Evaluating convergence between two data visualization literacy assessments

Erik Brockbank[1,2], Arnav Verma[1], Hannah Lloyd[2], Holly Huey[2], Lace Padilla[3], Judith E. Fan[1,2]

[1]Department of Psychology, Stanford University.
[2]Department of Psychology, University of California San Diego.
[3]Department of Computer Science, Northeastern University.

Contributing authors: ebrockbank@stanford.edu; arnavv@stanford.edu; hslloyd@ucsd.edu; hhuey@ucsd.edu; l.padilla@northeastern.edu; jefan@stanford.edu;

Abstract

Data visualizations play a crucial role in communicating patterns in quantitative data, making data visualization literacy a key target of STEM education. However, it is currently unclear to what degree different assessments of data visualization literacy measure the same underlying constructs. Here, we administered two widely used graph comprehension assessments (Galesic & Garcia-Retamero, 2011; Lee, Kim, & Kwon, 2016) to both a university-based convenience sample and a demographically representative sample of adult participants in the United States (N=1,113). Our analysis of individual variability in test performance suggests that overall scores are correlated between assessments and associated with the amount of prior coursework in mathematics. However, these assessments also seem to probe somewhat distinct components of data visualization literacy, and we do not find evidence these components correspond to the categories that guided the design of either test (e.g., questions that require retrieving values rather than making comparisons). Together, these findings suggest opportunities for development of more comprehensive assessments of data visualization literacy that are organized by components that better account for detailed behavioral patterns.

Keywords: graph comprehension, graphical literacy, data literacy, psychometric evaluation, STEM education

Significance statement

Data visualizations are indispensable for communicating patterns in quantitative data. However, while several test-based measures of data visualization literacy exist, there is not yet clear agreement on what the key components of data visualization literacy are and how to measure them. In this study, we administered two widely used assessments of data visualization literacy to multiple diverse groups of U.S. adult participants. Participants who performed well on one assessment also generally did so on the other, suggesting some degree of convergence between these two measures. Moreover, performance on the combined assessment was associated with how much formal education in mathematics an individual had received, a measure of their convergence with other measures of quantitative literacy. However, it was less clear what underlying components of data visualization literacy these assessments measure. While it seems natural to assume that the ability to answer any question correctly on this assessment would be predicted by the type of graph shown or the type of question being asked, we used tools from machine learning to discover a small (and different) set of latent factors that could explain these patterns much more effectively. These findings lay the groundwork for future efforts to characterize what aspects of data visualization literacy these latent factors represent and to develop improved and unified measures of data visualization literacy.

Introduction

Data visualizations — also commonly known as *graphs*, *charts*, and/or *plots* — provide a powerful and versatile medium for reasoning about data (Bertin, 1981; Tufte, 1983; Wilkinson, 2012). They do so by leveraging color, shape, size, position, and other visual variables to convey quantitative patterns and relationships that might otherwise be difficult to discern when inspecting raw data. Although a relatively recent invention (Playfair, 1801; Spence, 2006), data visualizations are now vital to communication both among scientists and between scientists and members of the general public (Börner, Bueckle, & Ginda, 2019; Franconeri, Padilla, Shah, Zacks, & Hullman, 2021). As such, the ability to use visualizations to explore and reason about data is a key priority in STEM education (Council, 2014; Garfield & Gal, 1999).

However, a key challenge to successfully addressing this priority is a clear definition of what specific competencies are constitutive of the ability to use data visualizations effectively. This ability, often termed "visualization literacy," encompasses a broad suite of skills involved in the process of linking questions about data (which are often not inherently visual in nature) to visual patterns in graphical representations of those data (Boy, Rensink, Bertini, & Fekete, 2014; Brehmer & Munzner, 2013; Börner, Maltese, Balliet, & Heimlich, 2016; Creamer, Padilla, & Borkin, 2024; Friel, Curcio, & Bright, 2001; Hedayati, Hunt, & Kay, 2024; Shah & Hoeffner, 2002). Visualization literacy has been operationalized in a number of different ways across disciplinary contexts, including STEM education (Börner et al., 2019; Friel et al., 2001; Maltese, Harsh, & Svetina, 2015; Shah & Hoeffner, 2002), cognitive psychology (Boy et al., 2014; Padilla, 2018), human-computer interaction (Brehmer & Munzner, 2013; Lee et

al., 2016), and public health (Ancker, Senathirajah, Kukafka, & Starren, 2006; Galesic & Garcia-Retamero, 2011; Padilla et al., 2022).

While much of the existing work on visualization literacy focuses on the ability to understand formal data visualizations, other lines of work have explored the ability to design new visualizations (Alper, Riche, Chevalier, Boy, & Sezgin, 2017; Berg & Smith, 1994; Bishop et al., 2019) or use an existing visualization to make a sound decision (Price, Crumley-Branyon, Leidheiser, & Pak, 2016; Ruginski et al., 2016). Measures of data visualization understanding are especially important because these comprehension skills are foundational for more complex activities involving visualizations, such as design and decision-making (Börner et al., 2019; Hedayati et al., 2024). Moreover, reliable and valid measures of data visualization literacy are critical for evaluating the success of any educational intervention intended to improve visualization literacy skills. Finally, reliable measurement is crucial for developing cognitive theories of data visualization literacy — that is, theories of how graphs are mentally represented that explain why people find some questions about them easier to answer than others, as well as how the ability to understand graphs develops over time (Padilla, 2018; Padilla, Creem-Regehr, Hegarty, & Stefanucci, 2018; Pinker, 1990; Shah & Hoeffner, 2002).

Generally speaking, an individual's ability to read and interpret a data visualization is assessed using a sequence of test items, each one posing a question and providing a data visualization used to answer it. While there are currently several assessments that adopt this general strategy (Boy et al., 2014; Börner et al., 2016; DelMas, Garfield, & Ooms, 2005; Galesic & Garcia-Retamero, 2011; Garcia-Retamero, Cokely, Ghazal, & Joeris, 2016; Ge, Cui, & Kay, 2023; Lee et al., 2016; Maltese et al., 2015; Okan, Janssen, Galesic, & Waters, 2019; Pandey & Ottley, 2023), they define and operationalize the component skills in different ways. For instance, some assessments group items into a compact hierarchy of abstract abilities, progressing from "reading the data" to "reading beyond the data" (Friel et al., 2001; Galesic & Garcia-Retamero, 2011). Others group items into a broader suite of tasks that do not necessarily imply strong dependencies between them, such as finding extreme values or making comparisons (Boy et al., 2014; Lee et al., 2016; Pandey & Ottley, 2023). Still others focus on the ability to overcome intentionally misleading data visualizations (Ge et al., 2023) or misconceptions about distributions that are common among students enrolled in introductory statistics courses (DelMas et al., 2005).

But because these assessments have not been compared directly, it is unknown to what degree they converge with one another or imply the same decomposition of data visualization literacy into underlying component skills. As such, it remains unclear on what basis any given assessment should be preferred to provide the most reliable and valid measure of data visualization literacy. Towards filling this gap, here we compare two widely used assessments that measure data visualization literacy in distinct ways: the 13-item assessment developed by Galesic and Garcia-Retamero (2011), which we refer to as *GGR*, and the 53-item Visualization Literacy Assessment Test (*VLAT*; Lee et al. (2016)). We focused on these two assessments because, at the time this work was being conducted, they were among the most influential measures of data visualization literacy that could also be combined into a single assessment that could be administered in one session.

The items in GGR are organized into a three-level hierarchy of skills (Friel et al., 2001): "Level 1: Read the Data" (i.e., finding specific values in a graph); "Level 2: Read Between the Data" (i.e., comparing values in a graph); and "Level 3: Read Beyond the Data" (i.e., extrapolation). On the other hand, the items in VLAT are organized into a suite of eight skills (Amar, Eagan, & Stasko, 2005; Brehmer & Munzner, 2013): retrieving a value, finding extreme values, finding anomalies, making comparisons, determining a range, finding correlations and trends, characterizing distributions, and finding clusters.

We administered both assessments to a large and diverse sample of adult participants in the United States (U.S.). Our analyses of task performance were guided by three objectives. *First*, to characterize performance on the assessments across different sample demographics, as well as to estimate the association between performance on these assessments and the amount of prior coursework in mathematics. *Second*, we investigated the degree to which these two assessments produced convergent estimates of overall data visualization literacy levels, despite having been designed in different ways. *Third,* we sought to measure how well individual variability in test performance could be explained by the skill-based categories used to design each assessment. Taken together, this study offers empirical insights that serve as a foundation for the future development of more comprehensive and well-validated assessments of data visualization literacy.

Method

Participants

A total of 1,176 participants were recruited: 726 were students recruited from the University of California, San Diego study pool (211 male; mean age = 21). 450 adults were recruited using Prolific to obtain a sample that is demographically representative of the U.S. based on age, sex, and ethnicity (206 male; mean age = 45). This total sample size is comparable to the study reported in Galesic and Garcia-Retamero (2011), which included 987 participants, and larger than the sample initially recruited in Lee et al. (2016), which included 46 participants. All participants provided informed consent in accordance with the University of California, San Diego IRB. Participants were excluded for failing to complete the full assessment and the post-experiment survey of demographics and prior math experience, as well as for any technical issues reported in the post-experiment survey which prevented them from answering the questions. Unless otherwise indicated, we report results from analyzing the combined sample after exclusions (N=1,113 participants; U.S. university sample: N=714 participants; U.S. general public: N=399 participants).

Materials

Two assessments were included in our study: *GGR* (Galesic & Garcia-Retamero, 2011) and *VLAT* (Lee et al., 2016); see Figure 1A.

GGR is a 13-item assessment containing eight graphs: three *bar charts*, one *pie chart*, three *line plots*, and one *icon array* (Figure 1A , left). All items were assigned by

Num math courses	U.S. representative	U.S. university
0	39	9
1	136	46
2	116	146
3	108	513
Total	399	714

Table 1 Number of participants grouped by how many high school-level math courses they reported having previously taken.

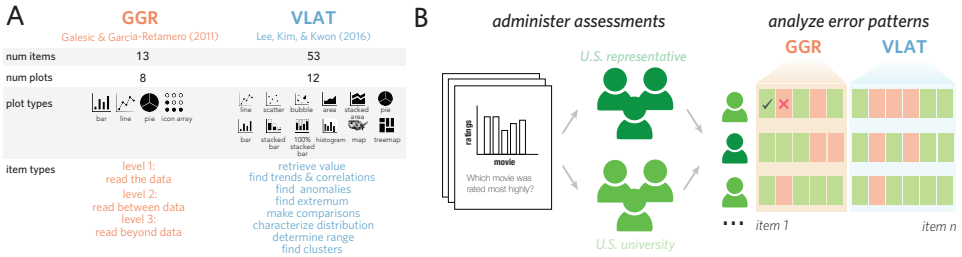

Fig. 1 **A:** The current study investigates two assessments of data visualization literacy: GGR (Galesic & Garcia-Retamero, 2011) and VLAT. (Lee et al., 2016). **B:** The combined assessment was administered to two groups of participants: a U.S. university sample recruited using a study pool and a U.S. demographically representative sample recruited using an online crowdsourcing platform.

the test developers to three categories: *Level 1: Read the Data*, *Level 2: Read Between the Data*, and *Level 3: Read Beyond the Data*. However, not all graph types were paired with all three types of questions. Nine items were fill-in-the-blank questions and four were multiple-choice questions with three response options. It was possible to skip any multiple-choice question, but not fill-in-the-blank questions, following the original test administration procedure. For all but one of the fill-in-the-blank items, it was necessary to provide a response that exactly matched the correct answer to be counted as correct; for the remaining item, the test developers allowed responses that fell within a range (i.e., between 23 and 25).

VLAT is a 53-item assessment containing 12 graph types: *line chart*, *bar chart*, *stacked bar chart*, *100% stacked bar chart*, *pie chart*, *histogram*, *scatter plot*, *bubble chart*, *area chart*, *stacked area chart*, *choropleth map*, and *tree map* (Figure 1A, right). These items were assigned by the test developers to eight question types: *retrieve value*, *find extremum*, *find anomalies*, *make comparisons*, *determine range*, *find correlations/trends*, *characterize distribution*, and *find clusters*. There were 16 True-False items; the remaining 37 multiple-choice items contained either three options (3 questions) or four options (34 questions). It was possible to skip any question.

Procedure

Participants completed the two assessments in a randomized order during a single session (see, Figure 1B), and each assessment was administered in a manner as similar to the original procedure as possible. Participants could spend as much time on each assessment as needed. After, they completed an optional post-study questionnaire that asked about their sex, age, ethnicity, and level of educational attainment. The possible responses to the educational attainment item were:"Have not graduated high school", "High school graduate, diploma or equivalent", "Associate degree", "Bachelor's degree", "Master's degree", "Professional degree (e.g., M.D., J.D.)", "Doctoral degree (e.g., Ph.D.)"). To obtain a proxy for the amount of prior mathematics knowledge participants had, which is especially relevant for tasks involving reasoning about data visualizations, we also prompted participants to indicate how many of the following high-school-level math courses they had previously taken: Algebra, Calculus, and Statistics.

Statistical Analyses

Overall, our statistical analyses aim to disentangle different potential sources of variation in how well participants performed on these assessments. The primary tool we use towards this end is linear regression, and the primary factors we consider are the type of graph used in an item, the type of question asked, prior mathematics knowledge, as well as the population from which participants were recruited (i.e., university study pool vs. demographically representative sample of U.S. crowd workers). To contextualize the degree to which these factors account for the explainable variance in these data, we additionally conduct exploratory factor analysis to infer the latent components that best predict participants' individual patterns of correct and incorrect responses. Below we describe the details of the statistical analysis strategy we use towards these ends.

Linear Models

We construct linear models to test the reliability of the association between several different variables of interest (i.e., graph type, question type, number of math courses, group membership) and test performance. We use nested model comparison because it provides a unified framework for hypothesis testing that generalizes beyond the narrower set of use cases that traditional hypothesis tests (e.g., t-tests, ANCOVA) were designed to independently handle. Specifically, to estimate the strength of the relationship between each of predictor variable of interest and test performance, we fit mixed-effects logistic-regression models to predict accuracy from that predictor variable, modeled as a fixed effect, and include random intercepts for each participant. We use logistic regression to predict the binary outcomes for individual items (i.e., correct vs. incorrect) because it provides more accurate estimates than using ordinary least squares regression to predict the proportion of correct responses across a set of items. To assess the explanatory value of any given predictor variable, we then use nested model comparison to assess how much more variance in performance could be explained when that variable was included in the linear model than when it was

omitted (i.e., the baseline model), accounting for the increase in model complexity when the variable was included. We implement these analyses using the *lmer* package in R (Baayen, Davidson, & Bates, 2008; Bates, Mächler, Bolker, & Walker, 2015) and report χ^2 statistics, degrees of freedom, and p-values for each model comparison.

Exploratory Factor Analysis

To investigate latent structure within and between assessments that was predictive of test performance, we employed exploratory factor analysis (EFA). EFA is a widely used dimensionality reduction method to uncover the set of latent *factors* that underlie observable patterns in data (Briggs & Cheek, 1986; Cowen & Keltner, 2017; Eisenberg et al., 2019; Haig, 2005). We apply EFA to the combined assessment in two ways. First, as a tool to infer how many factors are needed to account for the patterns of correct and incorrect responses generated by different participants. Second, we adopt the same formalization to compare existing methods of decomposing assessment items (by question type, graph type, and test) to explain the same error patterns.

To fit a factor model, each response on the combined assessment is modeled as a linear combination of latent factors and measurement error: $X - \mu = LF + \epsilon$, where X is an m (number of test items: 66) x n (number of participants) binary matrix of observed errors and μ is a matrix containing the mean score for each item. L is the m x f (number of factors) loading matrix, an estimate of how much each item contributes to each latent factor, and F is the f x n matrix of factor scores, an estimate of how much each participant's responses are predicted by each factor. ϵ represents measurement error, variance left unexplained by the latent factors.

We first apply EFA to the combined assessment to estimate the number of latent factors needed to account for error patterns while minimizing extraneous factors. Adding more factors improves prediction accuracy, but this often comes at a cost of interpretability. Researchers have proposed several methods for selecting the number of factors that best balances this tradeoff in a particular set of data (Preacher, Zhang, Kim, & Mels, 2013). We use Bayesian Information Criterion (BIC) to identify a minimal set of factors that account for individual patterns of error (Schwarz, 1978). We then compare this *latent factor* model to several *idealized* factor models based on graph type, question type, and test.

Rather than estimate the loading matrix L from participants' responses, our idealized factor models specify a loading matrix based on known decompositions of the assessment items. For example, the idealized loading matrix encoding question type information is a 66 x 11 (number of question types) matrix with binary values in each column encoding whether the item in each row belonged to that question type. This specification of L embodies the proposal that all of the items that involve the same question type "hang together" to explain error patterns, i.e., an individual who knows how to perform the operation for a particular question type is predicted to get all of those items correct or all of those items incorrect. In addition, encoding each item's question type independently in the loading matrix L allows for the *possibility* that different question types are entirely independent of one another, i.e., an individual who knows how to perform one question type task is not more likely to be able to perform another. Critically, structuring the loading matrix in this way does not *enforce* such

independence on the idealized factor models. The idealized factor models are derived by estimating the factor scores F given an idealized loading matrix L — in this way, systematic patterns in participants' responses that arise from, e.g., similarity across different question types can be expressed in the factor scores assigned to participants for those question types. Our analyses focus on how well idealized models with "manually" encoded loading matrices and freely varying factor scores are able to predict participants' responses.

We compare the performance of our fitted latent factor model to idealized factor models encoding test, question type, and graph type information. Each factor model predicts individual responses on all 66 assessment questions. These predictions can be compared to the actual responses to produce a vector of prediction errors for each participant. We calculate each participant's mean squared error (the average of item-level squared errors for each participant). The average of all participant mean-squared-error (MSE) values for a given model provides a group-level MSE value for that factor model, allowing us to compare models according to their overall predictive accuracy. For the latent factor model, which does not specify a particular factor loading matrix in advance, we obtain this MSE estimate using five-fold cross-validation. We fit a separate factor model to each set of training data folds, then use the loading matrix from the training set to estimate a factor score matrix for the held-out data. This model is used to then predict individual responses in the held-out data. The group-level MSE value for this model is calculated based on the held-out prediction error for each participant. This ensures that evaluation of model performance is always based on splits of the data that are independent from those used to fit the latent factor model's loading matrix.

Confidence Intervals

To provide quantitative estimates of effect size, we report 95% confidence intervals (CIs) for various quantities of interest (e.g., average test performance). Where we have used linear models to fit the data, these confidence intervals were constructed using estimates of standard error based on the linear model itself. For estimates of mean squared error (MSE) for predictions made by exploratory factor analysis model, we used bootstrap resampling methods, which have the advantage of not depending on parametric assumptions about the sampling distribution of the statistic (Efron & Tibshirani, 1994). This approach entailed resampling $N = 1,113$ individual participant MSE values with replacement and calculating a group-level MSE value from this sample. We repeated this process 10,000 times to estimate a sampling distribution of group-level MSE values from which the 2.5th and 97.5th percentile values could serve as confidence interval endpoints.

Results

Our analyses[1] were guided by three main objectives. First, we sought to estimate differences in performance on the combined assessment between groups of participants, depending on how they were recruited and how much prior coursework in mathematics they had completed, providing initial insights into potential sources of variability in

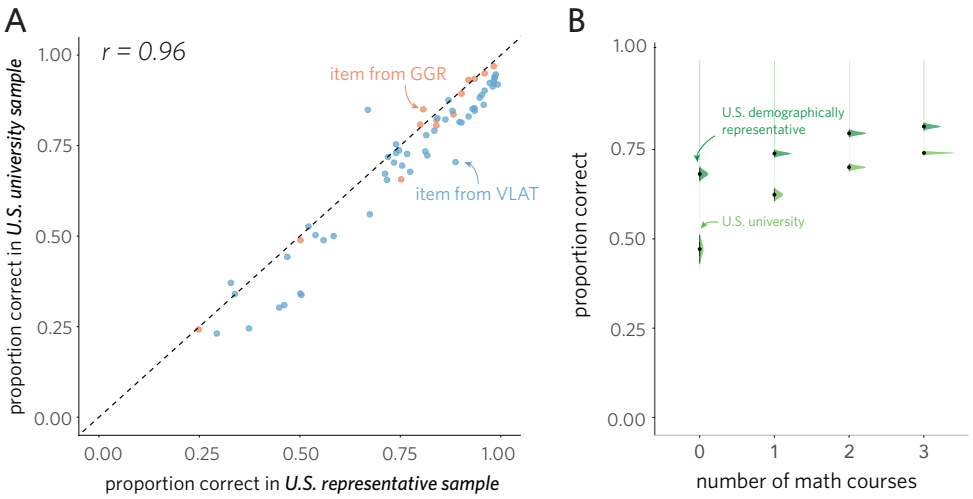

Fig. 2 **A:** Mean performance on each test item in both U.S. university and U.S. demographically representative samples. Each dot is a test item. **B:** Overall performance in each group as a function of the number of math courses (i.e., algebra, calculus, statistics) previously taken. Error bars represent standard error of the mean (SEM).

performance across individuals. Second, we assessed how strongly performance on one assessment was associated with performance on the other, providing a preliminary measure of these two assessments' convergent validity. Third, we evaluated how well variability in performance could be explained by the skill-based categories used to group items in each assessment, such as which type of graph was presented or what type of question was being asked. We compared the predictive value these skill-based categories to that of an alternative data-driven decomposition, providing a principled way of assessing the reliability of these categories based on their ability to predict individual error patterns.

Comparing performance across groups

On average, participants answered 75.7% of test items correctly (95% CI: [74.9%, 76.5%]), though test performance varied substantially across participants (SD: 14.0%; min: 3.0%; max: 97.0%). These findings indicate both that participants were neither at ceiling nor at floor on these assessments, and that there is meaningful individual variability in performance to explain. We also observed that accuracy for the demographically representative sample (78.6%, 95% CI: [77.4%, 79.7%]) was higher on the combined assessment than for the university sample (73.9%, 95% CI: [72.8%, 75.0%]). Nevertheless, we found that relative performance on individual test items was highly correlated between samples ($\rho = 0.96$, 95% CI = [0.94, 0.98], $p < .0001$; Figure 2A). Together, these results suggest that while the two groups of participants performed

[1]All data, along with analysis scripts used to generate the current results, are publicly available at the following github repository: https://github.com/erik-brockbank/visualization_literacy_public.git.

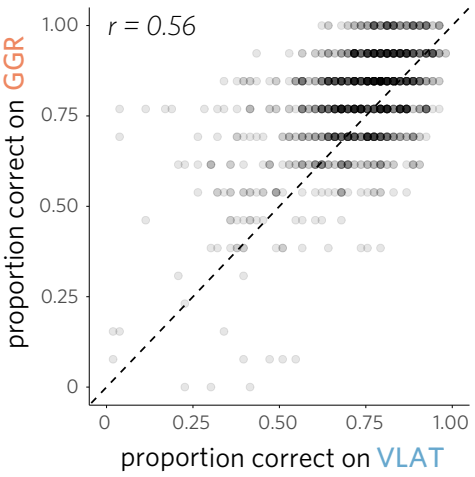

Fig. 3 Correlation between performance on VLAT and GGR assessments for individual participants.

at different levels on average, data from both samples provide convergent estimates of the test items' relative difficulty.

We next sought to explore the relationship between performance on these assessments and other relevant characteristics of these participants, with a focus on how much high-school-level coursework in mathematics they had completed (Table 1). Toward that end, we grouped participants by how many math courses they reported having previously taken (among Algebra, Calculus, and Statistics). We found that the number of math courses an individual had taken was a reliable predictor of overall test performance ($\chi^2(3) = 40.04$, $p < .0001$), with more prior math courses leading to higher predicted performance (0: 65.1%, 1: 72.5%, 2: 75.9%, 3: 77.1%); the strength of this association did not differ significantly between sample groups ($\chi^2(3) = 3.96$, $p = .27$; see Figure 2B).

Comparing performance across assessments

On average, participants achieved a level of performance that was reliably well above chance, yet below ceiling, on both tests (GGR: 80.4%, 95% CI: [79.5%, 81.3%]; VLAT: 74.6%, 95% CI: [73.7%, 75.5%]). However, there was also substantial individual variability in test performance (GGR $SD = 15.0\%$, VLAT $SD = 15.1\%$), with an individual's score on one test being moderately predictive of their score on the other ($\rho = 0.56$, 95% CI: [0.52, 0.60]; Figure 3). These findings provide an initial estimate of these two assessments' ability to reliably measure the same construct. Nevertheless, this analysis does not resolve what underlying factors account for the observed level of convergence between assessments and what accounts for the remainder of the gap between them.

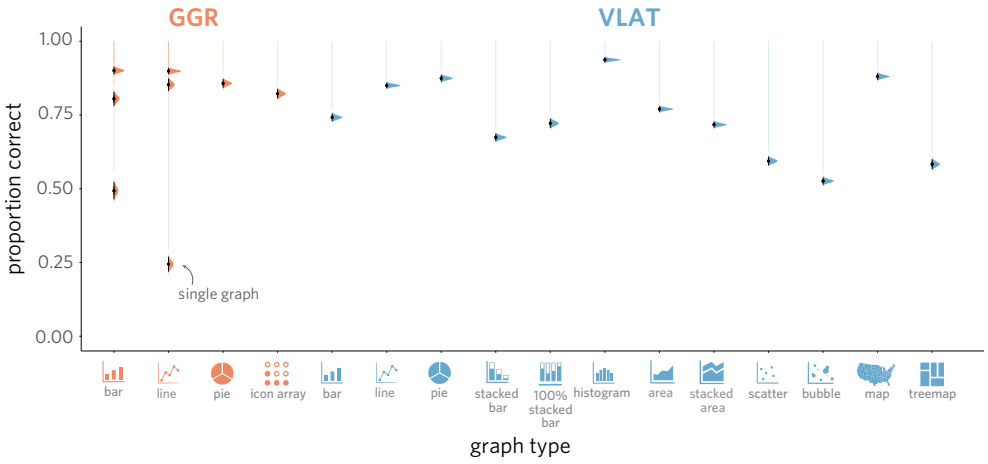

Fig. 4 Mean proportion correct for every type of graph in the combined assessment, disaggregated by test. Point estimates are plotted for each individual graph, aggregating questions that pertain to the same graph. GGR contains multiple instances of bar plots and line plots, one pie chart and one icon array. VLAT contains exactly one instance of each type graph. The sampling distributions for each point estimate are shown along with error bars representing the standard error of the mean (SEM).

Comparing performance across graph type

One possibility that could account for the moderate correlation between scores on each test is the use of similar types of graphs. For instance, some of these graph types might be ones that most individuals know how to interpret, while others are ones that only a minority of individuals are familiar with. Insofar as graph type drives variability in test performance, participants would be expected to achieve higher accuracy on questions involving more familiar graphs, and lower accuracy on questions on the less familiar ones.

To explore that possibility, we took an inventory of the types of graphs appearing in each assessment. We observed that three graph types appeared in both GGR and VLAT (i.e., *bar chart*, *line graph*, and *pie chart*), while there was one additional graph type that appeared only in GGR (i.e., *icon array*) and nine additional graph types appearing only in VLAT (i.e., *stacked bar*, *100% stacked bar*, *histogram*, *area chart*, *stacked area chart*, *scatter plot*, *bubble chart*, *map*, *treemap*). We found that performance reliably varied across graph types in the combined assessment ($\chi^2(12) = 5066.18$, $p < .0001$; Figure 4). While both assessments have some overlap in graph types (i.e., bar graphs, line graphs, and pie charts), VLAT uses a broad range of additional graphs; this raises the possibility that the observed effect of graph type on accuracy reflects the combination of the two assessments. However, we find that performance varies significantly by graph type even when considering each assessment individually (*GGR*: $\chi^2(3) = 125.18$, $p < .0001$; *VLAT*: $\chi^2(11) = 4919.30$, $p < .0001$). The magnitude of this effect also reliably differed between samples (combined: $\chi^2(12) = 130.93$, $p < .0001$; GGR: $\chi^2(3) = 8.67$, $p = .03$; VLAT:

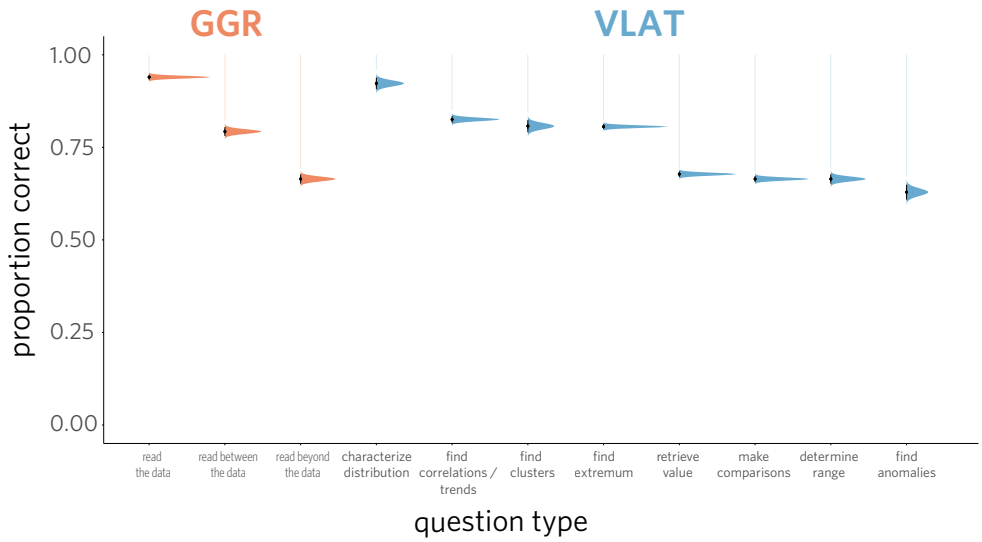

Fig. 5 Mean proportion correct for every question type in the combined assessment, disaggregated by test. The sampling distributions for each point estimate are shown along with error bars representing the standard error of the mean (SEM).

$\chi^2(11) = 112.01$, $p < .0001$), being larger in the demographically representative sample than in the university sample, perhaps reflecting the greater diversity in that sample relative to the university-based sample. Taken together, these results indicate that graph type accounts for a meaningful amount of variation in test performance, suggesting that participants found it easier to answer questions involving some kinds of graphs than others.

Comparing performance across question type

An additional factor that might account for variation in test performance is the type of question being asked. Perhaps some questions rely on skills that are more broadly shared across participants in the study, such as *Level 1: Read the Data* from GGR and *retrieve value* from VLAT, while other types of questions require understanding of more advanced statistical concepts that are familiar only to a minority of participants, such as the ability to *find correlations/trends* or *characterize distribution* in VLAT.

Insofar as question type is a driver of variability in test performance, participants would be expected to achieve a higher level of accuracy on some types of questions than others. Consistent with this possibility, we found that performance reliably varied by question type, both when each assessment was analyzed independently (*GGR*: $\chi^2(2) = 1331.61$, $p < .0001$; *VLAT*: $\chi^2(7) = 1981.92$, $p < .0001$; Figure 5) and when conducting the same analysis for all 11 question types in both assessments ($\chi^2(10) = 3585.91$, $p < .0001$). In further exploratory analyses, we found that variation in performance associated with question type was greater in the demographically representative sample (overall: $\chi^2(10) = 94.18$, $p < .0001$; GGR: $\chi^2(2) = 1.57$, $p = .46$;

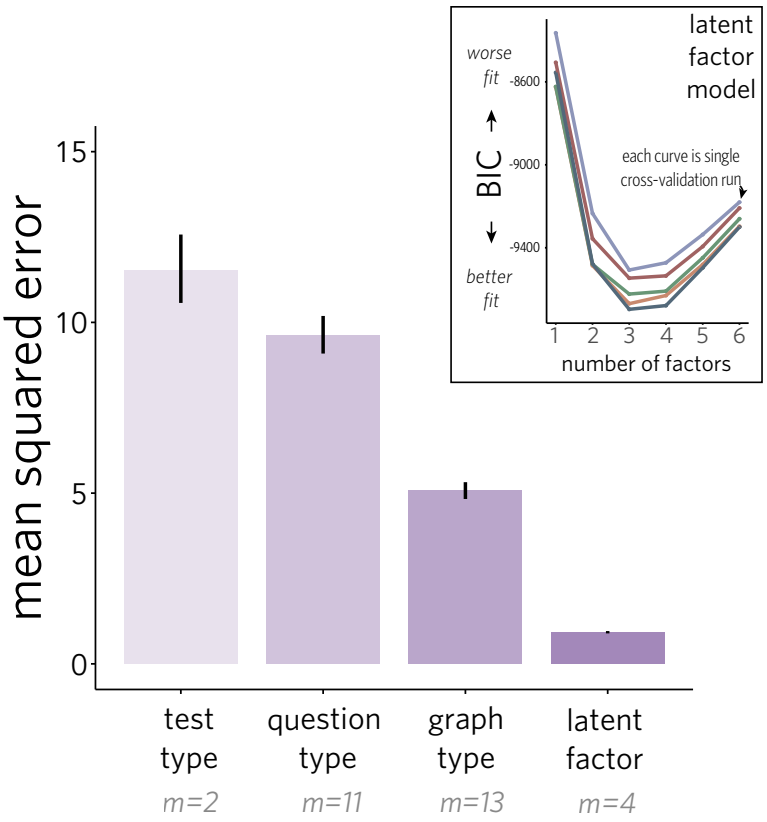

Fig. 6 Comparing the ability of different factor-based models to predict individual participant error patterns. Inset indicates number of factors needed by best-performing latent factor model (each curve depicts model fit with increasing factors for a different subset of the data).

VLAT: $\chi^2(7) = 71.03$, $p < .0001$), perhaps related to the greater heterogeneity in that sample. These results indicate that question type accounts for variation in test performance, suggesting that some types of questions are reliably more difficult than others.

Comparing predictive models of performance

Our findings so far provide evidence that both the type of graph used and the type of question being asked accounts for at least some of the explainable variance in average performance at the group level. However, an even stronger test of the explanatory value of these factors is their ability to account for variation in the *patterns* of errors that different individuals produce. For instance, insofar as the ability to "read the data" is distinct from the ability to "read beyond the data," with some individuals having mastered one of these, and other individuals having mastered both, we would expect to be able to predict *which* questions are more likely to be answered correctly by some individuals than others in terms of those skills.

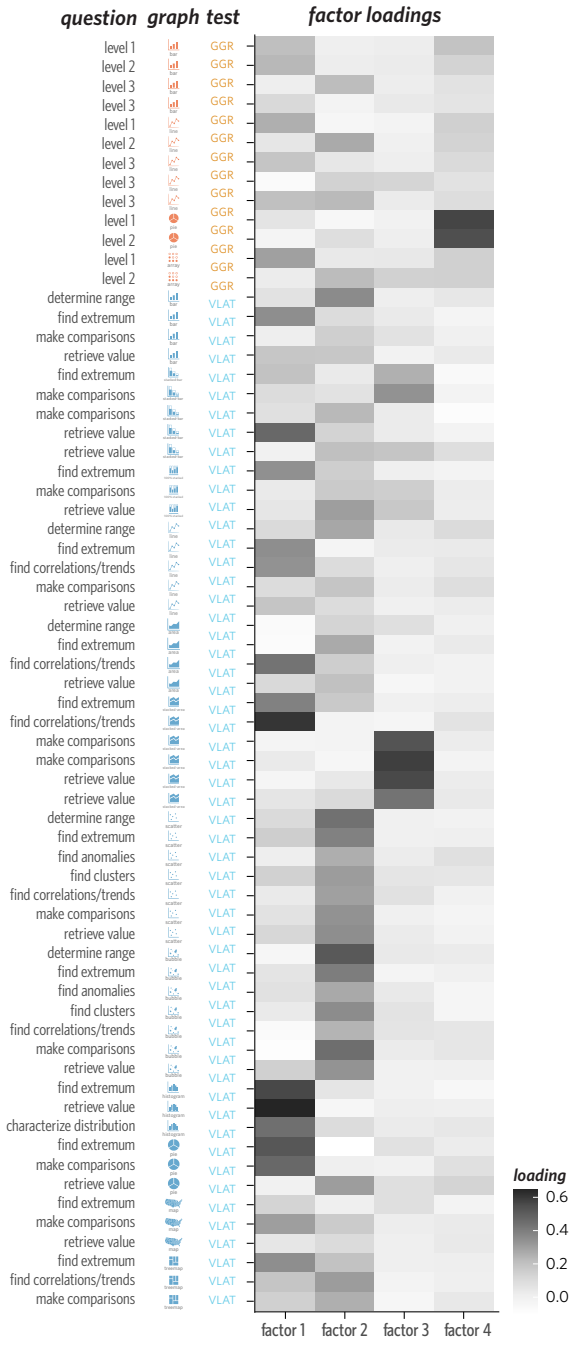

Fig. 7 Matrix indicating factor loading values across all items based on the latent factor model, grouped by test, question and graph type. Darker cells reflect higher factor loading values.

To explore how well graph type and question type predicts these individual differences in absolute terms, we sought to establish an upper bound for how well they could be explained by any decomposition of these assessments. Towards that end, we used exploratory factor analysis (EFA) to infer the decomposition of the combined assessment that best accounted for observed error patterns across all items (see *Exploratory Factor Analysis* in the Methods section for details).

First, we investigated how many factors would be needed to achieve strong out-of-sample behavioral predictivity without being too complex. When fitting the data with a variable number of factors, we found that a model with three to four factors consistently achieved the best performance, as measured using Bayesian Information Criterion (BIC; Figure 6, *inset*). These findings suggest that participants' error patterns can be explained by a relatively compact set of factors, and far fewer than there are unique types of questions and graphs across VLAT and GGR.

Next, we sought to directly compare the performance of idealized factor models based on graph type (13 factors), question type (11 factors), and test (2 factors) compared with that achieved by this *latent factor* model (4 factors) on held-out data under five-fold cross-validation. We compared the mean squared error (MSE) of the predictions made by each of our models (Figure 6). We find that both the *question type* (MSE = 9.62, 95% CI: [9.09, 10.19]) and *graph type* (MSE = 5.07, 95% CI: [4.83, 5.32]) models perform better than a 2-factor *test type* model (MSE = 11.52, 95% CI: [10.57, 12.57]), suggesting that these features of the assessment items allow for systematic predictions of participants' errors relative to the differentiation made by test alone. Further, we find that the 13-factor *graph type* model performs better than the 11-factor *question type* model, suggesting that fluency with some graphs and not others explains more variance in participant responses than their comfort with particular question types (Peebles & Cheng, 2003). Finally, both the *graph type* and *question type* models perform substantially worse than the 4-factor *latent factor* model (MSE = 0.93, 95% CI: [0.89, 0.96]). Taken together, these results suggest that neither graph type nor question type on their own can account for much of the explainable variation in individual error patterns.

Instead, these error patterns may reflect a more complex interaction between graph type and question type, which is captured by the latent factor model. If that were the case, then some combinations of question type and graph type (e.g., *find extremum* for a histogram) would be expected to load strongly on a latent factor, while other items involving either just that question type or just that graph type would not. While items involving scatterplots and bubble charts seem to load onto `factor 2`, perhaps reflecting the similarity between these types of graphs, it is far less clear what unifies the items that load onto the remaining factors (Figure 7). Indeed, it seems possible that there are features of these items beyond graph type and question type information that might be needed to better explain these behavioral data. In sum, a small number of factors seems sufficient to account for individual error patterns across VLAT and GGR, but these factors do not obviously reflect the categories often used to differentiate items in these assessments.

Discussion

In this study, we administered two widely used assessments of data visualization literacy (Galesic & Garcia-Retamero, 2011; Lee et al., 2016) to multiple independently recruited samples of U.S. adult participants. Participants who performed well on one assessment also generally performed well on the other, suggesting some degree of convergence between these two measures. Moreover, performance on the combined assessment was associated with how much formal education in mathematics participants had received, an indicator of these assessments' convergence with other measures of quantitative literacy. However, further investigation of individual variability in the patterns of mistakes that participants made suggests that these assessments probe a suite of skills that are only partially aligned with the grouping of items according to the type of question being asked or the type of graph being shown. That is, while there is some variance explained by these two variables, there remains substantial variance that remains unexplained by them and is better explained by an alternative set of factors that does not seem to have been explicitly encoded into the design of these two assessments.

To make sense of these results, it could be helpful to draw a distinction between an abstract framework for organizing the space of skills relevant to data visualization literacy (i.e., graph type, question type) and the concrete measures used to probe that space of skills. Our results could imply limitations in the measures, the underlying framework, or both. However, it is often not feasible to distinguish between these possibilities on the basis of any individual study, including the current one.

Nevertheless, clarifying these issues will likely involve conducting more thorough evaluations that include an expanded set of measures. For example, it might be useful to include assessments that focus on the ability to overcome potentially misleading data visualizations (Ge et al., 2023), as well as specific misconceptions that are common among students enrolled in introductory statistics courses (DelMas et al., 2005). However, even the full suite of existing assessments is still limited in several key ways. First, there are only a few examples of each type of graph represented across all of them. As such, the type of graph is almost always confounded with the variables being plotted, leaving it unclear whether variation in performance is attributable to core visualization literacy skills, rather than other factors (e.g., prior knowledge about those variables). Second, each test contains a relatively small and fixed set of items. With so few items, it is challenging to estimate the reliability with which any given skill is being measured. With only fixed sets of items, it is also not feasible to measure changes in data visualization literacy within the same individual, which is critical for assessing the impact of formal instruction.

Thus, it would be valuable to develop new assessments, leveraging both best practices in psychometric research already exemplified by the design of existing tests (Boy et al., 2014; Lee et al., 2016), as well as practical strategies from modern machine learning for scaling the generation of test items for inclusion in cognitive assessments (Masry, Long, Tan, Joty, & Hoque, 2022; Methani, Ganguly, Khapra, & Kumar, 2020; Zelikman et al., 2023). In addition, given the broad suite of skills that are recruited when interpreting a data visualization, it can be cumbersome to administer complete versions of every test, if the goal is to quickly assess an individual's literacy level.

A more efficient alternative might be to use a more targeted set of items identified using item-response theory that are particularly diagnostic (Pandey & Ottley, 2023), or even use adaptive testing protocols that dynamically propose sequences of items to administer that will be most informative about an individual's literacy level given their responses so far (Cui et al., 2023). The approach taken in the current study could be used in conjunction with these more targeted and adaptive test development strategies to identify multiple facets of visualization literacy that would be valuable to estimate, which would entail going beyond conceptualizing literacy level as varying along a single dimension. Such improved assessments would be valuable for advancing educational assessment — the understanding of how well core data literacy skills are being learned in real-world educational settings.

Separately, our findings are also consistent with the possibility that there might be alternative frameworks for decomposing data visualization literacy that provide both more detailed and generalizable ways of predicting quantitative patterns in task performance. For instance, recent work employing qualitative methods to analyze process-level barriers to correct interpretation of data visualizations in VLAT has emphasized distinctions between errors in translating verbal questions into visual queries and errors in the interpretation of plot elements (Nobre, Zhu, Mörth, Pfister, & Beyer, 2024). Extending such process-level analyses might be a promising route towards clarifying the relationship between the empirically derived decomposition uncovered in the current study and the typologies of data visualization literacy skills proposed in prior work (Börner et al., 2019; Brehmer & Munzner, 2013; Friel et al., 2001). Measuring those component skills is important because they enable differentiation between individuals who might otherwise seem equally proficient, but actually have different strengths and weaknesses. Reliably diagnosing those strengths and weaknesses makes it possible to then provide instruction that is more effectively tailored to each individual.

Beyond their role in educational assessment and instruction, new measures of data visualization literacy could also be instrumental for advancing fundamental understanding of the cognitive processes involved in the successful interpretation of data visualizations (Padilla, 2018; Pinker, 1990; Shah & Hoeffner, 2002). These processes include the rapid perceptual computations (Cleveland & McGill, 1984) performed with respect to a known graph schema (Pinker, 1990), explicit numerical operations (Gillan & Lewis, 1994) constrained by fintite working memory resources (Padilla et al., 2018), and interpretive processes that lead to more general insights (Carpenter & Shah, 1998), which may be influenced by prior content knowledge (Shah & Freedman, 2011). A more thorough understanding of each of these cognitive processes is a crucial step towards more unified cognitive models of data visualization understanding. One important purpose of such cognitive models is to explain why someone finds some questions easier to answer with one graph than another (Huey, Oey, Lloyd, & Fan, 2023; Shah & Hoeffner, 2002). Previously developed cognitive models have proposed qualitative accounts of how people reason about data visualizations (Carpenter & Shah, 1998; Padilla et al., 2018). A promising avenue for future work is to develop *computational* cognitive models that specify the operations performed in explicit and quantitative terms: the form of the input, the form of the output, and the exact operations applied

in between. Computational cognitive models have enabled major advances across several cognitive domains because they not only offer precise specifications of the mental processes involved, but also generate concrete behavioral outputs that can be directly compared to what people produce given the same inputs (Bear et al., 2021; Cao & Yamins, 2024; Hu, Floyd, Jouravlev, Fedorenko, & Gibson, 2022; Mukherjee et al., 2024; Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021). Future work employing these approaches are especially timely, as computational cognitive models — and in particular, AI systems that perform complex real-world tasks — have only recently advanced to the point that it is feasible to measure these models' behavior on tasks that approach the complexity of those that humans encounter in real-world settings (Bommasani et al., 2021).

In the current study, we measured a positive relationship between the number of mathematics courses an individual had previously taken and how well they performed on the combined assessment, consistent with prior work examining the association between formal education and behavior on tasks involving data visualizations (Harsh et al., 2019; Maltese et al., 2015). However, while such correlative findings are suggestive, experimental studies are needed to firmly establish any causal relationships between specific learning experiences and subsequent task performance (Bhatt, Guryan, Khan, LaForest-Tucker, & Mishra, 2024; Koedinger, Carvalho, Liu, & McLaughlin, 2023; Solomon et al., 2019). A promising approach that might be complementary to such studies with real students might be to conduct so-called *in silico* experiments with computational cognitive models. These models can be used to develop and test hypotheses about what kinds of experience are needed to acquire various data visualization literacy skills. A major advantage of *in silico* experiments is that they enable researchers to efficiently sweep through a wider range of possible learning conditions than can be practically and ethically implemented in real-world educational environments with human learners. By systematically manipulating the amount and type of experience a computational model receives, it is possible to investigate what kinds of experience are needed to succeed on some tasks and generalize to others (Gupta et al., 2024; Zamir et al., 2018; Zhuang et al., 2021). For example, a vision-language model might be pretrained on a suite of visual, language-based, and quantitative reasoning tasks before being evaluated on its ability to accurately interpret data visualizations (Gupta et al., 2024). Comparisons between this model and others that had been pretrained on only a subset of the same tasks could be used to assess the necessity of certain kinds of prior experience to generalize to reasoning tasks involving data visualizations.

Conclusions

In sum, the current study evaluated the convergent validity of two commonly used assessments of data visualization literacy. We found that these two measures exhibited a reasonable degree of convergence, such that people who achieved a high score on one assessment often also did on the other. In addition, we observed that individual variability in performance was related to how much formal education in mathematics an individual had received. To gain insight into the component skills that underlie

the observed relationship between assessments, we used tools from machine learning to identify latent factors that best predicted individual error patterns. These latent factors achieve high predictive accuracy but do not align with existing typologies that have been used to structure these assessments (i.e., the type of graph, the kind of task being performed), suggesting the need for future work to characterize what component skills these factors identify. In sum, this work lays the groundwork for future efforts to develop improved and unified assessments that provide accurate and reliable estimates of the underlying components of data visualization literacy (Uttal, McKee, Simms, Hegarty, & Newcombe, 2024). Over the long run, the development of unified measures of data visualization literacy is not only crucial for advancing cognitive theories of how people learn to extract meaning from abstract graphical representations, but might also lead to improved ways of teaching graphical literacy skills in real-world educational settings.

Open Practices

All data reported in this manuscript, along with analysis scripts used to generate the current results, are publicly available at the following GitHub repository: https://github.com/erik-brockbank/visualization_literacy_public.git.

Declarations

Ethics approval and consent to participate

All participants in the experiments reported here provided informed consent in accordance with the University of California, San Diego Institutional Review Board.

Consent for publication

Not applicable.

Availability of data and material

All data described in this manuscript, along with analysis scripts used to generate the current results, are publicly available at the following GitHub repository: https://github.com/erik-brockbank/visualization_literacy_public.git.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by NSF DRL award #2400471 and NSF CAREER award #2047191 to J.E.F. E.B. is supported by NSF SBE Postdoctoral Research Fellowship #2404706. L.P. is also supported by NSF CAREER Award #2238175 and NSF EAGER Award #2122174.

Authors' contributions

EB: conceptualization, methodology, software, validation, formal analysis, data curation, writing - original draft, writing - review and editing, visualization, supervision, project administration. **AV:** methodology, software, validation, formal analysis, data curation, visualization. **HL:** conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing - original draft, visualization, supervision, project administration. **HH:** methodology, software, investigation, writing - original draft, visualization. **LP:** conceptualization, methodology, writing - original draft, writing - review and editing, supervision, project administration, funding acquisition. **JEF:** conceptualization, methodology, validation, writing - original draft, writing - review and editing, visualization, supervision, project administration, funding acquisition.

Acknowledgements

Thanks to Amy Rae Fox for helpful feedback on this manuscript.

References

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

Amar, R., Eagan, J., Stasko, J. (2005). Low-level components of analytic activity in information visualization. *Ieee symposium on information visualization, 2005. infovis 2005.* (pp. 111–117).

Ancker, J.S., Senathirajah, Y., Kukafka, R., Starren, J.B. (2006). Design features of graphs in health risk communication: a systematic review. *Journal of the American Medical Informatics Association*, *13*(6), 608–618,

Baayen, R.H., Davidson, D.J., Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390–412,

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48, https://doi.org/10.18637/jss.v067.i01

Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.-Y.F., Pramod, R., ... others (2021). Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, ,

Berg, C.A., & Smith, P. (1994). Assessing students' abilities to construct and interpret line graphs: Disparities between multiple-choice and free-response instruments. *Science Education*, *78*(6), 527–554,

Bertin, J. (1981). *Graphics and graphic information processing.* Walter de Gruyter.

Bhatt, M.P., Guryan, J., Khan, S.A., LaForest-Tucker, M., Mishra, B. (2024, May). *Can technology facilitate scale? evidence from a randomized evaluation of high dosage tutoring* (Working Paper No. 32510). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w32510

Bishop, F., Zagermann, J., Pfeil, U., Sanderson, G., Reiterer, H., Hinrichs, U. (2019). Construct-a-vis: Exploring the free-form visualization processes of children. *IEEE transactions on visualization and computer graphics*, *26*(1), 451–460,

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, ,

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,

Briggs, S.R., & Cheek, J.M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of personality*, *54*(1), 106–148,

Börner, K., Maltese, A., Balliet, R.N., Heimlich, J. (2016). Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, *15*(3), 198-213,

Cao, R., & Yamins, D. (2024). Explanatory models in neuroscience, part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, 101244,

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,

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,

Council, N.R. (2014). Developing assessments for the next generation science standards.

Cowen, A.S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, *114*(38), E7900–E7909,

Creamer, M., Padilla, L., Borkin, M.A. (2024). Finding gaps in modern visualization literacy.

Cui, Y., Lily, W.G., Ding, Y., Yang, F., Harrison, L., Kay, M. (2023). Adaptive assessment of visualization literacy. *IEEE Transactions on Visualization and Computer Graphics*, ,

DelMas, R., Garfield, J., Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. *Proceedings of the fourth international research forum on statistical reasoning, thinking and literacy 2005.*

Efron, B., & Tibshirani, R.J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

Eisenberg, I.W., Bissett, P.G., Zeynep Enkavi, A., Li, J., MacKinnon, D.P., Marsch, L.A., Poldrack, R.A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature communications*, *10*(1), 2319,

Franconeri, S.L., Padilla, L., Shah, P., Zacks, J.M., Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest*, *22*(3), 110–161,

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,

Garcia-Retamero, R., Cokely, E.T., Ghazal, S., Joeris, A. (2016). Measuring graph literacy without a test: A brief subjective assessment. *Medical Decision Making*, *36*(7), 854–867,

Garfield, J.B., & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International statistical review*, *67*(1), 1–12,

Ge, L.W., Cui, Y., Kay, M. (2023). Calvi: Critical thinking assessment for literacy in visualizations. *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,

Gupta, A., Gupta, V., Zhang, S., He, Y., Zhang, N., Shah, S. (2024). Enhancing question answering on charts through effective pre-training tasks. *arXiv preprint arXiv:2406.10085*, ,

Haig, B.D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research*, *40*(3), 303–329,

Harsh, J.A., Campillo, M., Murray, C., Myers, C., Nguyen, J., Maltese, A.V. (2019). "seeing" data like an expert: An eye-tracking study using graphical data representations. *CBE—Life Sciences Education*, *18*(3), ar32,

Hedayati, M., Hunt, A., Kay, M. (2024). From pixels to practices: Reconceptualizing visualization literacy.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., Gibson, E. (2022). A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*, ,

Huey, H., Oey, L.A., Lloyd, H., Fan, J.E. (2023). How do communicative goals guide which data visualizations people think are effective? *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Koedinger, K.R., Carvalho, P.F., Liu, R., McLaughlin, E.A. (2023). An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, *120*(13), e2221311120,

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,

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,

Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E. (2022). Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, ,

Methani, N., Ganguly, P., Khapra, M.M., Kumar, P. (2020). Plotqa: Reasoning over scientific plots. *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 1527–1536).

Mukherjee, K., Huey, H., Lu, X., Vinker, Y., Aguina-Kang, R., Shamir, A., Fan, J. (2024). Seva: Leveraging sketches to evaluate alignment between human and machine visual abstraction. *Advances in Neural Information Processing Systems*, *36*, ,

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

Okan, Y., Janssen, E., Galesic, M., Waters, E.A. (2019). Using the short graph literacy scale to predict precursors of health behavior change. *Medical Decision Making*, *39*(3), 183–195,

Padilla, L. (2018). A case for cognitive models in visualization research: Position paper. *2018 ieee evaluation and beyond-methodological approaches for visualization (beliv)* (pp. 69–77).

Padilla, L., 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), 1–25,

Padilla, L., Hosseinpour, H., Fygenson, R., Howell, J., Chunara, R., Bertini, E. (2022). Impact of covid-19 forecast visualizations on pandemic risk perceptions. *Scientific reports*, *12*(1), 2014,

Pandey, S., & Ottley, A. (2023). Mini-vlat: A short and effective measure of visualization literacy. *arXiv preprint arXiv:2304.07905*, ,

Peebles, D., & Cheng, P.C.-H. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human factors*, *45*(1), 28–46,

Peterson, J.C., Bourgin, D.D., Agrawal, M., Reichman, D., Griffiths, T.L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214,

Pinker, S. (1990). A theory of graph comprehension. *Artificial intelligence and the future of testing*, 73–126,

Playfair, W. (1801). *The commercial and political atlas: Representing, by means of stained copper-plate charts, the progress of the commerce, revenues, expenditure and debts of england during the whole of the eighteenth century.* T. Burton.

Preacher, K.J., Zhang, G., Kim, C., Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate behavioral research*, *48*(1), 28–56,

Price, M.M., Crumley-Branyon, J.J., Leidheiser, W.R., Pak, R. (2016). Effects of information visualization on older adults' decision-making performance in a medicare plan selection task: a comparative usability study. *JMIR human factors*, *3*(1), e5106,

Ruginski, I.T., Boone, A.P., Padilla, L.M., Liu, L., Heydari, N., Kramer, H.S., ... Creem-Regehr, S.H. (2016). Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, *16*(2), 154–172,

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464,

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,

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational psychology review*, *14*(1), 47–69,

Solomon, T., Dupuis, A., O'Hara, A., Hockenberry, M.-N., Lam, J., Goco, G., ... Tannock, R. (2019). A cluster-randomized controlled trial of the effectiveness of the jump math program of math instruction for improving elementary math achievement. *PloS one*, *14*(10), e0223049,

Spence, I. (2006). William playfair and the psychology of graphs. *Proceedings of the american statistical association, section on statistical graphics* (pp. 2426–2436).

Tufte, E.R. (1983). *The visual display of quantitative information* (Vol. 2). Graphics press Cheshire, CT.

Uttal, D.H., McKee, K., Simms, N., Hegarty, M., Newcombe, N.S. (2024). How can we best assess spatial skills? practical and conceptual challenges. *Journal of Intelligence*, *12*(1), 8,

Wilkinson, L. (2012). *The grammar of graphics*. Springer.

Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3712–3722).

Zelikman, E., Ma, W.A., Tran, J.E., Yang, D., Yeatman, J.D., Haber, N. (2023). Generating and evaluating tests for k-12 students with language model simulations: A case study on sentence reading efficiency. *arXiv preprint arXiv:2310.06837*, ,

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M.C., DiCarlo, J.J., Yamins, D.L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118,