A Sanity Check of Wide-Spread Data Visualization Literacy Assessments

Adani Abutto
aabutto@stanford.edu
Stanford University
Stanford, CA, USA

Abstract

Data visualization literacy is key in science and STEM education. The present study examined U.S. adult test-takers' performance on five data visualization literacy assessments and investigated what may predict variability in average performance. Results suggested that the variability in performance across test items was predicted by item nature (designed to mislead vs. not designed to mislead) and question length. Additionally, results from an item-response-theoretic analysis indicated that for various test items, even test-takers with low latent ability levels had a considerable probability of obtaining a correct response. However, a model simulating a low-literacy test-taker who follows an exclusion-and-guessing strategy showed bad fit to the observed data. In sum, this evidence supports that beyond just item difficulty, test performance depends on additional 'commonsense' factors, and that the assessments successfully tap an underlying latent data visualization literacy ability, even if the structure and dimensionality of this ability remains unclear.

ACM Reference Format:
Adani Abutto. 2025. A Sanity Check of Wide-Spread Data Visualization Literacy Assessments. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The ability to read, interpret, and communicate the contents of graphs—a capacity often termed 'data visualization literacy'—has become a foundational skill not just for communication among researchers but also students and professionals in many STEM-related domains. However, it remains unclear what exactly constitutes the construct of data visualization literacy, to what extent different assessments capture it effectively, and which (other) factors predict performance on such assessments.

Assessments of data visualization literacy often present the test-taker with a series of items entailing a question and various graphs to answer said question (Brockbank et al., 2024). Widely used examples of such assessments are GGR (Galesic et al., 2013), VLAT (Lee et al., 2016), CALVI (Ge et al., 2023), BRBF (Boy et al., 2018), and WAN (Wainer, 1980). In the present study, I assess whether two 'commonsense' predictors of item difficulty—question length and misleading item nature—predict performance across items contained in these various data visualization literacy assessments. I also draw on Item Response Theory (IRT) tools to further investigate the properties of the assessments' test items and examine how well a model performs that guesses its way through the assessments in the absence of any substantial data visualization literacy. In sum, my approach serves as a "sanity check" on two fronts: 1) Checking whether the test items the designers intended to be more difficult are indeed more difficult as per the subject data, and 2), checking whether the tests are (reasonably) safe from low-literacy test-takers scoring highly through mere chance/guessing.

2 Methods

The recruited sample included $N = 426$ U.S. adult participants. No demographic information was available.

2.1 Data and Measures

Put together, the five administered tests (GGR, VLAT, CALVI, BRBF, WAN) spanned 230 items (178 questions; some questions involved multiple items). Each participant completed all five tests but was assigned only a subset of items from each assessment. All of subjects' individual responses were scored as correct (1) or incorrect (0). For cross-validation purposes, the participant data was split into two samples: Training (80%; 184 items) and test (20%; 46 items). The dataset also contained additional information on each item: Item presentation format (table shown vs. graph shown with item), graph shown with a given item (graph type; e.g., bar plot, line plot), type of task (task category; e.g., make comparisons, make predictions), and item nature (designed to mislead vs. *not* designed to mislead).

2.2 Inferential Statistical Analyses

My inferential statistical analyses assess the (additional) explanatory power of two factors that likely affect subjects' performance: Item nature (whether or not the item was designed to mislead), and question length (the amount of verbal information to process). To examine how much these sources of item-level variation were systematically related to variation in subject performance across items, I ran regression analyses. I then used non-nested model comparison to contrast how each model performed relative to an "empty" model (fitting the mean). More specifically, I fit mixed-effects logistic regression models to regress performance (correct response vs. incorrect response) on item nature (model A) and performance on question length (model B). For model A, I included item nature (categorical) as a fixed effect; for model B, I included question length (continuous) as a fixed effect; for both models, I additionally included random intercepts per participant.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

For each model, I report Chi-squared, degrees of freedom, regression coefficients, standard error (SE), p-values. In the present study, I did not conduct direct between-model contrasts using various model metrics (e.g., AIC, BIC), but computed Root Mean Squared Error (RMSE) for comparability. All of the above was done using the R packages *lme4* (Bates et al., 2003), *ggeffects* (Lüdecke, 2020), *broom.mixed* (Bolker et al., 2019), and *emmeans* (Lenth, 2018).

2.3 IRT Analysis

In addition to the above analyses, I analyzed individual items through an Item Response Theory (IRT) lens. In IRT, item difficulty is defined as the level of latent ability (here data visualization literacy) required for a 50% probability of achieving a correct response on a given item. Harder items require *more* ability to have a 50% probability of getting the item correct; easier items require *less* ability to have a 50% probability of getting the item correct. Building on this notion, I fit a two-parameter logistic (2PL) model where the expected difference in performance (i.e., the probability of getting the correct response to a given item) depends not just on ability level but also the difficulty of the item in question. That is, the 2PL model predicted item-level subject responses based on (hypothesized) latent ability level as well as the item difficulty.

I also evaluated 1) item information and 2) item discrimination for all test items. Item information represents each item's ability to differentiate between test-takers in terms of their latent ability levels (represented by theta). Higher item information is better. Item discrimination represents how strongly related a given test item is to the latent ability in question as assessed by the broader test it is a part of. Thus, items with higher discrimination are better at differentiating test-takers. I plotted both item information and item discrimination using Item Characteristic Curves (ICCs), where the x-axis represents ability level (theta) and the y-axis represents probability of getting a correct response. All analyses were conducted using R packages *ltm* (Rizopoulos, 2005) and *TAM* (Robitzsch, 2013).

2.4 'Guessing Participant' Simulation

To check whether the assessments used here are (reasonably) safe from low-literacy test-takers scoring highly through mere chance, I created a model simulating a participant with low data visualization literacy who pairs an exclusion strategy with guessing. Such a strategy is especially effective for a test-taker who does not actually possess the tested ability and thus does not know the answer to (most) administered items. If a test allows for high performance based on such a strategy, it will fail to effectively tap the latent ability of interest; in other word, the variability in item-level performance would no longer depend on (just) the construct in question.

I created separate simulations for binary items and multiple-choice (MC) items, since guessing a correct response is harder (less likely) when the number of response options is larger. For binary items, I used a 2PL model as described in the previous section. Based on the fitted 2PL model, I simulated a participant with a latent ability (or theta) of 1 SD below average (i.e., theta = -1). This effectively represents a test-taker low in data visualization literacy. On this basis, I then generated predictions in expected performance (% correct) for each item. Given this approach, for any item with an item difficulty around −1 (i.e., an "easier" item), the modeled student had about a 50% chance of obtaining the correct response, which matches the probability of getting a correct response when choosing between two response options at random: $P(\text{correct} \mid \theta = -1) = \text{logit}^{-1}\big(a(\theta - b)\big) = \text{logit}^{-1}(a(-1 - (-1)) = \text{logit}^{-1}(a(0)) = 0.50$, where b = item difficulty, and a = item discrimination. For easier items (i.e., b < −1), the model let the testtaker do better than chance (i.e., P(correct|θ = -1) > 50%), and for harder items (i.e., b > −1) it let them do worse (i.e., P(correct|θ = -1) < 50%).

For MC items, the modeled student followed a slightly more sophisticated guessing-and-exclusion strategy: They first eliminated up to 75% of the response options (with a 50% chance of eliminating any given option), and then guessed among the remaining options. Using these parameters, the model then computed an average predicted % correct for each item by simulating across 426 participants who all implemented the strategy outlined above. The details of the parameters and simulation code are available in the GitHub repository: https://github.com/adaniabutto/datasci294l_mini-project-1/tree/main/code.

3 Results

3.1 Descriptive Statistics

The average proportion of correct responses across all tests was 62.9%, with an average standard deviation (SD) of 27.4%. The table below shows means, SDs, min % correct, and maximum % correct per test (i.e., collapsed across items within a given test).

Table 1: Means, SDs, Min, and Max for % correct by test

Test	Mean	SD	Min	Max
BRBF	0.699	0.214	0.200	0.988
CALVI	0.435	0.285	0.012	0.964
GGR	0.620	0.291	0.129	0.943
VLAT	0.657	0.259	0.072	0.988
WAN	0.795	0.192	0.227	0.989

Figure 1 gives a visual overview of all item difficulty faceted by test and task type. Items are grouped by test and sorted by item difficulty in ascending order. Overall, we see that even within a single test, the items varied considerably in their difficulty (as measured by % correct). Notably, on some items, as few as 1% of participants selected the correct response (very difficult items), and on other items, as many as 99% of participants selected the correct response (very easy items). In sum, this suggests that the test-takers were neither at ceiling nor at floor on our five assessments, offering considerable individual variability to explain.

3.2 Item Difficulty and Item Nature

Next, I turned to assessing what predicted this observed item-level variability in proportion correct. To examine the predictive power of item nature (designed to mislead vs. not designed to mislead), I examined the subset of subject data from CALVI (4,082 responses from N = 425 participants), as this was the only assessment that included misleading items. I predicted that, on average, misleading items would yield a lower proportion correct (representing higher

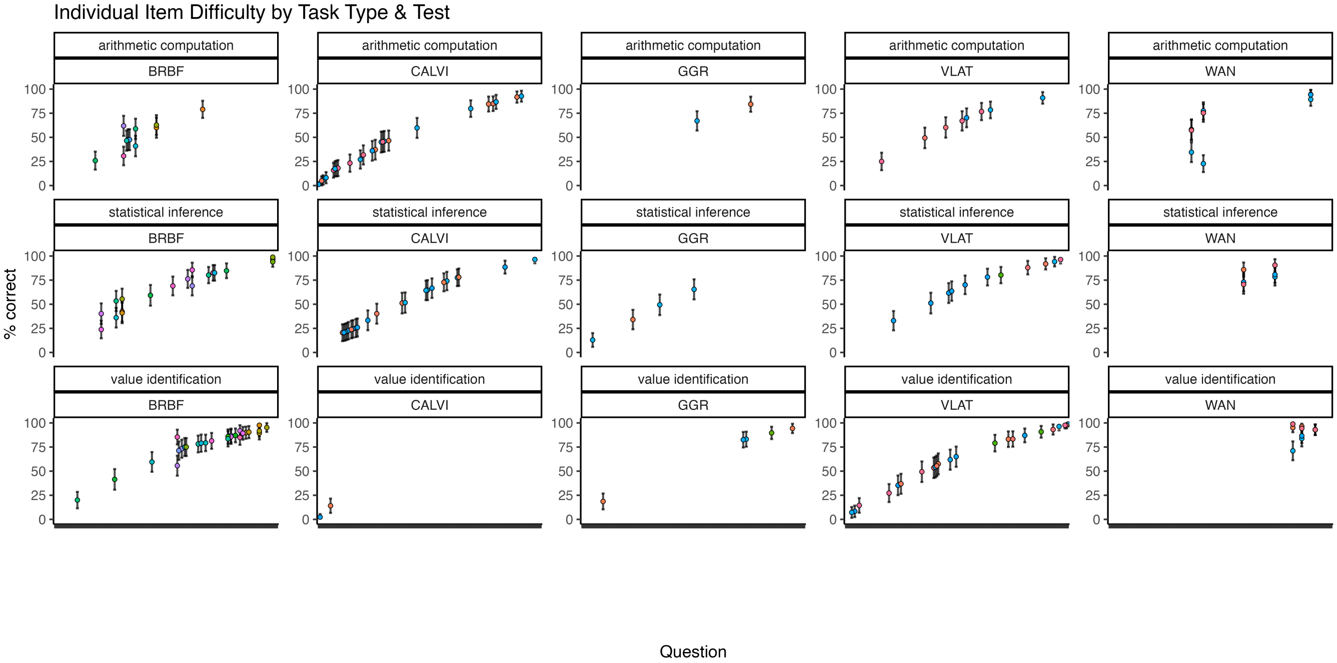

Figure 1: Individual item difficulty by task type and test

item difficulty) than non-misleading items. The results of the mixed-effects logistic regression supported this: The predicted % correct for non-misleading items was about twice as large (70%; 95% CIs = [67%, 73%]) as % correct for misleading items (34%; 95% CIs = [32%, 36%]). Compared to a minimal model that fits the mean regardless of item nature (misleading vs. not misleading), the model including item nature improved fit to data: χ^2 = 356.6, df = 1, β = -1.51, SE = .080, p < .001; RMSE.model = .463 vs RMSE.baseline = 0.496.

3.3 Item Difficulty and Question Length

Secondly, I examined whether question length helped explain the observed variability in proportion correct. To this end, I drew on the data from all tests (15,616 responses from N = 426 participants). There was some variation in question length both within and across tests, with the average question length ranging from 48 to 95 characters depending on test. I predicted that items with greater question length would tend to yield a lower proportion correct (representing higher item difficulty) than shorter questions. The results of the mixed-effects logistic regression supported this: χ^2 = 430, df = 1, standardized β = -.37, SE = .0001, p < .001; RMSE.model = .459 vs RMSE.baseline = 0.483.

3.4 IRT Analysis

Figure 2 shows the Item Characteristic Curves (ICCs) faceted by test. Altogether, items differed quite widely in their item information and item discrimination. For some items, item difficulty was high, meaning that even with high levels of ability, respondents had a low (estimated) probability of selecting the correct answer. For other items, item difficulty was low, meaning that respondents low on data visualization literacy still had a decent (estimated) probability of choosing the correct answer. For some items, ICCs were uninterpretable: Their fitted item response function suggested that, as ability level increased, respondents had a lower probability of selecting the correct answer. The 2PL model did not converge for the BRBF data subset; these ICCs should be interpreted with caution.

3.5 'Guessing Participant' Simulation

The item discrimination and item information analyses suggested that even students with relatively low data visualization literacy had a non-negligible chance of getting the correct response on various items. Given this, I next ran my model representing a student with low data visualization literacy who attempted to get correct answers by using an exclusion-and-guessing strategy. Running a simple bivariate correlation analysis showed that the model predictions were significantly but only weakly correlated with the observed data: $r(200)$ = .22, 95% CIs = [.09, .35], p = .002. Notably, evaluating the model's RMSE shows that the low-literacy guessing model performed worse than a baseline model that fit just the mean: RMSE.model = 35.17 vs RMSE.baseline = 27.22. This suggests that the guessing model was not a good explanation for the observed item-level variability. In other words, is highly unlikely that the high-performing respondents achieved their correct responses via (just) guessing. These results also held true for predicting the held-out data. Figure 3 visualizes the above results.

4 Discussion

Assessments of data visualization literacy often present the test-taker with a series of items entailing a question and various graphs to answer said question (Brockbank et al., 2024). Here, I examined if

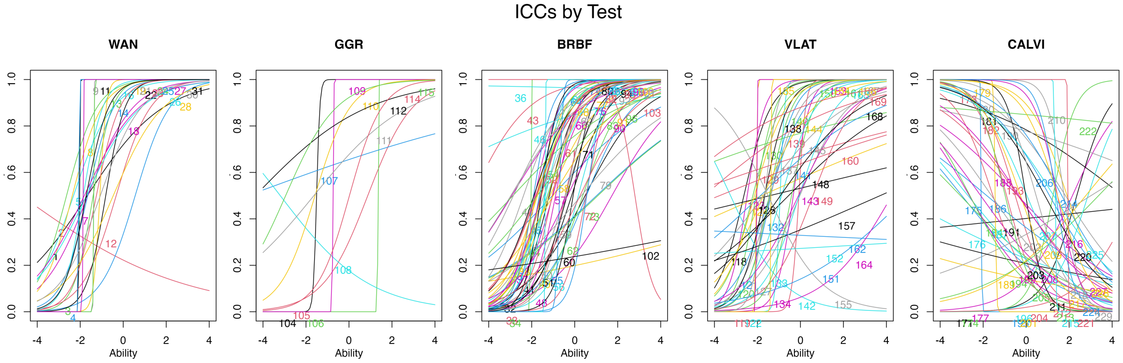

Figure 2: ICC Curves by Test

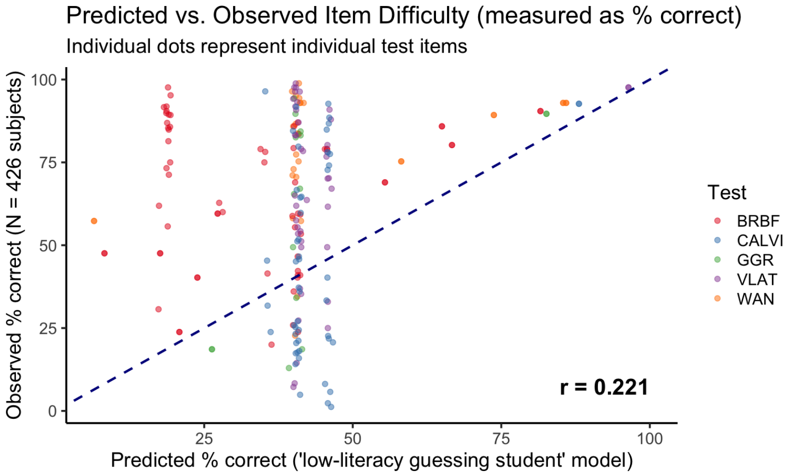

Figure 3: Comparison of Model Predictions and Observed Data

five widely-used data visualization literacy assessments pass a "sanity check" on two fronts: 1) Are the test items the designers intended to be more difficult indeed more difficult as per the test-takers' performance? And 2), are the tests (reasonably) safe from low-literacy test-takers scoring highly through mere chance/guessing? The answer to both questions was yes: Two common-sense predictors of item difficulty—question length and item nature—predicted item-level performance. Additionally, my model simulating a low-literacy subject who followed a exclusion-and-guessing strategy did not align well with the observed data.

4.1 Limitations and Future Directions

The present approach was limited in various ways. At the analysis and modeling level, I did not perform nested model comparisons or mixed-effects multiple linear regressions to assess the predictive power of my 'common-sense' predictors while holding the other predictor(s) constant, including predictors found to be significant by previous work (e.g., graph type and task type of the item; cf. Brockbank et al., 2024). Moreover, my 'exclusion-and-guessing' model had hard-coded parameters that were not empirically derived but rather a best guess for what a true test-taker with such a strategy could plausibly be characterized as. Future research should look into an empirical approach to setting these parameters to increase the robustness of this 'sanity check.'

Secondly, for the IRT analyses, I exclusively fit a 2PL model. Various literature discusses the advantages and drawbacks of such a model compared to a more traditional 1PL (Rasch) model and multi-parameter models of other kinds. Future research should fit a variety of models and conduct nested model comparisons to assess which one best captures the structure underlying the data on the data visualization literacy assessments used here.

Finally, these results are limited in generalizability. The participants in the dataset used here were recruited from a specific U.S. population, and each subject completed only a subset of the items from the five assessments. With this, all estimates and item parameter calibrations rest on a convenience sample and a partial item set. It is unclear whether and how the present findings apply to other age groups, cultures, languages, and how similar items from further assessments would behave psychometrically. Future work should expand its scope accordingly.

5 Code and Data Availability

All analysis scripts are publicly available in the following GitHub repository: https://github.com/adaniabutto/datasci294l_mini-project-1/tree/main/code.

References

[1] Margaret Wu Alexander Robitzsch, Thomas Kiefer. 2013. TAM: Test Analysis Modules. doi:10.32614/CRAN.package.TAM Institution: Comprehensive R Archive Network Pages: 4.2-21.

[2] Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. 2003. lme4: Linear Mixed-Effects Models using 'Eigen' and S4. doi:10.32614/CRAN.package.lme4

[3] Ben Bolker, David Robinson, Dieter Menne, Jonah Gabry, and Paul Buerkner. 2019. Package "broom. mixed". (2019).

[4] Jeremy Boy, Ronald A. Rensink, Enrico Bertini, and Jean-Daniel Fekete. 2014. A Principled Way of Assessing Visualization Literacy. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 1963–1972. doi:10.1109/TVCG.2014.2346984

[5] Mirta Galesic and Rocio Garcia-Retamero. 2011. Graph Literacy: A Cross-Cultural Comparison. *Medical Decision Making* 31, 3 (May 2011), 444–457. https://doi.org/10.1177/0272989X10373805

[6] Lily W. Ge, Yuan Cui, and Matthew Kay. 2023. CALVI: Critical Thinking Assessment for Literacy in Visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. doi:10.1145/3544548.3581406

[7] Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. 2017. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan. 2017), 551–560. doi:10.1109/TVCG.2016.2598920

[8] Dimitris Rizopoulos. 2005. ltm: Latent Trait Models under IRT. doi:10.32614/CRAN.package.ltm Institution: Comprehensive R Archive Network Pages: 1.2-0.