mechanistic terms why some operations with visual displays are more difficult for people than others (Nobre et al., 2024), and what strategies might be useful for overcoming those barriers. In addition, improved measures and models might help to account for reliable variation in data visualization literacy across individuals, who might have varying amounts of prior experience with mathematics, statistics, and other data-intensive subjects. In the long run, improved understanding of the mechanisms that support data visualization understanding might be leveraged to develop improved ways of helping more people acquire core quantitative literacy skills.

Acknowledgments

We thank members of the Cognitive Tools Lab for their input and feedback. This work was supported by NSF CAREER Award #2436199, NSF DRL #2400471, and awards from ONR Science of Autonomy, the Stanford Human-Centered AI Institute (HAI) and Stanford Accelerator for Learning.

Data and code availability

All experimental materials, data, and analysis code are available at `https://github.com/cogtoolslab/viz_item_measures_cogsci2025`.

References

Alper, B., Riche, N. H., Chevalier, F., Boy, J., & Sezgin, M. (2017). Visualization literacy at elementary school. In *Proceedings of the 2017 chi conference on human factors in computing systems* (pp. 5485–5497).

Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences*, *116*(6), 1857–1864.

Boy, J., Rensink, R. A., Bertini, E., & Fekete, J.-D. (2014). A principled way of assessing visualization literacy. *IEEE transactions on visualization and computer graphics*, *20*(12), 1963–1972.

Brehmer, M., & Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, *19*(12), 2376–2385.

Brockbank, E., Verma, A., Lloyd, H., Huey, H., Padilla, L., & Fan, J. E. (2025). Evaluating convergence between two data visualization literacy assessments. *Cognitive Research: Principles and Implications*, *10*(1), 15.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of experimental psychology: applied*, *4*(2), 75.

Ciccione, L., Sablé-Meyer, M., Boissin, E., Josserand, M., Potier-Watkins, C., Caparos, S., & Dehaene, S. (2023). Trend judgment as a perceptual building block of graphicacy and mathematics, across age, education, and culture. *Scientific Reports*, *13*(1), 10266.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, *79*(387), 531–554.

Cui, L., & Liu, Z. (2021). Synergy between research on ensemble perception, data visualization, and statistics education: A tutorial review. *Attention, Perception, & Psychophysics*, *83*, 1290–1311.

Delmas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In *Fourth forum on statistical reasoning, thinking, and literacy (srtl-4)* (Vol. 2).

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

Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison. *Medical decision making*, *31*(3), 444–457.

Ge, L. W., Cui, Y., & Kay, M. (2023). Calvi: Critical thinking assessment for literacy in visualizations. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–18).

Gillan, D. J., & Lewis, R. (1994). A componential model of human interaction with graphs: 1. linear regression modeling. *Human Factors*, *36*(3), 419–440.

Hedayati, M., Hunt, A., & Kay, M. (2024). From pixels to practices: Reconceptualizing visualization literacy. In *Chi 2024-workshop toward a more comprehensive understanding of visualization literacy.*

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). *Visual analytics: Definition, process, and challenges*. Springer.

Lee, C., Yang, T., Inchoco, G. D., Jones, G. M., & Satyanarayan, A. (2021). Viral visualizations: How coronavirus skeptics use orthodox data practices to promote unorthodox science online. In *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–18).

Lee, S., Kim, S.-H., & Kwon, B. C. (2016). Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics*, *23*(1), 551–560.

Lundgard, A., & Satyanarayan, A. (2021). Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, *28*(1), 1073–1083.

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

Mix, K. S., & Cheng, Y.-L. (2012). The relation between space and math: Developmental and educational implications. *Advances in child development and behavior*, *42*, 197–243.

Munzner, T. (2014). *Visualization analysis and design*. CRC press.

Nobre, C., Zhu, K., Mörth, E., Pfister, H., & Beyer, J. (2024). Reading between the pixels: Investigating the barriers to visualization literacy. In *Proceedings of the chi conference on human factors in computing systems* (pp. 1–17).

Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: a cognitive framework across disciplines. *Cognitive research: principles and implications*, *3*, 1–25.

Pandey, S., & Ottley, A. (2023). Mini-vlat: A short and effective measure of visualization literacy. In *Computer graphics forum* (Vol. 42, pp. 1–11).

Pinker, S. (2014). A theory of graph comprehension. In *Artificial intelligence and the future of testing* (pp. 73–126). Psychology Press.

Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in cognitive science*, *3*(3), 560–578.

Verma, A., Mukherjee, K., Potts, C., Kreiss, E., & Fan, J. E. (2024). Evaluating human and machine understanding of data visualizations. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

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