



NORTHWESTERN UNIVERSITY

Computer Science Department

Technical Report
Number: NU-CS-2026-40

September, 2026

Toward a More Comprehensive Understanding of Visualization Literacy Through Designing and Developing Measurements and Interventions

Lily Wanqian Ge

Abstract

Visualizations are highly effective at communicating data-relevant information. However, they can also exacerbate misinformation if not appropriately designed, making visualization literacy (i.e., the ability to both accurately interpret and create well-formed visualizations) essential for informed decision-making in today's society. Existing studies of visualization literacy largely focus on low-level readings of well-formed visualizations, which are not representative of what it means to be literate with visualizations in a world where not all information is well formed.

I advance the understanding of visualization literacy by developing enhanced measurements of higher-level skills (e.g., critical thinking, visual mapping), examining how visualization literacy measurements are designed and used, and grounding educational interventions in this broadened understanding. To address the specific challenges, such as how to reliably capture complex cognitive abilities, how to evaluate open-ended tasks with vast solution spaces, and how to scale assessment beyond manual expert judgment, I (1) develop a Critical Thinking Assessment for Literacy in Visualizations (CALVI), a validated measurement that tests people's susceptibility to misleading visualizations, (2) design an educational game with high school students for teaching critical thinking in visualization interpretation, (3) develop an Assessment for the Visual Encoding Ability in Visualization Construction (AVEC), a validated measurement that tests people's ability to visually encode data for creating well-formed visualizations, and (4) investigate factors that make visualization literacy hard to operationalize through conducting an

autoethnography with users and designers of assessments. My thesis clarifies what constitutes visualization literacy and outlines concrete steps for the research community to systematically advance this understanding.

Keywords

Visualization literacy, Measurement, Critical thinking, Human-data interaction, Human-computer interaction

NORTHWESTERN UNIVERSITY

Toward a More Comprehensive Understanding of Visualization Literacy Through
Designing and Developing Measurements and Interventions

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Lily Wanqian Ge

EVANSTON, ILLINOIS

September 2026

© Copyright by Lily Wanqian Ge, 2026
All Rights Reserved

Abstract

Visualizations are highly effective at communicating data-relevant information. However, they can also exacerbate misinformation if not appropriately designed, making visualization literacy (i.e., the ability to both accurately interpret and create well-formed visualizations) essential for informed decision-making in today's society. Existing studies of visualization literacy largely focus on low-level readings of well-formed visualizations, which are not representative of what it means to be literate with visualizations in a world where not all information is well formed.

I advance the understanding of visualization literacy by developing enhanced measurements of higher-level skills (e.g., critical thinking, visual mapping), examining how visualization literacy measurements are designed and used, and grounding educational interventions in this broadened understanding. To address the specific challenges, such as how to reliably capture complex cognitive abilities, how to evaluate open-ended tasks with vast solution spaces, and how to scale assessment beyond manual expert judgment, I (1) develop a Critical Thinking Assessment for Literacy in Visualizations (CALVI), a validated measurement that tests people's susceptibility to misleading visualizations, (2) design an educational game with high school students for teaching critical thinking in visualization interpretation, (3) develop an Assessment for the Visual Encoding Ability in Visualization Construction (AVEC), a validated measurement that tests people's ability to visually encode data for creating well-formed visualizations, and (4) investigate factors that make visualization literacy hard to operationalize through conducting an autoethnography with users and designers of assessments. My thesis clarifies what constitutes visualization literacy and outlines concrete steps for the research community to systematically advance this understanding.

Acknowledgment

I have always been a person who values the journey more than the destination (and what a journey it has been). I have never doubted that I would get here today. Whatever confidence I have is a gift from those who helped and believed in me along the way.

I am deeply grateful to my advisor, Matthew Kay, for giving me the space to find my own direction and generously offering his time and resources when I needed them. Matt has shown me what high-quality research looks like, never let me settle for a good-enough answer, and expected more of my work than I sometimes thought it had in it. The trust he has placed in me over the years has meant a great deal, and did much to build my confidence as a researcher. Many of the opportunities I have had would not have existed without his support, whether directly or in ways I likely never saw. Through detailed, often Socratic feedback, words of encouragement, and steady patience, he not only shaped how I approach research, but also how I approach life.

I would also like to thank my collaborators, who walked parts of this road with me. It has been truly satisfying to work and grow alongside Charles Cui, with whom I have collaborated most over the years. His drive for impact and his discipline have influenced me well beyond research. I am especially grateful to those who have been supporting me since the beginning of this journey. Steve Franconeri was the first faculty member I collaborated closely with other than Matt, and he practically acted as a co-advisor to me when I was just starting out. His exceptional communication skills made me realize what it takes to become a well-rounded researcher. Lane Harrison (Worcester Polytechnic Institute) was the first collaborator I had outside of Northwestern, and he has always been ready to offer his full support whenever I needed it, whether with research or career advice. Fumeng Yang joined Northwestern as a postdoc around the same time I started as

a Ph.D. student, and has since been someone I can always go to for feedback and conversation, big or small. I am also immensely grateful to my thesis committee members, Mike Horn, Marcelo Worsley, and Petra Isenberg, for their time and insight, which helped shape my dissertation beyond its individual projects. More broadly, I have been fortunate to work and exchange ideas with many others, across Northwestern and beyond, each of whom has contributed to my growth as a researcher in ways I will carry with me well beyond my Ph.D.

I feel extremely fortunate to have had support across many areas within and outside of Northwestern. Within Northwestern, the lab members I shared these years with (Mandi Cai, Ziyang Guo, Maryam Hedayati, Alex Kale, Paula Kayongo, Hyeok Kim, Tae-wook Kim, Sheng Long, Priyanka Nanayakkara, Xiaoying Pu, Emily Qiao, Abhraneel Sarma, Dongping Zhang) have offered immeasurable support and have been such an important part of my time here. The Computer Science staff members, including Katie Winters, Jensen Smith, Ethan Walles, Bella Barrios, Dru Redmond, Melissa Duong, and Madeleine Agaton, have done important work that made my life operate smoothly as a grad student. The student organizations I was a part of (CSSI, GWiC, CSPAC) contributed immensely to my wellbeing throughout grad school, and have been a source of lifelong friends. Anastasia Kurdia, Sara Sood, and other teaching faculty have also been very helpful during the latter portion of my Ph.D. for advice on teaching and mentoring. Outside of Northwestern, I would like to extend my gratitude to the members of the G-RIPS (Graduate-level Research in Industrial Projects for Students) program, Tohoku University, and Fujitsu, including Hiroshi Suito and Hiroyuki Higuchi, for their support and hospitality during my visits, where I grew both personally and professionally. Over the years, I have been fortunate to meet people in the visualization and HCI communities who offered advice at various points during my time in graduate school, including Cindy Xiong Bearfield, Alex

Kale, Lace Padilla, Evan Peck, Arvind Satyanarayan, Emily Wall, and Rebecca Williams. I also extend my gratitude to my friends, old and new, who have enabled me to grow into the person I am today.

I am fortunate to have received funding from the NSF Graduate Research Fellowship Program, Northwestern Computer Science department, and Northwestern HCI + Design Cluster Fellowship during the course of my Ph.D. Their support allowed me to pursue my passion, define my research path, and create a vision of my own.

Above all, I am grateful to my family, who have been there for me far before the start of my Ph.D. journey. Without them, none of this would even have been possible, so, from the bottom of my heart, thank you.

Table of Contents

Abstract	3
Acknowledgment	4
Table of Contents	7
1 Introduction	10
1.1 Thesis Background and Motivation	10
1.2 Methodological Overview	12
1.3 Thesis Contributions	12
2 Related Work	17
2.1 Visualization Literacy	17
2.2 Visualization Misinformation	19
2.3 Visualization Construction	20
2.4 Systematic Procedure for Test Development	21
3 CALVI: Critical Thinking Assessment for Literacy in Visualizations	23
3.1 Introduction	23
3.2 Test Conceptualization and Construction	26

3.3	Test Tryout	42
3.4	Item Analysis	44
3.5	Test Revision	51
3.6	Evidence of Validity	54
3.7	Finalized Bank and Reliability	56
3.8	Discussion	58
4	Inoculating Against Visualization Misinformation Through Gamification	64
4.1	Introduction	64
4.2	Related Work	66
4.3	Methods	69
4.4	Design Variants	71
4.5	Synthesis of Game Features and Design Implications	79
4.6	Discussion	81
5	AVEC: An Assessment of Visual Encoding Ability in Visualization Construction	84
5.1	Introduction	84
5.2	Related Work	87
5.3	Test Development: Items and Assessment Tool	88
5.4	Test Development: Visualization Answers	97
5.5	Expert Panel	101
5.6	Test Tryout	108
5.7	Discussion	116
6	An Autoethnography on Visualization Literacy: A Wicked Measurement Problem	121
6.1	Introduction	121

6.2	Related Work	123
6.3	Methods	126
6.4	Conceptual: What Exactly are We Measuring?	130
6.5	Operational: How Do We Use Existing Tests?	141
6.6	Methodological: How Do We Design Tests?	148
6.7	Discussion	157
6.8	Next Steps	160
7	Discussion	161
7.1	Assessment Tools for Complex Visualization Skills	161
7.2	How Are People Using CALVI?	162
7.3	Measurements for Understanding Versus Measurements for Evaluation .	165
7.4	Reflections and Future Directions	166
8	Conclusion	168
	References	169

Chapter 1

Introduction

Visualizations are often used to communicate high-stakes information to the public, such as through news outlets and social media. Given the wide variety of visual representations of data, it is important to know how well people can understand them. In particular, I was motivated by the fact that there is a lot more to visualization understanding than just *reading* values off a chart. Also interested in how different people might interpret the same visualization differently, I embarked on the journey of *visualization literacy*.

1.1 Thesis Background and Motivation

A person's ability to interpret and create visualizations is generally referred to as their *visualization literacy*. It comprises many skills, including extracting information from visualizations [106], identifying deception [33], and creating visualizations [23], and is essential for people to make informed decisions from data in today's information landscape. As illustrated by the visualization reference model from Card et al. [34], creating a visualization often involves a designer first processing the raw data, such as apply data transformations (e.g., aggregate, normalize, bin), before encoding data to visual channels (e.g., *x*-axis, *y*-axis, color, size) to answer a data-relevant question. Viewers then must decode

and form insights from the visualization, which often requires them to make judgments such as evaluating its quality.

Despite the variety of skills needed to master data visualization, many studies of visualization literacy largely focus on the *low-level reading* of visualizations (e.g., extracting individual values from charts), and more specifically, *well-formed* ones. This assumption not only simplifies visualization literacy to overly focus on interpretation, but also is not representative of the pool of visualizations in the real world, which is not guaranteed to be correct (e.g., [102, 52]). Research in visualization literacy must move beyond the focus on interpreting well-formed visualizations and seek to understand how we can reason about *ill-formed* visualizations and *create* well-formed ones.

My thesis work challenges the assumption that low-level readings of well-formed visualizations are representative of what constitutes visualization literacy. Through the design and development of new assessments that target higher-level skills, the development of an educational game for teaching critical thinking in visualization interpretation, and an autoethnographic study with users and designers of visualization literacy assessments, I draw a broader picture of what visualization literacy should encompass, how we can study the construct through enhanced measurements, and how we can approach developing engaging educational interventions for teaching the critical thinking aspect of visualization literacy. The findings throughout this dissertation can be summarized into the following thesis statement:

Broadening the skills and ways of measuring both the interpretation and creation of visualizations can better characterize and pinpoint challenges in visualization literacy, which in turn can also support the development of targeted interventions to help people overcome those challenges.

1.2 Methodological Overview

To study visualization literacy, I take a bottom-up approach rather than attempting to propose a single, unified model of literacy. Focusing on one component at a time lets me define it precisely enough to build a valid assessment or a targeted intervention around it, rather than around a broad or loosely specified notion of the skill. My approach to studying visualization literacy includes both qualitative and quantitative methods, such as drawing on psychometric testing methods (e.g., [44, 131]) to develop and validate assessments, and learning sciences (e.g., [99]) to ground the design of interventions in how people learn.

1.3 Thesis Contributions

This dissertation is composed of four projects across four chapters. I briefly describe each project and its contributions to the overall thesis.

1.3.1 CALVI: Critical Thinking Assessment for Literacy in Visualizations (Chapter 3)

This chapter is based on work published in ACM CHI 2023 [72] and is done in collaboration with Yuan Cui and Matthew Kay.

To broaden visualization literacy to include the critical thinking aspect of visualization literacy, we first define it to be “the ability to read, interpret, and reason about erroneous or potentially misleading visualizations”. We designed and validated CALVI, a Critical Thinking Assessment for Literacy in Visualizations to quantify this ability by: (1)

developing a precise definition of misleaders (decisions made in the construction of visualizations that can lead to conclusions not supported by the data), (2) constructing initial test items using a design space of misleaders and chart types, (3) trying out the provisional test on 497 participants, and (4) analyzing the test tryout results and refining the items using Item Response Theory, qualitative analysis, a wrong-due-to-misleader score, and the content validity index. CALVI consists of a bank of 45 multiple-choice questions that can be customized depending on the desired difficulty levels. The misleaders that are not solely visual distortions seems harder for people to detect, such as *Overplotting*, *Missing Data*, and *Missing Normalization*. CALVI broadens visualization literacy to higher-level skills such as critical thinking and offers researchers and educators an instrument that can assess critical visualization interpretation, allowing for the systematic study of misleading visualizations.

1.3.2 Inoculating Against Visualization Misinformation Through Gamification (Chapter 4)

This work is a working paper and is done in collaboration with Michael S Horn and Matthew Kay.

Measurements are essential for understanding the construct itself, but it is also important to improve visualization literacy through designing targeted interventions. Focusing on visualization misinformation, we designed an educational game, Newsroom, to help high school students build “immunity” against misleading visualizations. One benefit of CALVI is that it offered a clearer picture of the problem space of misleaders, which we used as a basis for designing our game. We ran a series of design workshops with high school students to test out design variants and receive feedback to iteratively

improve the game. We tried out individual, competitive, team-based, and competitive + team-based game designs. Through multiple rounds of refinement and playtests, we found that competitive and team-based designs were more engaging to students. From the design workshops, we reflect on the game features that worked especially well for our game for teaching the critical interpretation of data visualizations, such as allowing players to have secret agendas to build an evaluative mindset, demonstrating concepts concretely with visual illustrations, letting players engage in the creation process to elicit deeper reasoning.

1.3.3 AVEC: An Assessment of Visual Encoding Ability in Visualization Construction (Chapter 5)

This chapter is based on work published in ACM CHI 2025 [71] and is done in collaboration with Yuan Cui and Matthew Kay.

Visualization literacy as a construct is not just about interpretation. Skills involved in visualization creation are also important for understanding visualization literacy but are rarely studied systemically due to the lack of measurements available on the creation side. This gap hinders efforts in understanding and improving literacy in visualizations. To address this, we designed and developed AVEC, an Assessment of Visual Encoding Ability in Visualization Construction. We define the visual encoding ability to be the “ability to appropriately map data to visual channel(s) for effectively answering data-relevant questions”, which is a core component of the larger process of visualization construction. We developed AVEC by (1) creating an initial item bank using a design space of visualization tasks and chart types, (2) designing an assessment tool to support the combinatorial nature of selecting appropriate visual encodings, (3) building an autograder from expert

scores of answers to our items, and (4) refining and validating the item bank and auto-grader through an analysis of test tryout data with 95 participants and feedback from the expert panel. AVEC broadened visualization literacy again by enabling the systematic study of the design side of the construct.

1.3.4 An Autoethnography on Visualization Literacy: A Wicked Measurement Problem (Chapter 6)

This chapter is based on work published in IEEE VIS 2025 [73] and is done in collaboration with Anne-Flore Cabouat, Karen Bonilla, Yuan Cui, Yiren Ding, Noëlle Rakotondravony, Mackenzie Michael Creamer, Jasmine Tan Otto, Maryam Hedayati, Bum Chul Kwon, Angela Locoro, Lane Harrison, Petra Isenberg, Michael Correll, and Matthew Kay.

Although CALVI and AVEC offered ways to characterize additional components of visualization literacy, we are still far from solving the visualization literacy measurement problem due to reasons such as the complexity of the construct itself. We need more efforts from the research community to tackle this measurement problem. To surface some factors that contribute to the complexity of defining and measuring visualization literacy, I led a group of researchers who are users and designers of visualization literacy tests to conduct an autoethnography. This is a qualitative method to expose our tacit thoughts on what makes visualization literacy measurement a “wicked” problem. In this autoethnography, we gather insights from 14 internal interviews, reflect on the fluidity of visualization literacy, and discuss how this property may lead to misalignment between what the construct is and how measurements of it are used or designed. We also examine potential threats to measurement validity from conceptual, operational, and methodological perspectives. Based on our experiences and reflections, we propose several calls to ac-

tion aimed at tackling the wicked problem of visualization literacy measurement, such as by broadening test scopes and modalities, improving test ecological validity, making it easier to use tests, seeking interdisciplinary collaboration, and drawing from continued dialogue on visualization literacy to expect and be more comfortable with its fluidity.

Chapter 2

Related Work

I describe related work in the visualization literacy space, including prior definitions and assessments. Many of such definitions and assessments focus on the interpretation of well-formed charts, which overlook higher-level skills such as whether people can critically interpret a chart or whether people can construct a well-formed chart. I use visualization misinformation as a way to study the critical thinking aspect of visualization interpretation, which I detail in Section 2.2. I then review existing work on visualization construction and highlight the lack of assessments to systematically study such skills.

2.1 Visualization Literacy

There is no shortage of existing definitions of visualization literacy in the research community. For instance, Boy et al. [26] defined visualization literacy as “the ability to confidently use a given data visualization to translate questions specified in the data domain into visual queries in the visual domain, as well as interpreting visual patterns in the visual domain as properties in the data domain”. Focusing on the interpretation of visualizations, Börner et al. [24] defined literacy in visualizations as “the ability to make meaning from and interpret patterns, trends, and correlations in visual representations

of data”. Refining and building off of these definitions, Lee et al. [106] proposed to define it as “the ability and skill to read and interpret visually represented data in and to extract information from data visualizations”.

Based on these definitions, researchers have developed several assessments to test for people’s ability to interpret visualizations. Boy et al. [26], for example, devised a way of using Item Response Theory (IRT) to design assessments of interpreting basic charts such as line charts, bar charts, and scatterplots. Lee et al. [106] systematically developed a 53-item Visualization Literacy Assessment Test (VLAT) that measures people’s ability to perform basic visualization interpretation tasks, such as retrieve value or make comparisons. To improve upon VLAT, Pandey and Ottley [130] developed a shortened version of VLAT, MINI-VLAT, with 12 items. These assessments focus on the interpretation of well-formed visualizations and the low-level visualization tasks [7], despite the importance of higher-level tasks, such as identifying deception [33] or visualization creation [23]. Without valid measurements for these higher-level visualization tasks, there lack the prerequisite for really understanding and improving abilities related to the more complex tasks like critically interpreting and constructing visualizations.

Realizing this gap, I led the organization of a visualization literacy workshop at ACM CHI 2024, where we discussed topics including how to design assessments and interventions of visualization literacy. During this visualization literacy workshop, we called for the re-conceptualization of visualization literacy by including higher-level abilities, like visualization construction. A more comprehensive understanding of visualization literacy requires deeper investigations of this construct. This includes examining the more complex, higher-level tasks in *both* interpretation and construction, such as critically interpreting misleading visualizations and creating well-formed charts.

2.2 Visualization Misinformation

I use visualization misinformation as a lens to study the critical thinking aspect of visualization literacy. In general, misinformation can happen as a result of carelessness during communication, but it could also come from deliberate manipulation. More specifically, several aspects of visualization misinformation have been studied previously. For instance, some have looked at line charts and addressed deception in line charts by adding annotations [65]. Others have studied misalignment between a visualization and its title and how this impacts trust and recall of information [98]. When visualizations violate important design principles, they can become hallucinators (i.e., different representations of data causing a different impression) or confusers (i.e., ambiguous visualizations such that changes to the data makes no difference in the visual representation) [97]. McNutt et al. studied visualization mirages, which are visualizations that might appear to support a certain claim, but actually are not upon a closer examination [121]. McNutt et al. also compiled categories of errors that can lead to visualization mirages, such as missing records in the data curating process, overplotting, concealing uncertainty, or manipulating scales in the visualization phase [121]. Using a different set of visualizations as the corpus, Lo et al. developed a taxonomy of misinformative visualizations that included errors or issues that could be present on poorly designed visualizations [113]. Previous work on visualization misinformation has informed us ways of manipulation that can result in misleading visualizations. However, more work is required to understand how well the public can identify and interpret misleading visualizations, which is a question that can inform us of the critical thinking aspect of visualization literacy (Chapter 3).

2.3 Visualization Construction

Another major component of visualization literacy is visualization construction, and it is a complex process. Previous work has studied how people create visualizations [89] and how to improve the ability to create visualizations. For instance, both Alper et al. [5] and Gäbler et al. [69] developed interventions to help young children learn about bar charts; He and Adar [83] created VizItCards, a toolkit to facilitate collaborative visualization design in classroom instruction; Bishop et al. [21] developed Construct-A-Vis, a tool that scaffolds the visual mapping process and supports children in free-form visualization creation; and Adar and Lee-Robbins [1] created Roboviz, a game-based activity for visualization courses to facilitate the creation of interactive visualizations. However, limited work has focused on *measuring* how well people can construct visualizations, which requires valid measurements of construction-related abilities. Such measurements can help identify areas of improvement in novice learners and evaluate the effectiveness of proposed interventions.

To study visualization construction, I draw on the visualization reference model by Card et al. [34]. The reference model has been extended, adapted, and refined in prior work [42, 35, 152, 92]. The model presents a process of mapping data to a visual form that supports human sense-making. Before the viewers make sense of the visualization, a sequence of steps are involved in this process, including data transformations, visual mappings, and view transformation. I focus on the core of this model, which is the **visual mapping** transformation: the mapping from data to visual representation. Visual mappings involve encoding data variables to appropriate visual structures, such as visual marks (e.g., point, line) and visual channels (e.g., position, color, size). The mapping of data to visual structures are governed by established principles such as the ranking of

perceptual tasks [116], effectiveness and expressiveness criteria [116], and the Grammar of Graphics [163]. I draw on these principles to develop a systematic way of measuring the visual mapping component of visualization construction (Chapter 5).

2.4 Systematic Procedure for Test Development

A major component of my thesis is on the design and development of assessments to capture a latent ability. I employ an established way for test development, which is outlined in Psychological Testing and Assessment [44]. This is a method that prior work [106] in the visualization community has also adopted. There are five core stages, which I outline below with its core considerations.

The **Test Conceptualization** phase lays the foundation of the test by identifying a need for it and considering questions such as what the test intends to measure, what contents should be included in the test, who will be administering and taking the test, and how the test will be administered [44]. An *item* is a question in the test, and the main tasks in **Test Construction** include deciding the format of the items and designing the items. During **Test Tryout**, the developer tries out the test on a sample of the target audience [44]. The necessary sample size depends on the analysis method planned for the next phase, **Item Analysis**, where data from test tryout will be used to evaluate how well different questions separate participants of different ability levels [44]. Classical test theory (CTT) and IRT can both be used for item analysis, and both methods have been used in prior work for measuring visualization literacy (e.g., Lee et al. [106] used CTT; Boy et al. [26] used IRT). Depending on the IRT model, different samples sizes are recommended [59]. We opted for IRT because it can be extended further to develop computer adaptive testing (CAT) [59, 22]. We selected the 2-parameter logistic (2PL) model to conduct item

analysis, because we are most interested in the item easiness and item discrimination parameters. These parameters will be used in the **Test Revision** phase to determine whether certain items should be removed or rewritten. There are no standard ways of revising the test; developers must decide on the criteria and revise based on the aim of the test [44].

Chapter 3

CALVI: Critical Thinking Assessment for Literacy in Visualizations

This chapter is based on work published in ACM CHI 2023 [72] and is done in collaboration with Yuan Cui and Matthew Kay.

3.1 Introduction

Visualizations are constantly in the public eye: for example, news and media outlets often use visualizations to illustrate and support their claims [18, 151, 161]. While visualizations are capable of effectively conveying data-driven information, they can exacerbate the spread of misinformation if not designed appropriately. In 2014, a controversial visualization used an inverted y -axis to show the number of gun deaths in Florida after introducing the “Stand Your Ground” law (shown in Figure 3.1.A) [102]. Reading such visualizations can lead to a completely opposite understanding of the data. Fast forward to 2020, the Georgia Department of Public Health published a bar chart that presented COVID-19 cases for five counties (Figure 3.1.B), but the bars were sorted from highest to lowest, giving an impression that the cases were going down over time if one did not notice the dates on the x -axis [118]. Unfortunately, there is an abundance of real-world examples of

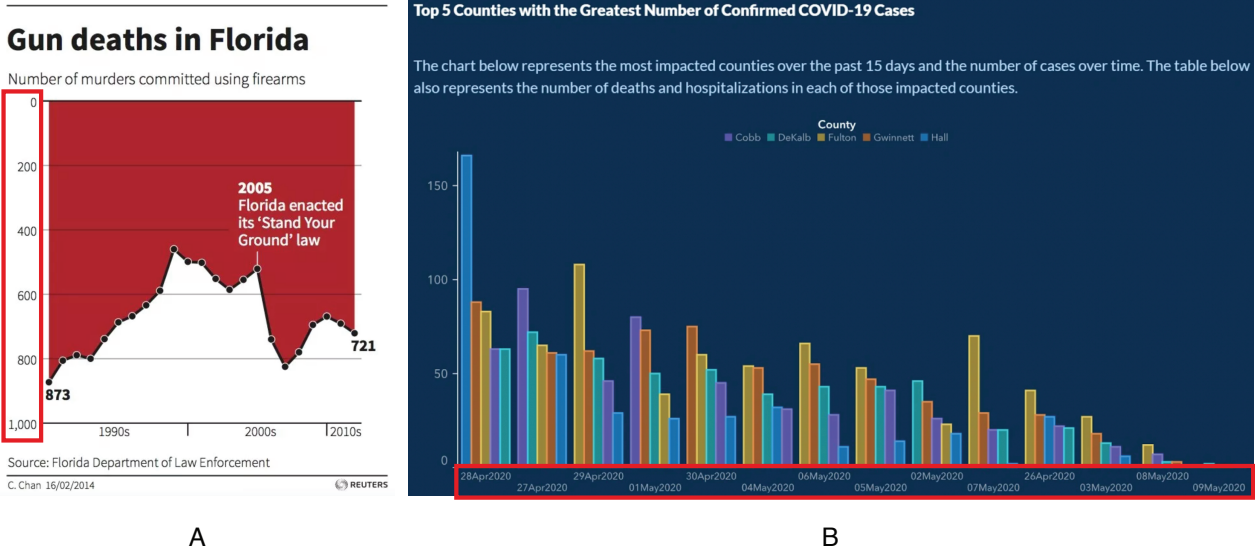


Figure 3.1: Real-world examples of visualization misinformation. The misleader in these examples is *Manipulation of Scales - Unconventional Scale Directions*. (A) The inverted y-axis gives a visual impression that gun deaths went down drastically after the “Stand Your Ground” law in 2005, when they actually increased [102]. (B) From a COVID-19 chart published by the Georgia Department of Public Health, the shuffled dates on the x -axis due to sorting the bars from highest to lowest seem to suggest that confirmed cases are going down [118].

this kind of *visualization misinformation*, which occurs when charts send misleading messages to the audience.¹ Some are unintentional errors due to negligence, but others are purposefully designed to misguide the viewers. As consumers of information, the public needs to be able to distinguish between accurate and inaccurate representations in order to navigate around these erroneous and potentially misleading visualizations.

To combat visualization misinformation and assess the effectiveness of potential interventions, we must be able to measure people’s ability to critically interpret visualizations. However, existing visualization literacy assessments [26, 106] mostly focus on the ability to read and extract information from *correctly-constructed* visualizations, which can only provide limited help in measuring the *critical thinking* skills necessary to iden-

¹Many such examples have been discussed at the VisLies meetups held in conjunction with the *IEEE VIS* conference [158, 159, 160].

tify and reason about visualization misinformation. Recently, Camba et al. pointed out that *deception* identification is an important aspect of visualization literacy that must be explicitly taught [33], which we consider a form of visualization misinformation.² However, they considered only one type of deception (inappropriate y-axis range) [33]. There remains no systematic way of measuring the ability to identify visualization misinformation as a component of visualization literacy.

We contribute (1) a design space of chart types and potential *misleaders* (decisions made in the construction of visualizations that can lead to conclusions not supported by the data) garnered from the literature and (2) CALVI, a test systematically developed using this design space to measure the critical thinking aspect of visualization literacy: *the ability to read, interpret, and reason about erroneous or potentially misleading visualizations*. We use Item Response Theory (IRT) and qualitative analysis on the results of a test tryout experiment on 497 participants to identify how different questions separate participants by this ability. We also revise the item bank using a wrong-due-to-misleader score and content validity index (CVI). Based on our analysis, we finalize a reliable ($\omega_t = 0.81$) question bank of 45 questions to measure critical thinking in visualization literacy. We provide recommendations on how to tailor tests based on researchers' specific needs; for example, how to construct shorter tests that are similarly reliable. Our work is a necessary step towards a line of future research that aims to design effective interventions to improve people's visualization literacy and battle visualization misinformation.

²In our view, deception implies intent; some visualizations are misinformative by accident—thus our use of the term *misinformation* instead of *deception*.

3.2 Test Conceptualization and Construction

In the development of our Critical Thinking Assessment for Literacy in Visualizations (CALVI), we extend previous definitions of visualization literacy and define **the critical thinking aspect of visualization literacy** to be *the ability to read, interpret, and reason about erroneous or potentially misleading visualizations*.

As visualization misinformation may be encountered in many contexts such as examples shown in Figure 3.1, we aim to design CALVI for the general public. For example, educators or employers may wish to assess and improve the critical thinking skills in the context of visualization interpretation of their students or employees; researchers may want to assess the effectiveness of interventions against visualization misinformation. For ease of access and data collection, we decided to develop an internet-based test and administer it through a survey management software. We chose Qualtrics because of its comprehensive functionalities and customization flexibility.

Subsequently in the test construction phase, we created a design space to systematically build our item bank, then we conducted a preliminary study to evaluate and iterate on the items.

3.2.1 Test Construction: CALVI Format

We chose to create our items in a selected-response format, which includes multiple-choice and true-or-false questions. This format provides a convenient and efficient way to aggregate results and conduct quantitative item analysis and is easy to use for large-scale testing [81]. Henceforth, we will use *items* to refer to selected-response questions in our test that are associated with visualizations. We conducted an initial pilot study to

qualitatively try out our test with members in our lab and observed that 30 items is a reasonable set for people to complete in one sitting. Thus, we limited the test to contain 30 items with an estimated completion time of 30 minutes.

In the real world, people likely encounter misleading visualizations mixed in with correctly construed visualizations, making it harder to identify and reason about the potentially misleading ones. To simulate this reality, CALVI contains two categories of items: (1) *trick* items using misleading and erroneous visualizations and (2) *normal* items using well-formed visualizations inspired by VLAT. Each participant sees 15 items from each category. We use only the *trick* items to assess their critical thinking ability in visualization interpretation. To be able to analyze a large bank of items, the 15 *trick* items each participant sees are randomly drawn from our item bank, which we describe below. The 15 *normal* items are fixed.

3.2.2 Test Construction: Visualization Content Design Space

To systematically generate the items in the bank, we constructed a design space that consisted of combinations of possible ways a chart can become misleading (we refer to these as *misleaders*) and *chart types*. Below, we describe the process of distilling the set of misleaders and chart types used to construct the design space (also shown in Figure 3.2).

Misleaders

To compile an initial set of ways a visualization can mislead, we drew from two main prior works: categorizations from McNutt et al. and Lo et al. [121, 113]. We reviewed each category from McNutt et al., extracted relevant categories based on main criteria such as **visually detectable** (we cannot test for misleaders that people cannot detect visually) and **not cognitive biases** (cognitive biases from readers do not fit in the definition

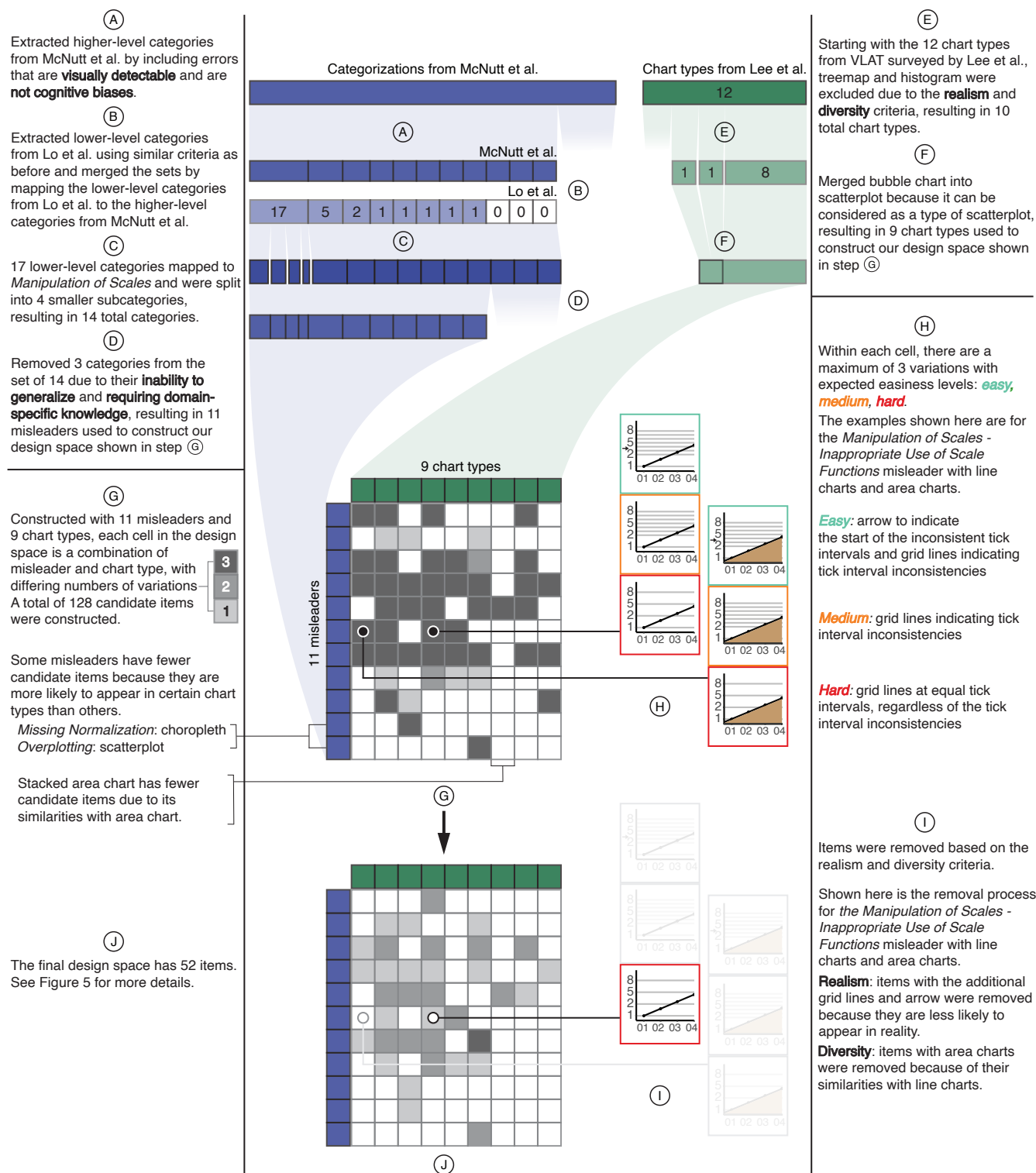


Figure 3.2: Process of distilling the set of misleaders and chart types used to create the design space. While preserving a similar distribution of items in the design space, we removed items from the set of 128 candidates using the realism and diversity criteria, leaving 52 items in the bank.

of misleader as they are not part of the visualization construction process³), and further categorized them into higher-level or lower-level categories ((A) in Figure 3.2). The same process is repeated with categorizations from Lo et al.⁴ Then, we merged the two sets by mapping the lower-level categories from Lo et al. to the higher-level categories from McNutt et al. ((B) in Figure 3.2). During the merge, we mapped 17 lower-level categories to the *Manipulation of Scales* higher-level category, making it the largest category. We then split it into four subcategories: *Inappropriate Order*, *Inappropriate Scale Range*, *Inappropriate Use of Scale Functions*, and *Unconventional Scale Directions*, resulting in 14 misleaders ((C) in Figure 3.2).

In order to systemically apply the misleaders to different chart types, the misleaders have to be generalizable across chart types. Thus, misleaders with an **inability to generalize** cannot be applied to a variety of chart types to populate the design space. Additionally, the items need to be self-contained, so they **cannot require domain-specific knowledge**. Three high-level categories from McNutt et al. [121] were removed from the set of 14 based on these two criteria ((D) in Figure 3.2). Namely, *within-the-bar-bias* and *misunderstanding area as quantity* categories were removed because of their inability to generalize to chart types without bars and without area encodings, respectively. *Assumptions of causality* category was removed due to it requiring domain-specific knowledge: e.g., to decide whether or not a correlation in a visual representation reflects a causal relationship requires knowing the causal structure of the domain, and is not a property of the visualization itself. Thus, the result is a set of 11 misleaders, whose descriptions are shown in Table 3.1.

³However, cognitive biases may be exacerbated by a misleader: for example, *Missing Normalization* may induce a *base rate bias*, so we include the former and exclude the latter.

⁴Details from the derivation process can be found in supplemental materials: <https://osf.io/pv67z>.

Table 3.1: The descriptions of the 11 misleaders in our design space. The presence of these misleaders can result in confusion and inaccurate conclusions of the data.

Misleader		Description
Cherry Picking		Selecting only a subset of data to display, which can be misleading if one is asked to infer something about the whole set of data. [160, 121, 113]
Concealed	Uncertainty	Not displaying uncertainty in visualizations may misrepresent the certainty in the underlying data. In the case of prediction making, this can misguide the viewers to falsely overconfident conclusions. [160, 121]
Inappropriate Aggregation		Aggregating data in an improper way that leads to inaccurate conclusions. [121, 159, 113]
Manipulation of Scales -		
Inappropriate Order	Order	The axis labels or legends appear to be in a random order due to manipulation of data ordering. [113]
Manipulation of Scales -		
Inappropriate Scale Range		Manipulating the range of the scales of axes or legends, such as stretching or truncating the axes or insufficient binning for color scales. [113, 121, 49, 50, 159, 158, 128]
Manipulation of Scales -		
Inappropriate Use of Scale Functions		Applying arbitrary non-linear functions to scales. [121]
Manipulation of Scales -		
Unconventional Scale Directions		The direction of scales of axes or legends is created against convention such as inverting axes or scales. [121, 128]
Misleading Annotations	Annotations	Annotations that contradict or make it harder to read the visualization. [98]
Missing Data		A visual representation implies data exist but the data is actually missing. [121, 96]
Missing Normalization	Normalization	Displaying unnormalized data in absolute quantity when normalized data in relative quantity is of interest. [121, 159, 51, 113]
Overplotting		Displaying too many things on a plot can obscure parts of the data. [121, 113]

Chart Types

We started with the 12 chart types from VLAT surveyed by Lee et al. [106] ((E) in Figure 3.2). Because we want the combinations of chart types and misleaders to create misinformative visualizations we might expect people to see in the wild, we remove chart types that are less likely to appear in reality (i.e., **realism** criteria). Hence, treemap was removed: it was ranked in the bottom half in the list of chart types in data visualization authoring tools and news outlets by Lee et al. [106]. In addition, to run an experiment in which each item is seen by a reasonable number of people, we need an item bank that is not too large. Yet we still want it to be diverse (i.e., **diversity** criteria). Thus, we removed histogram due to its similarities with bar chart after applying the misleaders. Additionally, we merged bubble chart into scatterplot as it is essentially a type of scatterplot with an additional dimension ((F) in Figure 3.2). The result is a set of 9 chart types.

Design Space Structure

To construct the skeleton of the design space, we generated a matrix with misleaders as rows and chart types as columns ((G) in Figure 3.2).⁵ This matrix helped us explore how misleaders can be applied across chart types to systematically construct misleading visualizations, and we refer a cell in this matrix as an *item type*.

Designing Visualizations

Next, we filled out the design space by applying misleaders to different chart types. We want to note one important consideration during the design process: visualizations alone are not necessarily misleading. To be misled means that the conclusion drawn from the visualization deviates from the correct conclusion, requiring a correct answer or conclusion to exist in the first place. Thus, a visualization can only be misleading when there is

⁵The initial design space can be found in supplemental materials: <https://osf.io/pv67z>.

a specific question or task that viewers need to answer or perform using the visualization. We acknowledge that certain types of visualization are implicitly associated with specific tasks because they are designed in such a way to highlight certain aspect of information from data, so those visualizations can be misleading even without explicitly asking viewers to perform a task based on them. Therefore, a misleading visualization cannot be divorced from its visualization task, and this is crucial to keep in mind while designing the visualizations in the design space (and writing the question text afterwards).

While filling out the design space, we also designed visual modifications that we believed would make some items easier or harder than others, and we systematically applied the alterations across chart types (Ⓜ in Figure 3.2). This is only our attempt to have items with varying levels of easiness, a property of items ultimately measured by IRT (see Section 3.4). We show another example alteration in Figure 3.3, which stemmed from applying the misleader *Manipulation of Scales - Inappropriate Scale Range* on bar charts. The base version of this is simply truncating the y-axis, and the alteration of adding an axis break should make it easier to identify the truncated axis. This specific item type (i.e., bar chart with *Manipulation of Scales - Inappropriate Scale Range*) has two variations. With the variations in each item type, a total of 128 candidate items were generated after applying misleaders across chart types, as shown in Ⓜ in Figure 3.2.

Within the set of 128 candidate items, there were redundancies due to generating up to three variations within each item type. Thus, to arrive at a diverse bank of items, we reviewed and eliminated misleading visualizations in the design space by following the same two criteria as before: **realism** and **diversity** (Ⓛ in Figure 3.2). By the realism criterion, if an item type is not likely to appear in a real world setting, then it is eliminated. For instance, inconsistent grid lines and an arrow next to them were less likely to appear

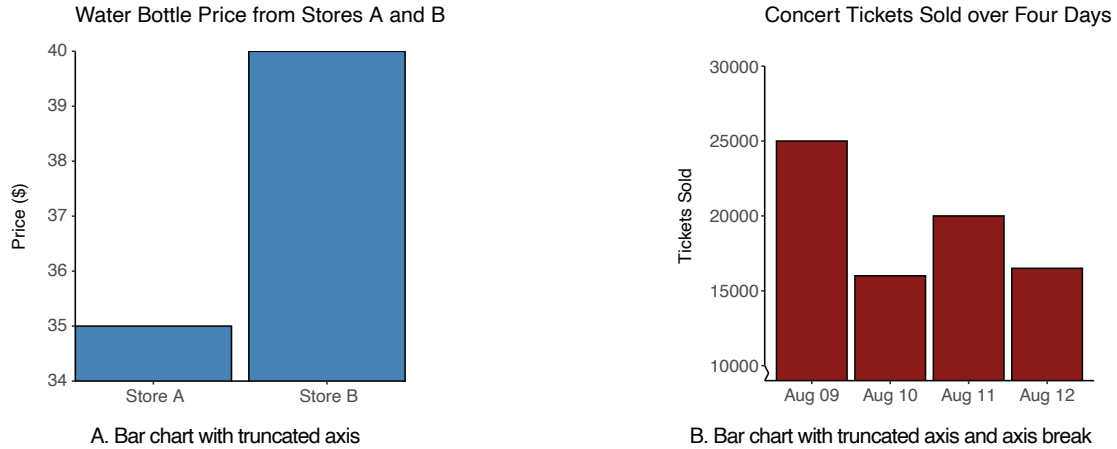


Figure 3.3: Two variations of the same item type: bar chart with misleader *Manipulation of Scales - Inappropriate Scale Range*.

in reality for the *Manipulation of Scales - Inappropriate Use of Scale Functions* misleader, so such items were removed (① in Figure 3.2). Per the diversity criterion, we removed item types that are redundant. One such example was *Manipulation of Scales - Unconventional Scale Directions* for stacked bar chart: including this item type would not add variety as it is essentially the same outcome design as a regular bar chart, so we eliminated it. After removing item types based on the realism and diversity criteria, we are left with a total of 35 item types that stemmed from 11 misleaders and 9 chart types. Along with the variations within each item type, the resulting bank contained 52 items associated with erroneous and misleading visualizations (② in Figure 3.2). An overview of our final design space is in Figure 3.4.

3.2.3 Test Construction: Writing *Trick* Items

Again, it is important to note that whether a visualization is misleading depends on the underlying visualization task one is asked to perform. Take the example of a line chart with an inverted y-axis: this visualization can be very misleading if one were asked to

	Area Chart	Bar Chart	Choropleth Map	Line Chart	Pie Chart	Scatterplot	Stacked Area Chart	Stacked Bar Chart	100% Stacked Bar Chart
									
Cherry Picking				2					
Concealed Uncertainty		1	1			1			
Inappropriate Aggregation	1	2		2		2		2	
[MS] Inappropriate Order	1	1	1	1		1			1
[MS] Inappropriate Scale Range		2	2	2			2	1	
[MS] Inappropriate Use of Scale Functions		1		1	2				
[MS] Unconventional Scale Directions	1	2	2	2		3			
Misleading Annotations		1		2	1	1			
Missing Data			1						
Missing Normalization			1						
Overplotting						2			

Figure 3.4: The design space with 11 misleaders and 9 chart types. The numbers in the cells indicate how many variations are associated with the combination of misleader and chart type. Four misleaders directly relate to the manipulation of scales (MS).

identify the trend of the line and did not notice the inverted axis. However, if the viewer was asked to retrieve the y value of the line at a specific x value, then it is not misleading because they would simply identify the point of interest on the x -axis and look up its corresponding y value. When we wrote the question text for each visualization in the design space, we ensured that the visualization task associated with each question is a relevant task for the visualization to be misleading. For example, *Concealed Uncertainty* is most salient in prediction-making, so all items in this category ask test takers to *make predictions*; to test whether people can detect *Inappropriate Aggregation*, the items must ask them to *aggregate values* from the visualization, such as finding the average. There are also misleaders that have multiple relevant tasks: in *Manipulation of Scales - Unconventional Scale Directions*, it is appropriate to ask people to *find correlations/trends* from a line chart with an inverted y -axis or to *make comparisons* of two regions in a choropleth map where the color scale is inverted. The items in our bank involve a total of six tasks: *retrieve value*, *find extremum*, *find correlations/trends*, *make comparisons*, *make predictions*, and *aggregate values*, all of which are visualization tasks rooted in literature [7, 27, 154].⁶ The first four tasks were taken from VLAT [106], and the latter two were added to support the *Concealed Uncertainty* and *Inappropriate Aggregation* misleaders. For further discussion on the relationship between tasks and misleaders, see Section 3.8.5.

To further increase the diversity of the items, we created a total of 52 different contexts (background stories), covering topics from the prevalence of a plant species to the market share of cell phone brands. Representative item examples from each of the 11 misleaders in our design space are shown in Figure 3.5. The contexts were intentionally made fictional in order to limit participants' use of prior knowledge. Thus, every item is self-contained and participants should be able to select the best answer solely based

⁶The specific task(s) of each item in the bank can be found in supplemental materials: <https://osf.io/pv67z>.

on what is given in the question text, answer choices, and the associated visualization. To isolate the effect of the specific visualization misleader of each item, we designed the choices for the items to be such that there is at least one correct (best) answer and at least one wrong but seemingly correct answer due to the misleader (i.e., wrong-due-to-misleader). Any other incorrect answers are more obviously wrong and are unrelated to the misleader (i.e., wrong-but-unrelated-to-misleader). The wrong-due-to-misleader answer(s) are needed to measure susceptibility to the misleader (see Figure 3.5).

For items with visualizations that do not present enough information for the test taker to choose a reasonable answer, such as the items in the *Inappropriate Aggregation* misleader category, we considered two ways of phrasing the correct answer: “Cannot be inferred” and “Inadequate information”. After trying both in pilots, we decided to keep it consistent throughout the test and used “Cannot be inferred / inadequate information” for these answers. Additionally, we added this option to items where the correct answer is not “Cannot be inferred / inadequate information” to ensure that this option is not the correct answer every time it appears in an item, so that the answer choices did not hint at the correct answer.

One difficulty while writing the items was to not make the question text and answer choices too long. The concern with long question text is that reading and parsing it might interfere with the goal of measuring the ability to identify the misleader (i.e., answering incorrectly due to errors in reading the question text or answer choices instead of not noticing the misleader). An example technique we used to shorten the texts is to label the points of interest on a chart to avoid using long descriptions to pinpoint those points.

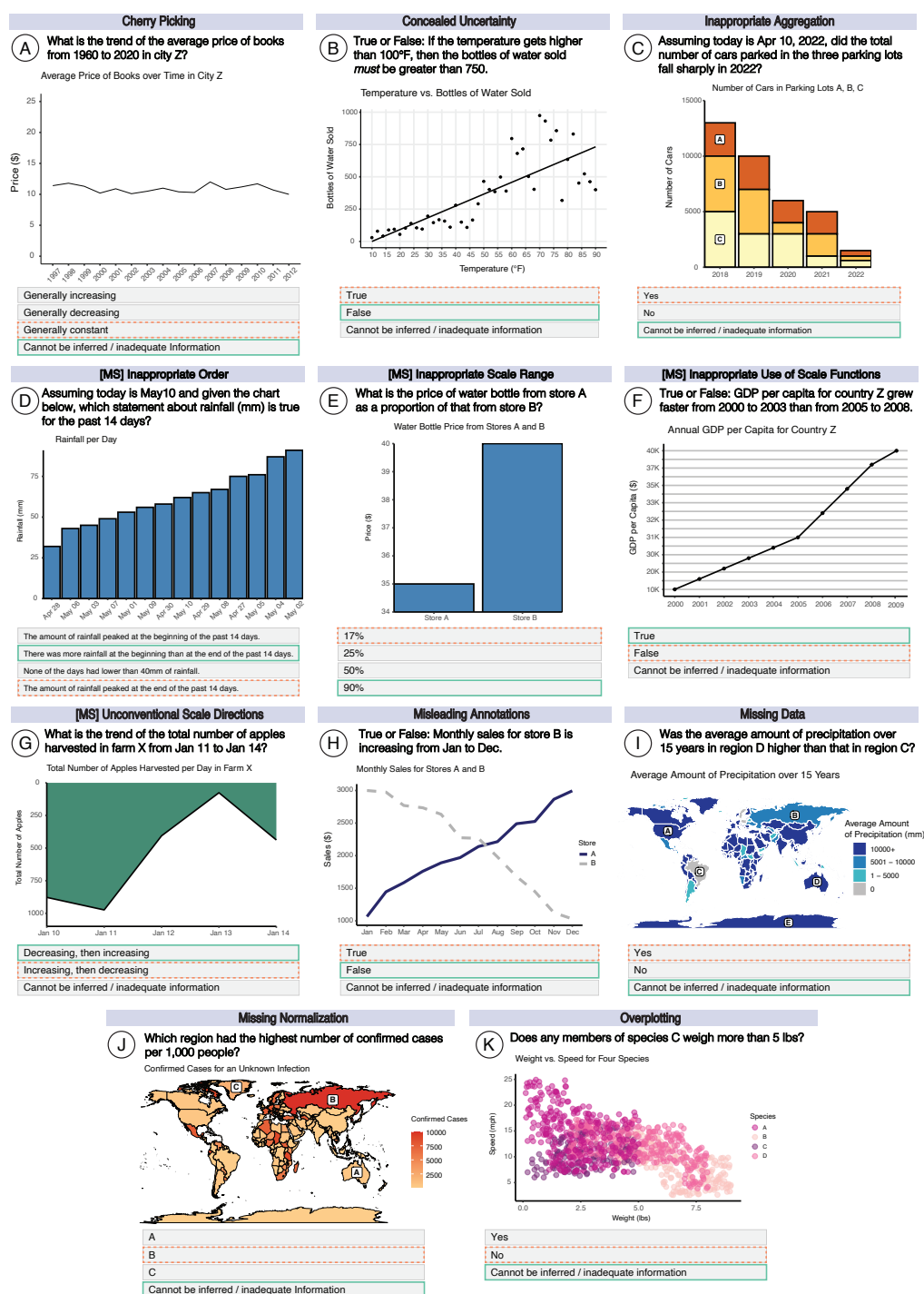


Figure 3.5: Representative items of the 11 misleaders from the bank. Correct answers are indicated with solid outlines, and wrong-due-to-misleader answers are indicated with dashed outlines. Arriving at the correct answers generally requires the test taker to reason about and reflect on the construction process after recognizing the presence of those misleaders, whereas arriving at the wrong-due-to-misleader answers suggest that the test taker did not recognize those misleaders at all or were not reflecting on the construction process to eliminate the effects of those misleaders.

3.2.4 Test Construction: Designing *Normal* items

In order to construct a diverse set of 15 *normal* items on well-formed visualizations that cover all chart types, we ensured that there is at least one item from each of the 9 chart types and created additional *normal* items for the chart types that appear more frequently in the item bank. As a result, 6 chart types have 2 items and 3 chart types have 1 item. We also kept the visualization tasks of the *normal* items consistent with those of the *trick* items, which mainly include comparing values and identifying trends. Most of the 15 *normal* items are essentially versions of *trick* items but with the misleaders removed. We did so to ensure the similarity and consistency between *trick* and *normal* items, so that participants would not be able to distinguish them just by looking at the style of the items. For the same reason, when writing the question text and answer choices for the *normal* items, we followed the same general principles as writing the *trick* items, including keeping the text concise and adding the option of “Cannot be inferred / inadequate information” to items where it is not the correct answer.

3.2.5 Test Construction: Preliminary Study

The goal of the preliminary study is twofold: (1) to qualitatively identify sources of ambiguity and misunderstanding in the question text and the visualizations in the test and (2) use the preliminary data to help determine the sample size needed for the test try-out phase. Therefore, in addition to asking each participant to answer 30 items, we also asked them to complete a set of open-ended questions related to the set of items that were randomly assigned to them. The open-ended questions asked participants to explain why they selected their answers for the items that they answered incorrectly (participants were oblivious to this logic). Thus, along with 3 attention check questions, each

participant received 33 selected-response items and a subset of open-ended questions depending on their responses in the selected-response section. There was no time limit.

Participants

We recruited 30 participants⁷ from Prolific for the preliminary study, whose average approval rate is 99.57%. All of the participants are located in the U.S. and speak fluent English. In addition, participants whose ages do not fall between 18 and 65 or who do not have normal or corrected-to-normal vision are excluded from the study. We collected a balanced sample of 15 males and 15 females with age ranging from 19 to 64.

Procedure

First, we presented participants with a consent form and information page describing the structure of the study.⁸ They were instructed that they are required to select an answer for the current item/question before moving on to the next one, and once they moved on, they could not return to previous ones.

Method

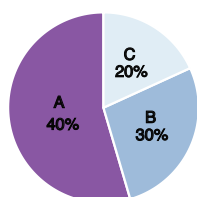
Because one of the goals of the preliminary study was to uncover any confusion in the question text or visualizations, for each selected-response item, we reviewed participants' responses to the corresponding open-ended question to understand their reasoning behind their incorrect choices and whether their decision was due to the intended misleader or problems with our question text or visualization.

⁷A sample size greater than common standards at CHI [32].

⁸The consent form, instructions, test items, and open-ended questions can be found in supplemental materials: <https://osf.io/pv67z>.

Does cell phone brand A have more than half of the total market share?

Market Share of Cell Phone Brands in City Z

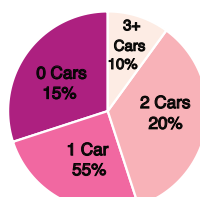


Yes
No
Cannot be inferred / inadequate Information

A

Are there more households with exactly 1 car than households with exactly 2 cars?

Household Car Ownership in City Z

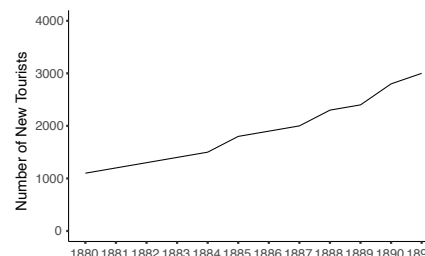


Yes
No
Cannot be inferred / inadequate Information

B

What is the trend of the number of new tourists in town Z from 1880 to 1891?

Decreasing Number of New Tourists in Town Z from 1880 to 1891



Generally increasing
Generally decreasing
Generally constant
Cannot be inferred / inadequate Information

C

Figure 3.6: The set of items the participants may receive with visualizations that contain conflicting information and have a corresponding question in the open-ended section during test tryout (see Supplemental Materials). In A and B, the percentage labels do not match the sizes of the pie slices. In C, the title says decreasing trend while the trend of the line is actually increasing.

Qualitative Results

The text responses revealed a few sources of ambiguity and inconsistency in the design of items. Two such items are on pie chart and *Manipulation of Scales - Inappropriate Use of Scale Functions*; in these items, the sizes of the pie slices are inconsistent with the percentages written on the slices. For both items, we initially designated the option of “Cannot be inferred / inadequate information” as the correct answer, because this inconsistency should suggest that the visualizations do not convey any reliable information. However, some participants expressed that they noticed this conflict between the percentages and the sizes of the pie slices, but they decided to trust one over the other nonetheless. Another item on line chart and *Misleading Annotations* has a similar quality: the title of the chart disagrees with the trend of the line. We were curious about how people understand such visualizations when there is conflicting information present, so we decided to im-

plement an open-ended questions section in the test tryout phase and asked participants who received the corresponding *trick* item(s) in the selected-response section to justify their answers for these three items (shown in Figure 3.6).

Another (slightly ironic) ambiguity came in the interpretation of stacked bar charts and stacked area charts. The key ambiguity for these charts is whether the chart was constructed using a position encoding or length encoding. If a position encoding is used, then the quantity of each segment of the bar (or area) should be the corresponding number on the y-axis. If a length encoding is used, then the quantity of each segment should be the difference between the top of the segment and the bottom of the segment (i.e., the top of the segment immediately below). We had disagreement ourselves during item construction: two different authors had used both types of interpretations in the construction of stacked bar charts and area charts, and many participants were able to understand both interpretations and make correct choices based on them. This reflected the ambiguous nature of stacked charts. To resolve this inconsistency, we unified our interpretation to using length encoding, which is the more common interpretation, and we designed the choices of the items in such a way that the “correct” answer for the position-encoding interpretation do not appear in the choices of most items related to stacked bar charts and area charts. To further reduce the ambiguity, we added visual cues with the goal of making the use of length encoding more obvious, such as add transparency to the stacked area charts as well as grid lines in the background to emphasize the interpretation should be based on length encoding.

For the rest of the items, there were no major ambiguities. We made stylistic modifications to a small subset of visualizations to improve their presentations, such as adding strokes to state and country borders in choropleth maps and making the colors more dis-

tinct from each other for certain visualizations with color encoding. Through reading the open-ended responses, we also noticed that the vast majority of participants who answered the wrong-due-to-misleader answers continued to explain their reasoning without recognizing the misleader.

Test Tryout Sample Size Determination

The data from the 30 participants were used to fit a preliminary model that was then used to simulate 500 participants, which is the sample size that would likely be sufficient for the 2PL IRT model [59]. We fit the simulated 500 participants to our preregistered model (details in Section 3.4.2) and checked that our model converged with a sample size of 500, posterior predictive checks were reasonable, and we could estimate the correlation between *trick* and *normal* items to a resolution of approximately ± 0.1 .⁹ Thus, we chose 500 as our sample size for test tryout.

3.3 Test Tryout

As before, the test in this phase consists of a selected-response section and an open-ended section. The selected-response section is the same as that of the preliminary study, which contains 15 *trick* items randomly sampled from the bank, 15 *normal* items, and 3 attention check questions. As a result of the preliminary experiment (paragraph 3.2.5.0.4), the open-ended section includes questions for three items that contained conflicting information (shown in Figure 3.6) to further understand why some preferred one choice over the other.

⁹The script for this simulation and model check analysis can be found in supplemental materials: <https://osf.io/pv67z>.

Participants

We originally recruited 500 participants from Prolific for test tryout. The same pre-screening in the preliminary study requirements were applied. We filtered out one participant who failed the attention check (i.e., provided incorrect answers for two of the three attention check questions). One participant who did not enter their Prolific ID at the beginning of the study was also removed. To rule out random clickers, we excluded participants who spent less than 5 seconds on more than half of the items; only one participant fit into this exclusion criterion. The result is a total of 497 participants whose data were used in our analysis. The final set of participants consists of 248 males and 249 females, with an average approval rate of 99.62% on Prolific. Their ages range from 18 to 65. Four participants reported that they were color blind. These participants also come from a wide variety of education levels: 97.59% of participants had an education level of high school or above, 15.49% graduated from a technical or community college, 37.83% hold a Bachelor's degree, and 16.50% hold a graduate degree or above.

Procedure

The same procedure in the preliminary study was applied in the test tryout phase. Participants received information about the test, and upon consenting, they were provided with clear instructions on how to proceed.

3.3.1 Descriptive Statistics

We calculated some basic statistics on participant level raw correctness and completion times.

Participant Correctness

Participants' correctness (proportion of correct answers) for the 15 *trick* items ranged from 0 to 0.93 ($M = 0.39$, $SD = 0.16$). Participants' correctness for the 15 *normal* items ranged from 0.27 to 1 ($M = 0.80$, $SD = 0.13$).

Completion Time

We examined the completion time for the selected-response section of the test, which included 15 *trick* items, 15 *normal* items, and 3 attention check questions. The total time in minutes (the complete distribution can be found in supplemental materials¹⁰) ranged from 4.62 to 89.60 ($M = 19.80$, $SD = 10.25$). This suggests that 30 minutes is reasonable to complete the test.

3.4 Item Analysis

3.4.1 2-Parameter IRT Model

We selected the 2PL IRT model, which characterizes each item in two dimensions: item easiness and item discrimination. Item easiness can be interpreted as the easiness (or negation of difficulty) of correctly answering the item. Item discrimination can be understood as an item's ability of differentiating participants of different levels of ability. The model is described by Equation 3.1: the left-hand side is the probability of a participant with ability θ answering item i correctly given item easiness b_i and item discrimination a_i ; the right-hand side is the formula of how to compute this probability given the values of θ , a_i , and b_i . Equation 3.1 is often referred to as the item response function, and the curve of this function is called the item characteristic curve (ICC).

¹⁰<https://osf.io/pv67z>

$$p_i(\theta) = \frac{1}{1 + \exp(-a_i(\theta + b_i))} \quad (3.1)$$

Item Easiness

To understand item easiness, consider the case where $\theta = -b_i$. Then by Equation 3.1, the probability of a person with ability $\theta = -b_i$ answering item i correctly is $\frac{1}{2}$. For example, the dot on the ICC curve of item A in Figure 3.7 has $\theta = -b_A$ and probability of correctness of 0.50. In other words, if a person's ability is the same as an item's difficulty, which is the negative of its easiness, then that person has a 50% chance of answering that item correctly. Therefore, the higher the person's ability θ is above the item difficulty $-b_i$, the more likely they will answer the item correctly.

Item Discrimination

To gain intuition for the discrimination parameter, observe that the maximum slope of Equation 3.1 is $\frac{a_i}{4}$ at $\theta = -b_i$. For instance, Figure 3.7 shows the slope of the tangent line at the point on the ICC curve of item B where $\theta = -b_B$, which is $\frac{a_B}{4}$. As a result, a_i is associated with the rate at which the probability of answering the item correctly changes with ability values, hence interpreted as item discrimination. If an item has high discrimination, then people with higher ability should have a much greater chance of answering the item correctly than people with lower ability.

3.4.2 Bayesian IRT

We estimate the parameters of our IRT model using Bayesian modeling. In Bayesian IRT, we are interested in obtaining the posterior distribution of the parameters given the observations. This contrasts traditional IRT where methods such as maximum likelihood

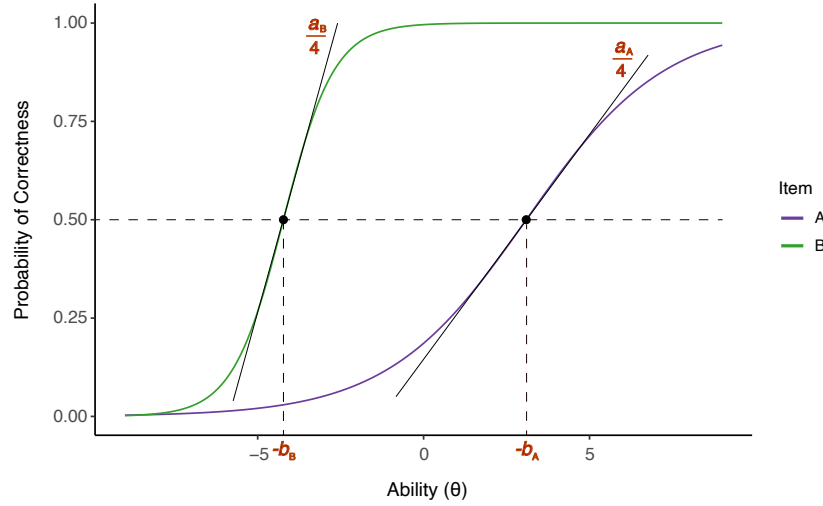


Figure 3.7: ICC curves for two example items, A and B. The θ values corresponding to the two points on the curves are the difficulty values (negation of easiness) for those two items, and the slopes of the curves at those points are proportional to the discrimination values of the items. In this example, item A is more difficult than item B, while item B is more discriminating than item A.

estimation are used to find the best point estimates of the parameters. We chose to use Bayesian IRT because it has better modeling flexibility and provides more informative results [29].

After we obtain the posterior distributions of the ability, item easiness, and item discrimination parameters, we take the medians of these distributions as our point estimates.

For our analysis, we used the R package **brms**, which provides flexible ways to conduct Bayesian IRT modeling [29], which was preregistered on OSF (see <https://osf.io/pv67z/>).

3.4.2.1 Selection of Prior Distributions

In addition to getting posterior distributions rather than point estimates as output, Bayesian IRT differs from traditional IRT in that prior distributions have to be specified in the Bayesian model. To ensure identifiability, we set the priors of the standard deviations of the slopes of the *trick* and *normal* items to be constant 1 [29]. We selected the LKJ distribution with parameter $\eta = 2$ as the prior for the correlation matrix, and used $\mathcal{N}(0, 1)$ as the priors for the means and the standard deviations of ability and easiness, and $\text{lognormal}(0, 1)$ for discrimination. This suggests that these means are between -2 and 2 in logit space with high probability, and when discrimination is 1 (the median of its prior), the average item correctness would be approximately between 11.92% and 88.08%, which is a reasonable range to expect for average item correctness.

We also conducted sensitivity analysis of whether setting the prior of one of the standard deviations of the slopes of the *trick* and *normal* items to be constant 1 and the other to be $\mathcal{N}(0, 1)$ would have a significant effect on the results. Our analysis showed that the difference between models with these different prior distributions are negligible, so we decided to use our original prior selection.

Before the test tryout experiment, a dataset of 500 participants was simulated using data from the preliminary study to test our model specifications (explained in paragraph 3.2.5.0.5). We ran the Bayesian IRT model with the selected prior distributions and achieved stable results.

We ran our final model with 4 chains, each with 20,000 iterations. We discarded 10,000 warmup iterations per chain and thinned the final sample by 5, yielding 8,000 total post-warmup draws. The minimum bulk effective sample size is 3,067 and the minimum

tail effective sample size is 5,494, and all $\hat{R}$ values are approximately 1.

3.4.3 Results

Figure 3.9 contains the analysis results of each item, including the item easiness and discrimination estimates, correctness, $rate_{wm}$ (see Section 3.6.1), $chance_{wm}$ (see Section 3.6.1), and the content validity index (CVI) (see Section 3.6.2).

Item Easiness and Discrimination

The median item easiness parameter estimates for all *trick* items range from -5.32 to 4.12, with an average of -1.12. The median item discrimination parameter estimates for all *trick* items range from 0.45 to 1.26, with an average of 0.76. Figure 3.9 shows the coefficient plots displaying the median easiness and discrimination of each item with 95% and 66% credible intervals (CI).

In Figure 3.8, we show the average item easiness and discrimination estimates for each misleader. *Overplotting*, *Missing Data*, and *Missing Normalization* are the most dif-

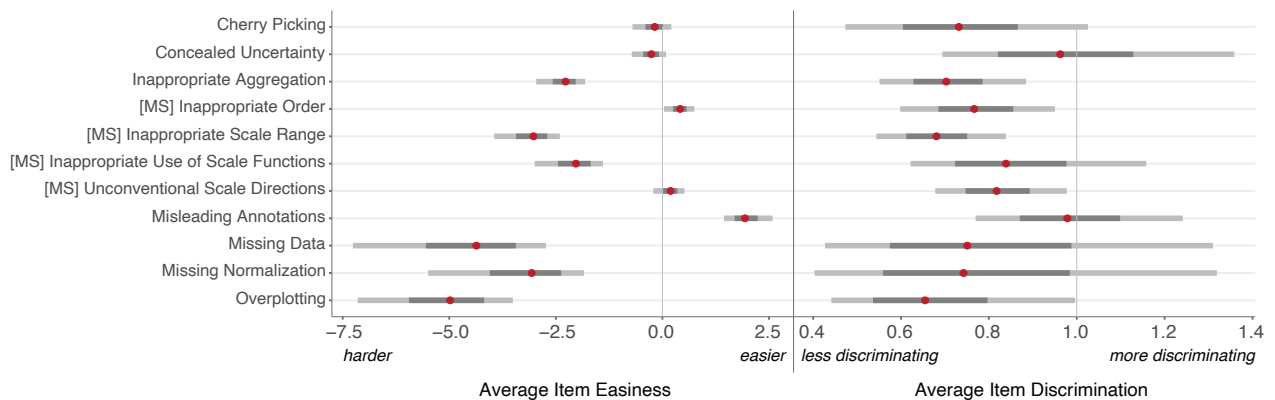


Figure 3.8: Average item easiness and discrimination estimates for each misleader. The dots represent the median average easiness/discrimination estimates, the lighter bar represents the 95% CI, and the darker bar represents the 66% CI.

difficult misleaders, while *Misleading Annotations* and *Manipulation of Scales - Inappropriate Order* are the easiest. The most discriminating misleaders include *Misleading Annotations*, *Concealed Uncertainty*, *Manipulation of Scales - Inappropriate Use of Scale Function*, while *Overplotting* and *Manipulation of Scales - Inappropriate Scale Range* have relatively weak discriminating power.

Performance on trick vs. normal items

In addition to analyzing the items, we also investigated the relationship between the performance of participants on *trick* items and *normal* items. Their performance on the *normal* items can be considered as a rough proxy of their basic visualization interpretation ability. The correlation coefficient is 0.63 with 95% CI: [0.48, 0.76], showing a moderately strong correlation. However, these abilities are still qualitatively different, as participant correctness for *normal* items is generally much higher than that for *trick* items.

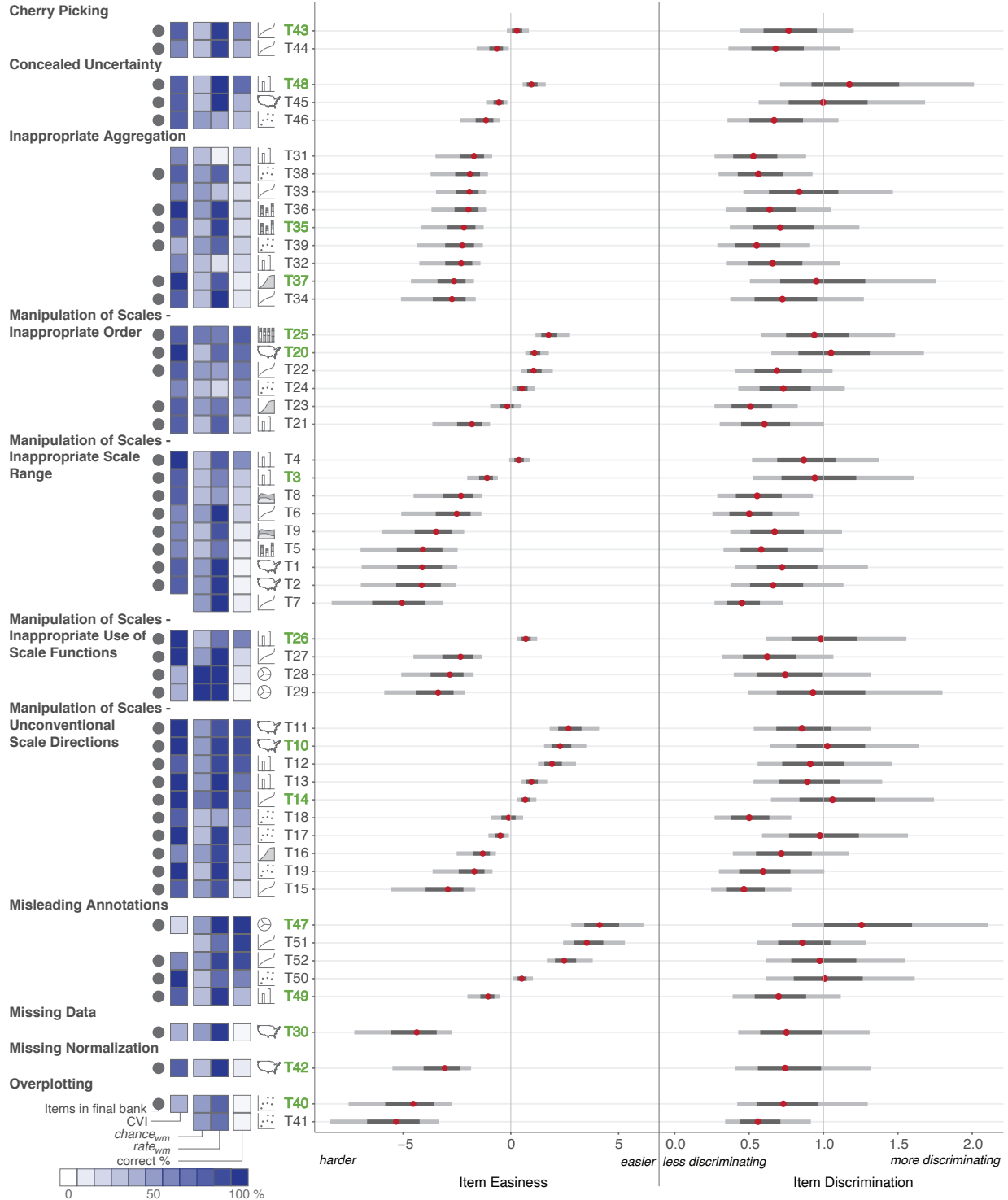


Figure 3.9: Item easiness and discrimination for all of the items in our item bank during test tryout. For each item, the dot represents the median estimate of its easiness/discrimination parameter, the lighter bar represents the 95% CI, and the darker bar represents the 66% CI. Additionally, shown on the left of the item IDs are chart type, correctness, $rate_{wm}$ (see Section 3.6.1), $chance_{wm}$ (see Section 3.6.1), CVI (see Section 3.6.2), and leftmost dots indicate the items that are in the finalized bank. The items in the selected set of 15 are indicated with colored item IDs (see Section 3.8.1).

3.5 Test Revision

To revise our test, we use results from IRT analysis and qualitative analysis of the open-ended questions from test tryout. We removed 2 items based on results from Section 3.4.3 and 1 item based on the responses to the open-ended questions. Below, we explain our reasons for these revisions.

3.5.1 Revision from IRT Analysis

Results from Section 3.4.3 were used to revise the items in the item bank. The revision process is necessarily holistic with many factors to balance [44], such as considering item discrimination, item easiness, and preserving the diversity of the item bank. The two most difficult items have very low discriminating power, namely items T7 and T41: T7 is the second most difficult and the least discriminating item, and T41 is the most difficult item with only 8 items that are less discriminating than it. In addition, both of these items have variations in the design space matrix (T6 for T7 and T40 for T41) that are easier and more discriminating than themselves, so removing them does not reduce the diversity of the item bank.

3.5.2 Revision from Qualitative Analysis of Open-ended Questions

As described in Section 3.2.5, we included open-ended questions for items T28 and T29 (pie charts where numbers and pie slices mismatch) to investigate whether “Cannot be inferred / inadequate information” is the only reasonable correct answer by examining the justifications of participants who did not select that option as their answer.

For both items, most participants answered according to the percentage labels on

the charts. For item T28 shown in Figure 3.6.A, 124 participants were asked “Does cell phone brand A have more than half of the total market share?”, and

- 86 answered “No”, indicating they answered based on the percentage labels rather than the sizes of the pie slices, and their reasoning aligns with this;
- 22 answered “Yes”, showing that they based their answer on the size of the pie slices;
- 16 answered “Cannot be inferred / inadequate information”.

For item T29 shown in Figure 3.6.B, 146 participants were asked “Are there more households with exactly 1 car than households with exactly 2 cars?”, and

- 131 answered “Yes”, indicating they answered based the percentage labels rather than the sizes of the pie slices;
- 7 answered “No”, showing that they relied more on the pie slice sizes when attempting this item;
- 8 answered “Cannot be inferred / inadequate information”.

We found that only a small fraction of such participants acknowledged the conflict between the percentage labels and pie slice sizes in their responses, and out of these people, no one provided a clear and convincing argument for why an alternative choice should be correct. This suggests that there is insufficient evidence that percentage label-based or size of pie slices-based answers should be correct. Therefore, the best answers are still “Cannot be inferred / inadequate information” for both items. This also suggests that the misleader *Manipulation of Scales - Inappropriate Use of Scale Functions* can be hard to detect and reason about. We decided to keep these two items in the bank because they

have good item discrimination and reasonable easiness.

The third item we included an open-ended question for was a line chart with a contradicting title, as shown in Figure 3.6.C (T51). Out of 141 participants who were asked “What is the trend of the number of new tourists in Town Z from 1880 to 1891?”,

- 134 answered “Generally increasing”, indicating that they based their answer on the visual representation of the line rather than the title;
- 5 answered “Generally decreasing”, showing that they answered according to the title of the chart;
- 2 answered “Cannot be inferred / inadequate information”.

We observed that the majority of participants answered this item based on the visual representation rather than the title. Upon examining the justifications of participants who chose “Generally increasing”, we found that most of them did not even notice that the title contained conflicting information at first. Since the aim of this item is to study how people make decisions when reading *misleading annotations* in visualizations and the majority of participants did not even notice the annotation, this suggests that T51 was not measuring susceptibility to misleadingness. Thus, we decided to remove it from the bank.

Interestingly, a previous study that directly investigated trust and recall when titles contradict the visual representation, found that people tend to recall the visualization title more [98]. Perhaps the salience of titles differs when people are asked to recall information from a visualization versus making a decision on the spot (i.e., choose a correct answer).

3.6 Evidence of Validity

Validity of a test is essentially making sure that there is a causal relationship between the ability being measured and people’s performance on the test [25], meaning that differences in ability (i.e., the ability to read, interpret, and reason about erroneous or potentially misleading visualizations) should lead to differences in measurement outcomes (i.e., correctness). We assess the validity of the remaining 49 items in CALVI and remove some items with low validity using two criteria: (1) a wrong-due-to-misleader score (RR_{wm}) and (2) the content validity index (CVI) [131] for each item.

3.6.1 Wrong-due-to-misleader Score

We constructed (Section 3.2.2) and wrote the items (Section 3.2.3) to have three types of answer choices: correct answers, wrong-but-unrelated-to-misleader answers, and wrong-due-to-misleader answers. The wrong-due-to-misleader answers were specifically designed for people who did not recognize or correctly reason about the associated misleader (see Figure 3.5). Thus, if the majority of people who get an item incorrect are choosing a wrong-due-to-misleader answer rather than a wrong-but-unrelated-to-misleader answer, this is evidence that people have been misled (i.e., that the item measures susceptibility to the misleader). We see evidence of this in responses from the qualitative preliminary study (Section 3.2.5): in explaining their reasoning, participants who chose the wrong-due-to-misleader answers generally continued to justify their answers without recognizing the misleader.

For the test tryout study, we use a wrong-due-to-misleader score, RR_{wm} , as one measure of validity. We define RR_{wm} of each item to be $\frac{rate_{wm}}{chance_{wm}}$, where $rate_{wm}$ is the proportion

of wrong answers that people chose that are wrong-due-to-misleader and $chance_{wm}$ is the probability of choosing the wrong-due-to-misleader answer among all wrong answers if one were to choose randomly. We consider items with a $rate_{wm}$ greater than $chance_{wm}$ to be valid, meaning the RR_{wm} is greater than 1. Items with a RR_{wm} below 1 are candidates for revision.

3.6.2 Content Validity Index

We also evaluated items using CVI. The CVI of an item is the proportion of domain experts who rated the item a 3 (quite relevant) or 4 (highly relevant) on a 4-point relevance scale [131]. We invited five domain experts to rate the 49 items in our bank on a scale of 1 (not relevant) to 4 (highly relevant) [131]. Each domain expert has a doctorate and actively conducts research in Information Visualization; three are in academia, and two are in industry. We calculated the CVI for each item based on the expert ratings, and items below 78% are candidates for revision (i.e., should be re-examined for whether they will be retained in the final bank) [131].

3.6.3 Revision from Wrong-due-to-misleader Score and Content Validity Index

We used both RR_{wm} and CVI to evaluate the items:

- 30 items had a RR_{wm} greater than 1 and CVI greater than 78%, so we retain these in the bank.
- 10 items were deemed candidates for revision by the CVI criterion only; we decided to retain these items because their rates of wrong-due-to-misleader answers ($rate_{wm}$) were all well above $chance_{wm}$ ($RR_{wm} > 1.5$).

- 3 items were deemed candidates for revision by the RR_{wm} criterion only. Four of five experts rated these items 3 (quite relevant) or 4 (highly relevant), and at least half of those four experts rated these 4 (highly relevant). We also re-examined the 3 items in detail and agree with the expert ratings, so we retain them.
- 6 items were deemed candidates for revision by both criteria. This includes T28 and T29, already discussed in Section 3.5.2.

For these two items, all wrong answers are wrong-due-to-misleader answers, so the RR_{wm} is 1 by construction, which sits at the threshold of validity by this measure. However, this does not suggest that these two items are invalid because this is an edge case where $chance_{wm}$ is 1 by construction. Additionally, as described in Section 3.5.2, we observed that the majority of people who chose the wrong-due-to-misleader answers for these two items were indeed misled by the associated misleader in their justifications. Thus, we retain T28 and T29. We removed the remaining 4 items deemed candidates for revision by both criteria.

This leaves 45 items in the final bank (indicated with the leftmost dots in Figure 3.9).

3.7 Finalized Bank and Reliability

3.7.1 Finalized Bank of Items

The finalized bank of items spans a wide range of difficulty and discrimination (as shown in Figure 3.10 (top)), which is appropriate for a general audience for which we initially designed the test. Moreover, test administrators can also conveniently customize their tests based on their target audience by selecting the appropriate items from the item bank

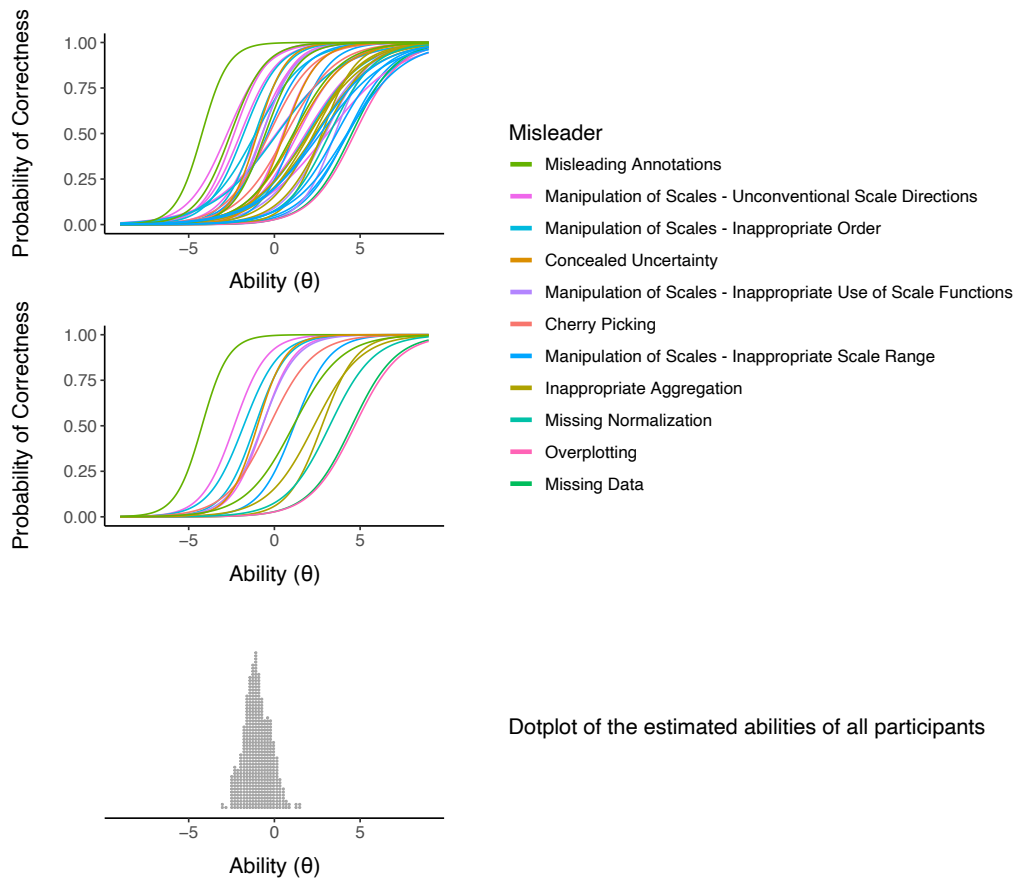


Figure 3.10: ICC curves for all 45 items (top) in the bank show that the bank covers a wide range of abilities. The 15 items (middle) we selected also cover a similarly wide range of abilities, which can be used to test a general audience. The ICC curves are colored by the 11 misleaders. The dotplot (bottom) shows the distribution of the estimated abilities of all participants from test tryout.

(see Section 3.8.1.2).¹¹

As we mentioned in Section 2.4, there is no strict criteria or gold standard for selecting the final set of items. Test administrators should select items most suited to the purpose of the test [44]. It is also worth noting that the test development process we applied in this paper is an iterative process: over time, as the test gets deployed in practice, the item bank may need to be revised to fit its purpose when new data arrives or related

¹¹The item bank can be found in supplemental materials: <https://osf.io/pv67z>.

research emerge in the field [44].

3.7.2 Reliability

Reliability is a fundamental concept in psychometrics, and its basic idea is simple: observed test results are a combination of signal and noise. The higher the proportion that is due to signal rather than noise, the more reliable the test. Psychometricians have developed many measures of reliability, such as ω and Cronbach's coefficient α [54]. However, Dunn et al. pointed out some difficulties with using α and outlined advantages for using ω [62]. Revelle and Condon also showed that ω is a more reliable measure because it does not consistently underestimate reliability like α and captures total variance common to all test items [135]. There are several forms of ω , and ω_t is the total reliable variance of the test, representing the overall reliability of the test. Thus, we used ω_t to measure reliability, and our analysis demonstrated that our finalized item bank with 45 items has a high reliability score ($\omega_t = 0.81$).

3.8 Discussion

3.8.1 Recommendations for Using CALVI

To make our test easily accessible without the need to use an item bank, we provide a set of 15 items that can be used to cover a wide range of abilities. These 15 *trick* items would be combined with the 15 fixed *normal* items to construct a test of 30 items. To select a representative set of 15 *trick* items, we adopt the following method. For each of the 11 misleaders, we select the item with the highest item discrimination. For the remaining 4 items, we prioritize the misleaders that have a large representation in the bank and items with high discrimination at ability levels not covered by the previously-selected 11 items.

The resulting 15 *trick* items cover all misleaders and 8 out of 9 chart types in the design space (colored item IDs in Figure 3.9); the 1 absent chart types is stacked area chart, an acceptable compromise since the similar 100% stacked bar chart, stacked bar chart, and area chart are in our selected set. The ICC curves of the 15 items are shown in Figure 3.10 (middle), and compared to the ICC curves of all 45 items in Figure 3.10 (top), we conclude that this recommended set of items covers a wide range of ability levels and is appropriate for a general audience. This selected set of 15 items also shows high reliability ($\omega_t = 0.82$).

3.8.1.1 Scoring

In order to obtain scores for the test that uses the 15 recommended *trick* items in real time, raw score (i.e., percentage of correctness) can be used as a proxy for the ability of a test taker. The limitation of this approach is that it treats all items equally, but they have varying easiness and discrimination which should be factored into scoring. Although raw score is not a perfect measure, it is highly correlated with ability θ ($r = 0.88$) in our test.¹² Alternatively, test administrators can use IRT to directly obtain ability estimates and use them as scores; the disadvantage of this approach is that it requires technical knowledge of IRT and is computationally expensive.

3.8.1.2 Customizing Future Tests

Although the target test taker population is the general public for our study, it could potentially be interesting to study the performance of participants from a wider variety of populations. Our item bank allows a lot of flexibility in customizing tests. Depending on the goal of test administrators, they should select a subset of the items most suited to their needs. For instance, if one wishes to use the test to filter out low-ability subjects,

¹²A scatterplot of raw score vs. ability and the correlation coefficient can be found in supplemental materials: <https://osf.io/pv67z>.

they should select a set of items that have high discriminating power at the low-ability range. Alternatively, by selecting more difficult items from CALVI's bank, one could design tests that suit more expert audiences.

When customizing, high discrimination items should generally be preferred over low discrimination items. Test administrators can also adjust the length of the test: for example, one can select 5 *trick* items that cover an ability range of interest along with 5 *normal* items to create a shorter and faster test. Reliability of such a shorter test should be assessed prior to use. Future work could experiment with (and validate) these and other shorter formats.

3.8.2 Applications of CALVI in Future Research

CALVI has many interesting applications beyond assessing the ability to reason about misleading visualizations of the general public. In educational settings, CALVI could be used to investigate when students learn to identify visualization misinformation by studying changes in that skill over time (e.g., from high school through college). Similarly, one could use CALVI to investigate whether consumers of different types of news have different critical thinking abilities for visualization interpretation (e.g., are people who read data journalism better at this skill?). Moreover, CALVI can also be used to test the effectiveness of an intervention on visualization misinformation by asking participants to take the test before and after the intervention to observe change in their performance. The item bank offers a resource to construct a pre- and post-test without asking the participants the same items (see Section 3.8.1).

3.8.3 Is Attention All We Need to Detect Visualization Misinformation?

We believe attention is indispensable in identifying visualization misinformation. Some misleaders involve direct manipulations to emphasize or distort certain visual features, such as y -axis inversion that gives an opposite visual impression (see Figure 3.5.G). Thus, for these manipulations, paying attention to the right part of the visualization matters. As with most tests, if one pays more attention to the right part of the question, one is likely to perform better. In the case of CALVI, adding salient features to draw people's attention to the misleading part of the visualization design, such as adding a downward arrow next to the inverted y -axis, seem to make it easier for people to detect and interpret the visualization.

While attention appears to be a key component of the ability to detect visualization misinformation, we want to emphasize that only attention is not enough. For about 20 items in the bank of CALVI (e.g., from *Missing Data*, *Concealed Uncertainty*, *Overplotting*), merely paying attention to the misleading part of the visualization is insufficient — reasoning about them requires thinking critically about the visualization construction process beyond just reading the visual representations. Therefore, the ability to detect and reason about visualization misinformation is not equivalent to the amount of attention one pays in viewing visualizations. We saw evidence in our data that suggests a positive relationship between attention and the ability to read, interpret, and reason about erroneous and potentially misleading visualizations; the connection between the two seem to exist in some real-world examples too: in Figure 3.1.B, if one paid more attention to the x -axis, then it would be easier to identify the inappropriate ordering of the dates. Future work is needed to further investigate the relationship between attention and the ability to detect visualization misinformation.

3.8.4 Context of Use May Affect Interpretation and Misleadingness

In reality, visualizations are rarely consumed alone in media: they are often accompanied by written text or verbal commentaries, which influence viewers' interpretations of the visualizations. Therefore, studying and understanding the interplay between visualization and text is crucial to advance our knowledge of visual reasoning and visualization literacy. In the context of developing a test, an interesting consideration is if we were to ask the question in a different way on the same visualization, would that have a noticeable effect on how people understand and reason about the visualization? Is a particular *misleader* always misleading, or is it only misleading in the face of certain tasks or questions?

3.8.5 Are All Misleaders Actually Misleading?

As we explained in Section 3.2.2, to what extent a visualization can be misleading depends highly on the visualization task. Existing literature lacks clear guidelines connecting misleading visualizations to visualization tasks. Thus, we did so using our best judgment and knowledge of the literature. A taxonomy mapping misleading visualizations to visualization tasks would be a useful direction for future research, and new findings may yield different items in the test.

There may also be disagreement on whether every potential misleader should be considered as a visualization *error*. For instance, some evidence suggests that whether a truncated *y*-axis is “deceptive” or “truthful” is domain-specific and depends on what effect sizes are meaningful in that domain [49]; so some may deem a truncated axis an error while others may not, or may disagree on when it is an error. In our case, if an item was found to be easy, then it may suggest that the associated misleader is not really

a misleader at all. Our design space and item bank offer a starting point for determining what *putative* misleaders are *actual* misleaders: future work could start from the easiest misleaders in our item bank, and conduct studies specifically examining their effect on interpretation across a range of chart types. Such an effort would help systematize the study of visualization misinformation, and could generate more empirically-grounded guidelines for visualization construction.

Chapter 4

Inoculating Against Visualization Misinformation Through Gamification

This work is a working paper and is done in collaboration with Michael S Horn and Matthew Kay.

4.1 Introduction

Visualization misinformation, or misleaders from Chapter 3, refers to any visual representation of data that can send misleading messages, through both graphical and contextual elements within the visualization [136, 75, 72, 109]. Ill-formed visualizations misinform through specific visual manipulations that lead to exaggerations or distortions of data (e.g., inverted y-axis in Figure 4.1.A, y-axis truncation [49]). Well-formed visualizations contain no apparent visual flaws but can still misinform through other contextual elements (e.g., not normalizing data in Figure 4.1.B, annotations [148]). Given the vast and diverse collections of misleading visualizations in the real world (e.g., [121, 113]), viewers must be able to discern problematic data representations to ensure the quality of their information consumption.

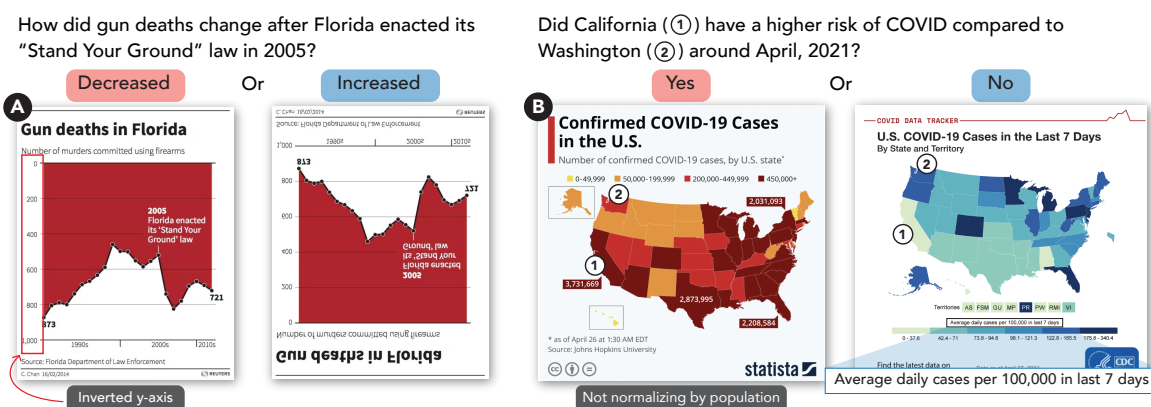


Figure 4.1: Real world examples of visualization misinformation. (A) shows an area chart with an inverted y-axis [102]. (B) shows a map with absolute numbers of COVID-19 case counts (left [52]) when it would be better to normalize by population (right [37]).

One approach to combating visualization misinformation is developing targeted interventions to help viewers critically engage with data visualizations (i.e., improving the critical thinking aspect of visualization literacy). Existing interventions often take the form of external aids (e.g., annotations or visual cues) or education interventions (e.g., perceptual training) [136]. While external aids can offer immediate help in interpreting the visualization at hand, they still occur after exposure to false information (i.e., debunking intervention [149]). The negative effects that are present after exposure can be harder to eliminate [108], suggesting that educational (i.e., prebunking [149]) interventions could be more effective. Although there are visualization literacy interventions that teach visualization skills, a majority of them focus on well-formed visualizations (e.g., [5, 21, 69]), which can only offer limited support in helping viewers critically interpret misleading visualizations. The ones that target misleading features in visualizations largely focus on a limited set (e.g., axis truncation [33, 133, 134, 165]), despite the variety of techniques that can lead to misleading visualizations in the real world [121, 114, 72, 75]. There remains a need to develop educational interventions specifically targeting visualization misinformation.

We address the lack of engaging educational interventions against visualization misinformation through the iterative development of an educational serious game over a series of workshops. This gamified approach was inspired by inoculation theory (i.e., developing resistance to false persuasion through pre-exposure to weakened forms of similar challenges [119]). Prior research also demonstrated the potential of serious games in the textual misinformation landscape (e.g., [141, 140, 78]). Our game development process involved running workshops with our target user group (i.e., high school students) to iteratively refine our game design. During the workshops, we tested out four design variants: individual, competitive, team-based, competitive + team-based designs. We describe each game mechanic and what we learned while observing high school students playing each of our game variants. We outline the challenges we observed and how they informed the design of the next variant(s). Collectively, our playtests offer insights into how we can design a serious game for improving the ability to critically interpret visualizations.

4.2 Related Work

4.2.1 Combating Visualization Misinformation

To combat visualization misinformation, researchers have designed interventions to help viewers detect problematic representations as they come up and enhance viewers' ability to better engage with visualizations. These efforts can be broadly summarized into two categories: retroactive (i.e., debunking [149]) and pre-emptive (i.e., prebunking [149]) interventions against visualization misinformation. Debunking interventions often occur after exposure to the actual scenario(s) of misinformation and take the form of external aids, such as adding textual annotations [114, 49] or visual cues [87, 49] to draw viewers'

attention to a specific portion of the visualization. Others presented alternative forms (e.g., a corrected version) of the visualization for viewers to explore [65]. Prebunking interventions occur before viewers are exposed to the misinformation and often take the form of additional education and training, such as perceptual training with feedback to improve viewers' ability to correctly extract information from visualizations [133]. Others developed lecture-style educational videos [143] and hands-on activity sessions [33] to teach learners how to accurately interpret visualizations and avoid misleading techniques. However, each educational intervention is limited to small sets of misleading factors. Compared to external aids, fewer studies have examined the potential of educational interventions [136]. While debunking interventions (e.g., external aids) offer immediate support, they are generally challenging to generalize as they are often specifically designed for a particular chart context [136]. Prior work has also suggested that prebunking techniques can be more effective than debunking because it is more difficult to eliminate negative effects after exposure [108]. We therefore focus on the development of an educational intervention that can enhance people's ability to engage with visual representations of data and prebunk visualization misinformation.

4.2.2 Improving Visualization Literacy

Visualization literacy and education researchers have developed interventions that enhance people's ability in interpreting visualizations. Many of the existing interventions to improve visualization literacy have targeted younger audiences and focused on simple chart types. For instance, Alper et al. [5] developed C'est La Vis and focused on helping students in elementary school learn pictographs and bar charts. Bishop et al. [21] developed Construct-A-Vis, a tool to help elementary school children to explore and create free-form visualizations. Some have designed for an informal setting, combining

play with learning. Gäbler et al. [69], for example, took a game-based approach and developed Diagram Safari to improve data and visualization literacy in young children. Similarly, Huynh et al. [91], also exploring gamification for learning, developed a narrative role-playing game. To foster creativity, Bae et al. [11] designed a toolkit to improve visualization literacy through physicalizations. Others have investigated and evaluated more general techniques for improving literacy in visualizations, such as online learning through tutorials [100], feedback mechanisms [20], visualization morphing (teaching through analogy) [142], and visualization adaptations [144]. Although Camba et al. [33] studied the effectiveness of a peer challenge intervention in teaching students about one type of deception (inappropriate y-axis range), existing tools and techniques for improving visualization literacy primarily focus on *well-formed* visualizations and often low-level, straightforward tasks. Higher-level tasks like the ability to *critically* interpret visualizations is essential for both improving visualization literacy and combating visualization misinformation. We explore the effectiveness of teaching skills such as critical interpretation through active intervention (e.g., engaging students in the creation and problem-solving process [33]), as prior work [33] has suggested that this approach has potential for training in deception identification.

4.2.3 Serious Games

A serious game is a game that is not for pure entertainment and is often applied in education, health, business [61], with the main function of training players to learn and improve on real-world skills [139]. In the context of misinformation, researchers have built serious games of various kind that aim to teach people the skills to combat misinformation, based on inoculation theory, which suggests that people can develop resistance to false persuasion through pre-exposure to weakened forms of similar challenges (e.g.,

misinformation) [119]. For instance, the Bad News game interacts with players mainly through dialogue and asks them to use tactics that can spread misinformation in a simulated game setting [141]. Another serious game targeted at combating political misinformation tactics, Harmony Square [140] uses more narrative and asks players to use the tactics to disrupt the harmony of a fictional town. GO VIRAL! is serious game that aims to combat COVID-19 misinformation and has been spotlighted by the World Health Organization (WHO) [78]. Cranky Uncle [45], on the other hand, relies more on quizzes, where it first explains the rhetorical techniques and fallacies often used in misleading content and asks players to answer questions to cumulate points. Within the visualization community, researchers have explored using games for education [157, 6], some of which have also been mentioned in Section 4.2.2 and focus primarily on the teaching of well-formed visualizations. While these prior efforts offer valuable insights on using serious games in education, they mainly focused on text-based misinformation (e.g., textual fake news) and well-formed visualizations. The power of persuasion that data visualizations have and the prevalence of chart-induced misinformation demand specific intervention methods targeting visualization misinformation.

4.3 Methods

We conducted a formative design study with with high school students from an after school program ran by a non-profit organization in the Chicago area to iteratively develop our game. The game is designed to be played in person. This is to foster more interaction between players and reduce the limitations that might exist with computer-based games (e.g., access to computers). The design workshops took the form of in-person playtests with the researcher prompting for feedback during and after gameplay. The components of each workshop consist of *materials* (e.g., dataset, worksheet) and *mechanics*, which are

Table 4.1: The descriptions of the learning objectives mapped to corresponding Bloom’s taxonomy levels.

Level	Description	Learning Objective
Knowledge	Recall basic facts and definitions	Be able to correctly identify what the misleader is
Comprehension	Understand the information in context	Be able to articulate a surface level reason why the misleader is misleading
Application	Apply knowledge to a new problem or represent it differently	Be able to apply the same misleader in new chart types not seen before
Analysis	Break down a concept into parts and understand their relationship	Be able to break down how the misleader misleads based on specific vis principles/conventions/perception, etc
Synthesis	Use knowledge to create something new	Be able to synthesize classes of misleaders
Evaluation	Judge the value of information, backed by evidence	Be able to evaluate the fit, existence, and/or severity of a misleader

the instructional rules that the players are asked to follow during gameplay. Each game lasted about 15 to 20 minutes. In the beginning of each workshop, we briefly explained the basics of data visualizations (e.g., what is a data visualizations, why is it important for it to not be misleading) then provided the materials and described the game mechanics.

4.3.1 Learning Objectives

We used Bloom’s taxonomy [99] as a basis for developing the set of learning objectives relevant for gaining expertise in reasoning about misleading data visualizations. We mapped our learning objectives to each level, as we describe in Table 4.1:

Early design workshops suggested that our group of high schools students were fa-

miliar with basic chart types but not as familiar with the ways charts can be misleading. We therefore target the *Knowledge* and *Comprehension* levels in the game to avoid overwhelming the students with too much content. In Section 4.4, we describe the different game designs we tested out and what we observed from actual gameplay.

4.4 Design Variants

We ran four design workshops¹ to test out our design variants, each informing the design of the next variant. This iterative process allowed us to test out individual (variant I), competitive (variant II), team-based (variant III), and competitive + team-based designs (variant IV). We used commonly known games as inspiration so that players can learn the rules quickly. Variants I and II were part of the first workshop we ran with our group of high school students. We ran variant III in the second workshop and variant IV in the third workshop.

4.4.1 Variant I: Individual Design

We started with a design that only requires simple interactions. One of the first game designs we tested was TeleCharts, inspired by a party game commonly known as Telephone.² In Telephone, players are lined up to whisper a message to the next player. The core of Telephone is observing how the transmitted message is distorted at the end after the first and final player compares their messages. We based the design of TeleCharts on Telephone because players are likely to already be familiar with the rules, which would allow them to focus on the content of the game rather than learning the rules. In the context of TeleCharts, the players are (mis)transmitting a chart, and the goal is for players

¹The workshop materials and outcomes can be found in supplemental materials: <https://osf.io/56ac4/>.

²<https://www.wikihow.com/Play-the-Telephone-Game>

to realize how distorted (misleading) the visualization can become. This is an *individual design* because each player is on their own, requiring minimal interaction.

4.4.1.1 Mechanics and Materials

Three players participated at a time. In the game, each player is given a blank sheet of paper and colored pencils. Only the first player sees the correctly constructed chart, which is the seed chart³ for the game. The subsequent players only see the chart produced by the player immediately before them. Each player is asked to draw a misleading chart based on the chart they see. After the last player draws their chart, the moderator reveals the seed chart to everyone and facilitates a discussion on how misleading the resulting chart became, including its progression through intermediate charts.

4.4.1.2 Observations and Findings

The entire gameplay was relatively quiet, as each player is waiting for the previous player to finish in sequential order. Because there were no particular instructions on what techniques to apply to the chart, the players were expected to be creative when they make the chart misleading. This led to some players having difficulty coming up with a technique to apply, especially for players toward the end of the line. For example, a player who was unsure of what to do to a visualization that already had an inverted x -axis ended up swapping the x and y axes, leading to “date” on the y -axis and a quantitative variable on the x -axis. However, the player kept the base of the bar chart on the x -axis, which was nonsensical. Some ways to remedy this include not asking players to repeatedly apply misleading techniques to the same chart, which can lead to exhaustion of techniques, or providing a list of techniques players can choose from.

³The seed chart can be found in supplemental materials: <https://osf.io/56ac4/>.

4.4.2 Variant II: Competitive Design

Another game design we tested was ChartWolf Mafia, inspired by a party game commonly known as Werewolf or Mafia.⁴ We aimed to playtest a variety of possible designs, so we chose to test this because it contrasts the individual design and requires more interaction among players. Similar to variant I, we chose a well-known party game for ease of understanding. Werewolf is a competitive game between players who have a secret agenda and players who do not, and the core of this game is to deduce who among the group are working against everyone else. In our version of ChartWolf Mafia, a subset of players receive the secret mission of creating a misleading chart while the rest of the group tries to create a correct chart that allows them to arrive at the right answer to a given question. This is a *competitive design* because the majority of the gameplay involves two competing groups with hidden roles.

4.4.2.1 Mechanics and Materials

Three players participated in ChartWolf Mafia at a time. The materials of ChartWolf Mafia consisted of blank sheets of paper, colored pencils, a multiple-choice visualization question adapted from CALVI from Chapter 3, and a mechanism for randomly assigning the ChartWolf, which in our case was sticky notes. Each sticky note contained either a check mark or the letter corresponding to the correct answer of the multiple-choice question. The player who drew the correct answer is the ChartWolf, meaning their goal was to mislead the group into choosing a wrong answer.

⁴<https://www.wikihow.com/Play-Mafia>

4.4.2.2 Observations and Findings

Because we tested variant II during the same workshop as variant I, the same set of players played both variants, and players were more engaged with ChartWolf Mafia compared to TeleChart. This is likely due to the competitive nature of ChartWolf Mafia. Because two sides are competing against each other, players were actively trying to make a case for their argument, so they were keen on finding evidence that supported their arguments. We identified several challenges while observing the gameplay. For example, the ChartWolf could use strategies other than manipulating the visual elements of the chart, such as making the question or the options ambiguous to shift the focus. Players can get too caught up on the options, making the argument about testing strategies instead (e.g., “how often would you choose ‘none of the above’”). The ChartWolf can also run out of misleading techniques to try. This may be due to a limited repository of techniques and can lead to them giving up early. Both variant I and II encountered this issue of running out of techniques to apply, which is unsatisfactory and affects the gameplay. We detail how we tried to resolve this issue in the next variant.

4.4.3 Variant III: Team-Based Design

The third variant we designed was ChartDetective. Players split into two teams to collaboratively create a misleading chart by applying techniques from a given list. Their goal is to mislead the other team. One of the challenges observed from variant II was players focusing too much on the options instead of visualization design. In variant III, we used open-ended questions rather than multiple-choice questions. This is a *team-based design* because each team is working together to collaboratively design a misleading chart.

4.4.3.1 Mechanics and Materials

Each team consisted of three to four players, and players were randomly assigned to each team. The materials of ChartDetective consisted of an instruction sheet, a dataset with a question of interest, a deck of cards from 1 to 10 representing different misleading techniques⁵ inspired by prior work [72, 75, 113, 121] (e.g., use inconsistent or confusing scale intervals, flip the axis), blank sheets of paper, and colored pencils. Each card corresponds to a misleading technique, which are listed on the instruction sheet. Players first draw 4 cards from the deck and then discard one card, leaving 3 technique cards they must apply to the final chart they design. This is to give some freedom to the players in case some of the techniques are difficult to apply. The team is given 15 minutes to design their chart before swapping with another team. Within each team, players discuss among themselves to decide how to apply the techniques they drew in a non-obvious way then collaboratively draw their chart on the given blank sheets of paper. The team that successfully tricks the other team into answering a wrong answer gains a point.

4.4.3.2 Observations and Findings

The design change to open-ended questions seemed to allow for more discussion among the players, and the discussions were about the data rather than properties irrelevant to the question and data (e.g., multiple-choice options, testing strategy). To address the challenge of running out of misleading techniques in both variants I and II, we built in the card system to give the players a techniques bank to work with. We intentionally kept the technique descriptions broad to avoid limiting the players to specific ways of applying the techniques. However, we observed some techniques were harder for high school students to grasp just from a simple description. For example, some players were unclear

⁵The set of techniques can be found in supplemental materials: <https://osf.io/56ac4/>.

about “aspect ratio” in the technique “manipulate aspect ratio to distort relationships”. Additionally, because players drew the techniques randomly, some techniques may not apply in some questions. For example, “combine/split categories to change grouping” was harder for players to apply when the dataset contained years and sales with no obvious categories. Some teams used the strategy of brainstorming together then drawing the chart on one sheet of paper, while other teams decided to draw their own versions first then choose one among all of the versions as the final answer. The teams that brainstormed together generally seemed more engaged. This suggests that we should implement mechanisms (e.g., clarifying in game instructions) to promote more collaboration.

4.4.4 Variant IV: Competitive and Team-Based Design

Based on feedback from the prior design variants, we refined the game to incorporate properties from both competitive and team-based designs. In this variant, Newsroom, each player receives a role of either *Journalist* or *Informant*. The goal of the journalists is to uphold the newsroom’s reputation by publishing good quality charts, whereas the informants’ goal is to publish misleading charts to damage the newsroom’s reputation. One major challenge observed from the previous design was that simple technique descriptions were too abstract for the high school students to translate into concrete applications. We therefore included examples of each technique in the form of cards to visually illustrate how a technique might be applied to a particular chart type. This is both *competitive* and *team-based* design because there are at least two players with the same role, but the informants are secretly competing with the journalists.

4.4.5 Mechanics and Materials

Each player drew a role card and got assigned as either a journalist or an informant (Figure 4.2 A and B). The materials of Newsroom consisted of a blank sheet of paper, colored pencils, and three decks of cards: dataset, visualization, and action cards. In the beginning of the first round, each player draws one dataset card, one visualization card, and three action cards. The dataset deck contains distinct datasets (e.g., Figure 4.2.C). The visualization deck contains visual illustrations of a chart type and a misleading technique (e.g., Figure 4.2.D). We compiled a total of 7 commonly used chart types and 7 misleading categories from Chapter 3.⁶ The action cards consist of “publish” (publish a chart), “audit” (initiates discussion on a chart and vote on whether to remove the chart), “source” (publicly reveal the dataset of a chart). The visualization card acted as a suggestion with a visual example that the player could choose to use. The game begins with each player drawing a chart based on their role. Once every player finishes drawing the chart, each player then plays an action card on a chart they choose to either “publish”, “audit”, or “source”. During subsequent rounds, each chart on the table gains one influence point. In the round that a chart gets published or removed, the team gains its associated points if a good quality chart is published or loses its associated points if a misleading chart is published. The opposite is true in terms of points if the charts are removed after auditing. If the total number of points drops below a pre-determined threshold, then the informants win. If the total number of points rises above a pre-determined threshold, then the journalists win.

⁶The categories included: inappropriate scale direction, inappropriate scale order, inappropriate scale range, inappropriate scale intervals, inappropriate selection, inappropriate aggregation, inappropriate annotations.

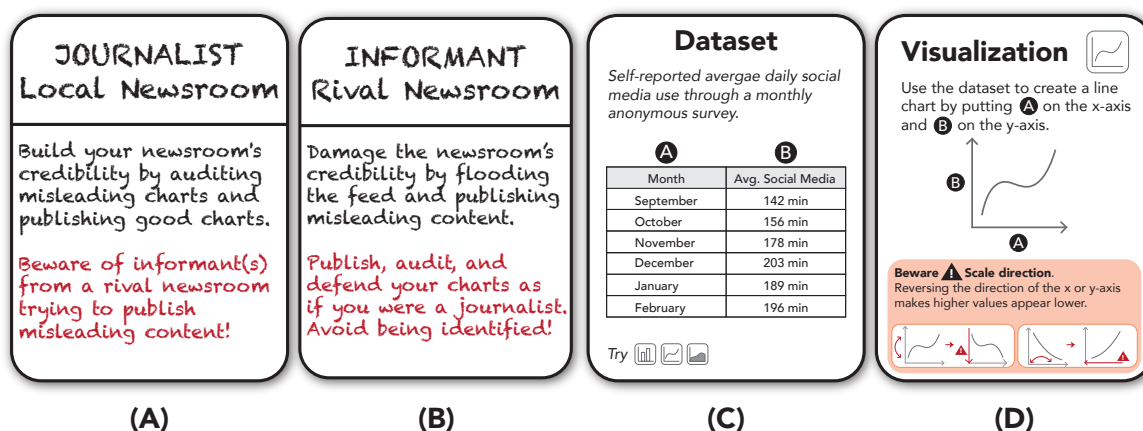


Figure 4.2: Representative examples from the cards deck. (A) and (B) are the role cards. (C) is an example of a dataset card. (D) is an example of a visualization cards, containing the chart type (top) and the visually illustrated misleader example (bottom).

4.4.6 Observations and Findings

The card deck offered more structure to the game and added randomness, which can contribute to the game re-playability. However, this also added more information for the players to track, which complicated the instructions compared to previous variants. Players frequently expressed difficulty in keeping track of the game instructions and scoring mechanism during gameplay. While the visualization cards were more helpful than text descriptions, them acting as mere suggestions led to several challenges: (1) it was difficult for the moderator to keep track of the submitted charts while keeping score and running the game, (2) allowing players to be creative increased the chances of creating nonsensical charts. One player who was the informant ended up modifying the data when creating the misleading chart, which led to them keeping their cover because it would be almost impossible to tell its misleadingness without carefully examining the dataset. This is not a misleader that we aimed to teach. Another challenge we did not anticipate was that hand-drawn charts were identifiable, especially when the players are already familiar with each other. This made it hard for the informants to keep their cover across multiple

rounds, because one suspicion can propagate across the rounds. Overall, players were noticeably engaged with the game and wanted to finish, suggesting that the both competitive and team-based design preserved the engagement property. Using what we learned throughout the design workshops, we describe the refined game features and reflect on the key takeaways for developing serious games that aim to teach critical thinking.

4.5 Synthesis of Game Features and Design Implications

We outline several game features that have worked in our Newsroom game and reflect on the insights learned through our iterative design process with high school students. To summarize these game features, we recorded what worked and did not work across the four variants. We identified features that (1) worked well across multiple design variants or (2) solved a previous challenge identified in the design workshops.

Secret agendas encourage players to adopt the evaluative mindset needed to identify misleaders. One of the learning objectives was to enable students to identify misleaders, which first requires an evaluative mindset. In other words, students should not read charts with the assumption that every chart they encounter is well formed. The secret agendas embedded in the game served as a way to break the assumption that every chart they see will be well formed. For instance, in the game Newsroom, each player has a secret role throughout the game, and they know that both good and misleading charts are being introduced into the pool. This incentivizes the players to not only evaluate the charts but also the identity of the other players. This continuous evaluation builds the evaluative mindset needed to identify misleading visualizations.

Demonstrating concepts concretely prepares players to identify misleaders. In order

to identify misleaders, students need to be able to first gain a sense of the space of misleaders. One major challenge we encountered during the game development process was that it was hard for our high school students to translate abstract concepts to concrete applications, making it hard for them to build a vocabulary of the types of misleaders. This persisted until we created the visualization deck in Newsroom, where it contained concrete examples that illustrate how a particular misleader works on a chart type. These concrete illustrations can help students build a better understanding of the space of misleaders, which they can draw upon when they need to identify misleading visualizations.

Creating a chart lead to deeper reasoning behind misleadingness. Letting students engage in the chart creation process allows them to see how a misleader can be misleading. When students were only looking at the potentially misleading charts, their reasoning for “why” was often limited to “it looks weird”, whereas during chart creation, they are forced to pay attention to particular parts of the chart such as scale. This was especially evident with the visualization cards that contained visual examples. Creating charts by hand also allowed students to slow down and engage with the chart more. One challenge with hand-drawn charts is that they are identifiable among groups of players who are familiar with each other, as we described in our observations from playtesting Newsroom. This could be resolved by limiting to one round so that charts being identifiable does not affect the gameplay. For example, a round would end once all charts are either published or removed from the pool using their respective action cards.

The action cards deck allows players to practice the learning objectives. The refined action deck includes “publish”, “discard”, “re-examine”, and “source” (Figure 4.3). Both “publish” and “discard” automatically publishes or discards a chart, respectively. Because players must perform an action, they must engage with the activity of deciding

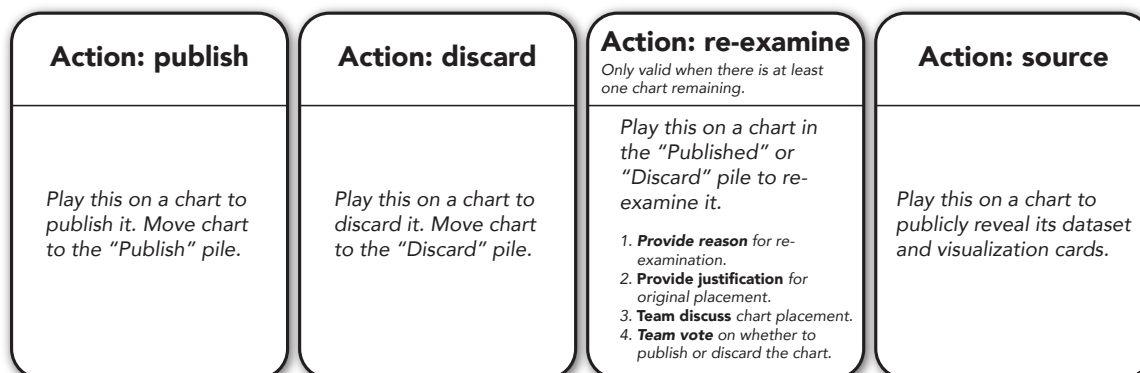


Figure 4.3: Cards in the refined action deck.

the status of a particular chart to evaluate whether it is misleading. “Re-examine” initiates discussion on a published or discarded chart with the possibility of changing the status of the chart, depending on the majority vote. The step-by-step process outlined on the “re-examine” card allows players to actively engage in “reasoning”, “justifying”, and “discussing” charts with each other, which help them learn to reason about why a chart might be misleading (and convince others).

4.6 Discussion

4.6.1 Broadening the Game and Coverage of Learning Objectives

Although we described a learning objective for each level in the Bloom’s taxonomy, it would be challenging to cover every learning objective in a single game, especially given limited time we have during each workshop. We thus chose to focus on the *Knowledge* and *Comprehension* levels for our target user group. The Newsroom game could be extended and made more challenging to cover more learning objectives. For example, the visualization cards in Newsroom could be used to incorporate the learning objective in the Application level (“be able to apply the same misleader in new chart types not seen before”).

Instead of enforcing the use of the chart types shown on the visualization card, the player must apply the misleader onto a chart type not shown on the visualization card. The learning objective in the Evaluation level (“be able to evaluate the fit, existence, and/or severity of a misleader”) could be incorporated by associating different numbers of points to the visualization cards, such as based on prior work on the difficulty of misleaders [72]; more difficult misleaders would worth more points. This could prompt players to tackle charts with more challenging misleaders to earn more points. Components of Newsroom could be extended and adjusted based on the focus of learning objectives. Future work could build on the basic version of Newsroom to broaden or customize the focus based on educational needs.

4.6.2 Challenges and Opportunities for Visualization Literacy

Assessments

CALVI served as a major inspiration and prior work for this game. However, the game development process also surfaced some challenges and opportunities useful for advancing visualization literacy measurements. Part of the reason why we could not apply CALVI directly to evaluate the effectiveness of the game is that CALVI does not cover some of the learning objectives. For example, CALVI is less suited for capturing reasoning relevant for the *Comprehension* or *Analysis* levels because it only contains multiple-choice questions. This echoes the call for modalities other than multiple-choice in Chapter 6. We would have to extend CALVI to include reasoning questions, which increases the test time and leads to another challenge in assessments: test format and ecological validity. In the context of a game or other informal learning environments, a standardized test would take the players out of the experience, which goes against the goal of informal learning, where the experience is often part of learning. CALVI, with its current standardized format, may

not be well-suited for assessing informal learning. Future work should expand test format to complement the intervention, such as a method for integrating assessments into interventions rather than a post-hoc evaluation. This would make evaluation more seamless and potentially improve the ecological validity of assessments and studies overall.

Chapter 5

AVEC: An Assessment of Visual Encoding Ability in Visualization Construction

This chapter is based on work published in ACM CHI 2025 [71] and is done in collaboration with Yuan Cui and Matthew Kay.

5.1 Introduction

Beyond critical interpretation skills, the ability to construct effective visualizations is also a critical component of *visualization literacy* [10, 23], analogous to the way in which writing is inseparable from textual literacy. With the rapid increase in data availability and accessibility, visualizations are not only created by professional designers [68, 151, 161], but also by members of the general public; for example, people who are not in data-oriented jobs or who might not be trained in visualization are nevertheless constructing visualizations at work (e.g., through Excel) or in a personal context (e.g., through personal visualization and visual analytics [88]). Despite increased participation in constructing visualizations and efforts to support non-experts [90, 156], the current focus of visualization literacy studies and assessments is still on visualization interpretation [106, 26, 72, 55, 130]. We lack valid measurements of construction-related abilities, which we need to effectively

measure, teach, and improve these skills in novices and amongst the general public.

Constructing visualizations, however, is a complex and multi-step process [34, 92], which includes transforming raw data into a form suitable for visualization (e.g., cleaning, filtering, and aggregating as necessary [162]), translating data variables to visual variables (e.g., setting the position of a data point by mapping data to desired x and y axes [116]), and transforming the visual encodings to convey a meaningful message (e.g., applying appropriate scaling functions or applying creative visual metaphors [167] to facilitate interpretation). All of these components make up the broad notion of a *visualization construction ability*.

Although there are many steps involved in visualization construction, one of its core components is selecting appropriate *visual mappings* [34]. We designate a person's *ability to appropriately map data to visual channel(s) for effectively answering data-relevant questions* their **visual encoding ability**. This ability, which is governed by established visualization principles [116, 34] and rules (e.g., the Grammar of Graphics [163]) and common to all visualization design, is critical for constructing well-formed visualizations and for visualization creators to master. We therefore focus on the visual encoding ability as a first step in systematically assessing construction-related abilities.

The complexity of visualization construction also translates into challenges in measuring construction-related abilities. The traditional multiple-choice question format cannot accurately represent the design space a visualization creator has to work with. Visually encoding data requires a designer to appropriately apply grammatical rules of visualization design and consider multiple data variables and visual encodings at once; the design space quickly explodes to thousands of possible combinations of data mappings. This *combinatorial* nature is inherent to selecting appropriate visual encodings and must

also be present in measurements of such abilities.

To create an assessment of the visual encoding ability in visualization construction, we:

1. **Created 9 initial items to measure the *visual encoding ability*.** To create the items, we first constructed an item design space consisting of 3 visualization tasks and 8 chart types. We used it to generate 38 candidate constructed-response items, each of which asks the test-taker to create a visualization for answering a true-or-false (T/F) statement about a dataset. We selected the 9 items that make up the initial item bank by varying the number of necessary data variables, mark types, and size of datasets.
2. **Developed a web-based assessment tool that can support the combinatorial nature of selecting appropriate visual encodings.** This tool allows test-takers to create a wide variety of visualization answers for the items, while still being **easy to use** without requiring prior knowledge of a specific visualization tool or programming language. This ensures that differences in test-takers' scores are due to differences in their abilities and not the usability of the tool, which contributes to the validity of the assessment.
3. **Created an autograder based on expert scores of visualization answers to our assessment items.** We invited an expert panel of 17 Information Visualization researchers to assess the content validity of candidate items and provide grading criteria for clusters of similar answers to the items.
4. **Conducted test tryout to assess the quality of our items and to refine the autograder.** Ninety-five online participants used our web-based assessment tool to an-

swer the 9 candidate items in a test tryout study. We used Item Response Theory (IRT) analysis to measure how easy or difficult each item is and how reliably each item separates people of different abilities. We also used the test tryout data to refine the autograder and assign final grades to participants' answers.

5. **Refined the bank to a final set of 8 constructed-response items, which have high content validity, have varying difficulty, and can differentiate people of different levels of ability.** We finalized the item bank of AVEC using feedback from the expert panel and results of the IRT analysis, removing one item with low content validity and high variance in experts' scores.

Reflecting on our efforts to go beyond multiple-choice formats, AVEC demonstrates the importance of and the complexities involved in measuring higher-level abilities in visualization literacy. Dealing with these complexities necessitates a line of future research that could transform both the ways that we assess visualization skills and our understanding of visualization literacy.

5.2 Related Work

5.2.1 Item Formats Beyond Multiple Choice

Existing assessments of visualization literacy often use multiple-choice items [26, 106, 72, 130, 55, 56]. However, simply choosing from a short list of answers is not an accurate representation of construction-related abilities, and thus is not a format that can sufficiently support measuring the ability to appropriately map data to visual channel(s). Constructed-response items, on the other hand, require a test-taker to “generate an answer rather than select from a short list of options” [15]. They are often used to measure

higher-level and more complex skills in literature, mathematics, and music [111]. While these items can better capture the skill of interest, they are harder to grade. The Educational Testing Service (ETS), for example, recognizes this and has applied automated scoring to support the grading of the constructed-response items [111]. In the context of a visualization assessment, we might opt to use a computer-mediated constructed-response item format, rather than purely asking test-takers to draw visualizations by hand, for example. Allowing test-takers to directly control the mapping of data variables to visual encodings serves as a proxy to hand-drawing visualizations and also preserves the combinatorial nature of selecting appropriate visual encodings (more details in Section 5.3.3).

5.3 Test Development: Items and Assessment Tool

Visual mappings are a core component of both the visualization reference model [34] and visualization grammar [163], making them essential for visualization creators to master. We dub *the ability to appropriately map data to visual channel(s) for effectively answering data-relevant questions* a person’s *visual encoding ability*, and focus on developing an assessment that measures this ability. We aim to design the assessment for non-experts, including members of the general public and students who are beginning to learn how to construct visualizations.

We use constructed-response items in our assessment because generating an answer—rather than choosing from a short list of answers—aligns better with the combinatorial nature of mapping data to visual channels. An *item* of AVEC consists of a constructed-response question that asks the test-taker to create a chart for answering a true-or-false (T/F) statement about a dataset. The format of each item is as follows:

Table 5.1: The data variables in each dataset. The size of the dataset for each item may be different, but each item asks about a subset of these variables.

Data variable	Description
Date	A recorded date (year-month-date) to track daily spending, income, or savings.
Month	The month.
Year	The year.
Income	The amount of income (\$) on each recorded date.
Savings	The amount of savings (\$) on each recorded date.
Spending	The amount of spending (\$) on each recorded date. In this hypothetical dataset, the person spends money on only one spending category each day.
Spending category	Daily spending category: Entertainment, Food, Transport, or Utilities.

”Create the best chart to use to determine whether this statement is true: [*T/F statement*]”.

Thus, for a diverse bank of items, we need a set of T/F statements whose reasonable visualization answers cover a diverse set of commonly-used visualization tasks and chart types (Section 5.3.1 and Section 5.3.2). Additionally, the items must be administered via an easy-to-use assessment tool that also supports the combinatorial nature of selecting appropriate visual encodings (Section 5.3.3).

5.3.1 Item Design Space

To systematically generate a set of suitable candidate T/F statements that will allow for a diverse set of visualization answers, we created a design space composed of a variety of visualization tasks and chart types.

Visualization Tasks

We started with the initial set of 8 tasks from VLAT [106], which was designed to assess the visualization interpretation ability of the general public. We **excluded 2 tasks that are lower-level (involving local point judgements)** to focus on higher-level tasks (involving visual patterns with many data points), because the visual encoding ability that we intend to measure involves whether people can create an appropriate *visual form* to answer a data-relevant question. Specifically, we excluded *Retrieve Value* and *Find Extremum* because they are single-point judgements, so we consider them lower-level. Additionally, we aimed to create a diverse set of items that would lead test-takers to construct a variety of visualization types. To this end, we **merged tasks that would generally lead to similar visualizations**. Namely, we merged *Determine Range*, *Find Anomalies*, and *Find Clusters* into *Describe Distributions*.¹ We kept the remaining 2 tasks from the original set of 8 as-is (i.e., *Find Trends / Correlations* and *Make Pattern Comparisons*) because they are higher-level tasks and are distinct from the task of describing distributions. As a result, our reduced set contains 3 tasks:

- *Describe Distributions*: describe the spread of the values of a variable and/or how frequently different values occur.
- *Find Trends / Correlations*: find the relationship between two variables or how a variable changes over time.
- *Make Pattern Comparisons*: compare patterns (e.g., data distributions, trends, or relationships) in the dataset.

¹We renamed the original *Characterize Distribution* from VLAT to better reflect this merged set, which includes tasks related to describing the spread of data values rather than only characterizing distributions.

Chart Types

To identify a set of chart types that the visualization answers should cover, we started with a set of 12 chart types from VLAT [106]. We aimed to compile a diverse set of chart types that cover our set of 3 tasks and can support varying the difficulty levels of items in the bank. However, for usability purposes, we also wanted to keep our assessment tool's interface consistent across items so that test-takers do not have to learn different interfaces for different items—in order to reduce the influence of the interface on a test-taker's performance. Out of the 12 initial chart types, a majority of chart types have x and y axes, which could be kept consistent across items. Thus, we retained 8 Cartesian chart types,² as this would allow us to **have a diverse set of chart types while maintaining a consistent chart-creation interface**. Our resulting set contains 8 total chart types,³ and each supports a maximum of 4 encoding channels (i.e., x -axis, y -axis, color, size).

5.3.2 Item Generation

We generated one T/F statement for each candidate item using the item design space. For all T/F statements, we used a dataset in the context of personal finance, because this topic is relevant to most people. We generated datasets of varying sizes (e.g., a year's or a month's data) to introduce some diversity between items, but all of the datasets contain the same set of data variables, as described in Table 5.1. Each T/F statement asks about a subset of the variables in this dataset. This reduced the need for test-takers to learn new datasets during the assessment: how well one can get familiar with a new dataset is not necessarily part of our measure of the ability to map data to visual channels.

²Between stacked bar chart and 100% stacked bar chart, we only kept stacked bar chart. Because of their similarity, keeping both would add little value to having a diverse set.

³The 8 chart types are bar chart, stacked bar chart, histogram, line chart, scatterplot, bubble chart, area chart, and stacked area chart.

Task	ID	True-or-false statements	Mark Types
Describe Distributions	DD.1	Daily income rarely went over \$300.	   
	DD.2	Total monthly utilities spending was generally lower than \$300.	 
	DD.3	Total yearly spending rarely exceeded \$10,000.	 
Find Trends Correlations	FTC.1	Daily spending often fluctuated between about \$10 to about \$30.	   
	FTC.2	Daily savings was generally higher when income was higher.	
	FTC.3	Total monthly utilities spending strictly decreased from Jul to Oct.	 
Make Pattern Comparisons	MPC.1	The range of daily transport spending was wider than food spending.	  
	MPC.2	The number of times that daily spending was around \$50 increased from 2020 to 2023.	  
	MPC.3	Daily food spending consistently decreased at the same rate as transport spending.	 

Figure 5.1: The set of T/F statements in our initial item bank. Each statement has its own set of mark types that we think are most relevant based on its relevant data type(s) and task, which covers a variety of chart types.

We first generated T/F statements with a differing number of *necessary data variables* (i.e., variables that must be visualized to answer the T/F question correctly) for each combination of visualization task and chart type. This was our attempt at generating items with varying difficulty. The ultimate difficulty of these items will be determined by IRT analysis (Section 5.6.5). For 6 of the 8 chart types, we generated two candidate T/F statements per combination of task and chart type, resulting in 36 items ($2 \text{ items} \times 6 \text{ chart types} \times 3 \text{ visualization tasks}$). For stacked bar chart and stacked area chart, we generated one candidate T/F statement each (under the task of *Make Pattern Comparisons*), because they require three and only three necessary variables. Thus, we created $36 + 2 = 38$ candidate T/F statements. We referred to a summarization of suitable chart types for different data types by authors of VLAT [106] and inferred the set of most relevant mark types for each T/F statement. For instance, scatterplots are less likely to be useful for items with T/F statements that require yearly aggregations, and line charts are more likely to be useful

for T/F statements that require order in the data (e.g., trend over time).

We then reviewed the set of candidate T/F statements and noticed redundancies that involved the same type of statement that only differ in the quantitative variable of interest (e.g., “[*quantitative variable*] rarely went over [*numeric value as threshold*]”). To remove redundancies and ensure a variety of items, we selected T/F statements from each visualization task category while ensuring a diverse set of relevant chart types,⁴ a varying number of necessary variables, and a varying size of datasets. This allowed us to arrive at a set of T/F statements that can lead to a diverse set of visualization answers that also cover all 3 visualization tasks. The result is a set of 9 items, with 3 items per task, as shown in Figure 5.1.

5.3.3 Assessment Tool

We need an assessment tool that is easy to use, able to support the combinatorial nature of selecting appropriate visual encodings, and expressive enough to allow test-takers to construct a good answer to the items. This led to the following design considerations:

- **(DC.1) Does not require coding expertise:** The target test population for AVEC includes non-experts in visualization, so we cannot assume they have any programming experience.
- **(DC.2) Easy to use:** Our goal is to assess visualization-related abilities, not expertise with the interface—therefore we need an assessment tool that non-experts can pick up quickly with little training. This will help ensure that differences in test-takers’

⁴We did not end up including T/F statements from bubble chart combinations due to a conflicting property of such T/F statements: the statements tend to be overly long and difficult to parse, but the visualization answer always includes all 3 of the quantitative variables in the dataset, which can make it easy to arrive at the answer.

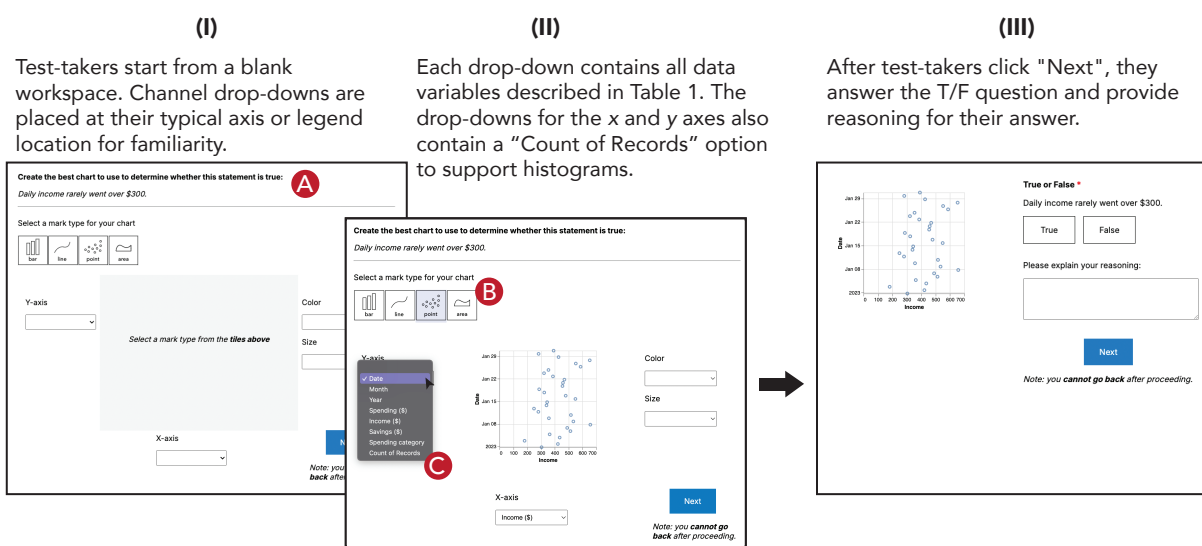


Figure 5.2: The interface of the assessment tool, which displays the item's question text (A) and includes an interactive area for constructing the visualization answer that allows test-takers to select their desired mark type (B) and contains drop-downs (C) to control the mapping between data to visual channels. On a separate page after a test-taker constructs their answer, they are asked to select an answer for the T/F question.

scores are due to differences in their abilities and not the usability of the tool.

- **(DC.3) Familiarity of user interface:** Similar to the previous consideration, we do not want our measurement confounded by a test-taker's familiarity with a specific visualization tool—therefore, we require an interface that is familiar to a broad set of users (e.g., one which uses standard UI widgets).
- **(DC.4) Consistency across items:** An assessment interface that is consistent across items reduces the need for the test-taker to learn new parts of the interface during the test, reducing noise in our measurement.
- **(DC.5) Expressive:** Even though we need to keep the assessment interface simple, it must be able to support the construction of a wide variety of chart types to give test-takers the flexibility to construct visualization answers of varying quality.

We went through an iterative design process to develop an assessment tool that meets all of the above design considerations. To broaden the user group for our assessment, we eliminated the need for programming experience by keeping controls at a high level: users only need to decide which visual channels they want to map data variables to (**DC.1**). Behind the scenes, we manipulate Vega-Lite [145] specifications according to the test-taker's desired mappings. The color schemes we used are color blind safe.⁵

Early versions⁶ of the interface included individual controls in a drag-and-drop format (similar to visualization tools like Tableau) for different data transformations (e.g., binning, sorting, mean, sum, count), scale transformations (e.g., reversing or truncating scales), and a larger set of encoding choices (e.g., sequential, diverging, qualitative colors). However, when piloting the interface, we found that the drag-and-drop features were too overwhelming for some users to navigate and were difficult to learn through short training. As our goal was to create an interface that did not require extensive training (**DC.2**) and which did not simply measure a user's familiarity with some particular visualization tool (**DC.3**), we changed the interface to use more typical drop-down widgets and refined our scope to focus on visual mappings, the core component of the visualization reference model [34]. We refined the tool to support a maximum of 4 encodings (*x*-axis, *y*-axis, color, and size) and 4 mark types, which together determine the resulting chart type. The 4 mark types are *bar* (for bar chart, stacked bar chart, histogram), *line* (for line chart), *point* (for scatterplot, bubble chart), and *area* (for area chart, stacked area chart) (**DC.5**, Figure 5.2.B).

In the refined version of the assessment tool, we automatically apply aggregation

⁵See supplemental materials for the results from Adobe's color accessibility checker: <https://osf.io/hg7kx>.

⁶See supplemental materials for details of early iterations: <https://osf.io/hg7kx>.

functions for the purpose of supporting the creation of all of the chart types in our design space. In cases where the desired chart type requires aggregation, we manipulate Vega-Lite specifications to apply aggregation depending on the selection of mark types and data variable types. We included 2 aggregation functions:

- *count* is included to support histograms. If a test-taker selects the *Count of Records* option for one of the positional axes, then we apply the count aggregation and bin the variable on the other axis.
- *sum* is included to support stacked bar and stacked area charts. If a test-taker selects a *bar* or *area* mark, maps a categorical variable to one of the positional axes and/or color and/or size, and maps a quantitative variable to the other axis, then we automatically sum the quantitative variable within combinations of the categorical variables.

Having a minimum set of built-in aggregation functions simplified the interface compared to earlier prototypes, which required test-takers to manually specify aggregation functions, but which participants found difficult to learn in pilots (**DC.2**). We further discuss the importance of the role of the assessment tool in measurements of complex construction-related skills in Section 7.1.

Throughout the assessment, we display each item in the same format for consistency (**DC.4**, Figure 5.2.A): the question text followed by an interactive area for the test-taker to construct their visualization answer (Figure 5.2). We placed the drop-downs close to the channel they control (or, for color and size, where their legends might often appear on simple charts), taking advantage of familiar spatial associations (**DC.3**, Figure 5.2.C). On a separate page, after the test-taker constructs the chart, we ask them to select the

answer to the T/F question and provide their reasoning (Figure 5.2.III). We included this for us (and future test administrators) to use as a resource to revise the T/F statement if needed, but the test-takers' answers to the T/F questions do not affect the assessment of their visual encoding ability.

5.4 Test Development: Visualization Answers

The combinatorial nature of selecting appropriate visual encodings implies there should be a large number of possible answers of varying quality for each item in our assessment. Unlike the traditional multiple-choice format—where the correctness of the answer is binary—we need a grading approach that allows test administrators to evaluate the quality of a broad spectrum of possible visualization answers.

It would be time-consuming to manually grade each possible visualization answer our assessment tool can generate. Therefore, we built an autograder for each item. Specifically, we (1) generated the complete set of possible visualization answers (as Vega-Lite specifications) to each question (Section 5.4.1); (2) used established rules and design principles from the literature [40, 116] to eliminate nonsensical answers, which will receive the lowest score (Section 5.4.2); (3) grouped the remaining answers into clusters we believe might receive similar scores from experts (Section 5.4.3); (4) randomly selected visualization answers from each cluster to form representative sets of visualization answers for each item. These representative sets of answers will be scored by an expert panel (Section 5.5). These expert scores will allow us to refine clusters (if needed) and will serve as the basis for our autograder.

5.4.1 Complete Sets of Visualization Answers

Our assessment tool, as explained in Section 5.3.3, gives test-takers the flexibility to map 7 data variables to 4 encoding channels and choose among 4 mark types. The test-taker manipulates each encoding channel through a drop-down list and is allowed to leave any channel “empty”. Thus, with the “empty” option, there are $(4 \text{ mark types} \times 8 \text{ options on } x\text{-axis} \times 8 \text{ options on } y\text{-axis} \times 8 \text{ options on color} \times 8 \text{ options on size})$ 16,384 combinations. In addition to data variables, test-takers can also map *Count of Records* to one positional axis and a data variable to another to create histograms. This generates an additional $(4 \text{ mark types} \times 7 \text{ data variables on } x\text{-axis} \times 8 \text{ options on color} \times 8 \text{ options on size})$ 1,792 combinations,⁷ yielding $16,384 + 1,792 = 18,176$ possible visualization answers for each item.

5.4.2 Initial Elimination

It would be impractical to ask the expert panel to rate the complete set of 163,584 visualization answers (18,176 answers for each of the 9 items), so we reduced the set of visualization answers for each item using the following criteria:

- **Whether all of the necessary data variables for that item are in the visualization specification:** Visualizations that do not contain all of the variables asked about in the T/F question statement would not be able answer that item, so we eliminate these answers.

⁷We fixed *Count of Records* to the y -axis when generating the combinations because versions with *Count of Records* on the x -axis would be redundancies that add little value to summarizing the grading rubric, which were combinations we could eliminate immediately to reduce the answers under consideration.

- **Whether the visualization passes the linter rules:** We reviewed linter rules from VizLinter [40] and only included the rules that pertain to the functionalities of our assessment tool. For example, the rule of *not using both bin and aggregate on the data at the same time* was excluded because our assessment tool automatically bins or aggregates depending on the data variable mapped, so it would not be possible to violate this rule. As a result, we curated the following three rules to use for this initial elimination:

- Use different fields for x -axis and y -axis
- Use no more than one continuous data in the x and y channels for mark ‘bar’ and ‘tick’
- Mark ‘bar’, ‘tick’, ‘line’, ‘area’ require some continuous variable on x or y

Additionally, the assessment tool allows for some combinations of mark type and data variable mappings that would lead to visualizations that make it *theoretically impossible* to answer the corresponding T/F question; we eliminated these answers. For example, visualization answers that use the *bar* or *area* mark and contain *Month* or *Year* were eliminated for items that ask about correlation between daily values of two variables (FTC.2), as aggregating by *Month* or *Year* would not allow the test-taker to judge daily correlation (see supplemental materials⁸ for a representative *theoretically impossible* answer for each rule). The visualization answers eliminated during this phase are of poor quality.

In total, 160,475 visualization answers failed the initial elimination criteria, leaving 3,109 answers that must be graded (see Figure 5.3 for a breakdown by item).

⁸<https://osf.io/hg7kx>

QID	Eliminated	Clusters (no particular order)																							
DD.1	16,921	24	4	30	30	60	60	72	6	52	14	240	256	78	128	2	40	10	18	54	50	6	12	9	
DD.2	18,144	2	9	9	4	2	2	2	1	1															
DD.3	17,940	20	1	1	28	2	18	1	1	112	10	22	1	1	2	1	1	4	1	1	6	2			
FTC.1	17,880	72	6	14	78	2	40	10	18	50	6														
FTC.2	17,648	4	204	60	4	64	64	78	50																
FTC.3	18,126	18	2	9	9	4	2	2	2	1	1														
MPC.1	17,785	12	24	22	13	54	60	22	96	30	6	2	6	8	32	2	2								
MPC.2	17,931	28	2	10	14	2	28	30	110	2	18	1													
MPC.3	18,100	20	8	20	8	2	6	10	2																

Figure 5.3: The left-most column in gray indicates the number of answers that were initially eliminated for each item. The set of numbers in colored boxes indicates the number of visualization answers that were in each item-specific cluster. The clusters are unordered; we use the expert scores to determine the relative order after conducting the expert panel (Section 5.5).

5.4.3 Answer Clusters

It would still be impractical to have experts rate all 3,109 remaining non-eliminated answers. Instead, we first created clusters of visualization answers that we consider to be similar to each other, have experts score answers sampled from those clusters, then build our autograder based on those scores.

To build the clusters, we first partitioned the visualization answers to each item into broad clusters using high-level rules, such as *Do core quantitative variable(s) use positional encodings?*. We then built a tool to display all of the visualization answers in a given cluster, which allowed us to quickly look through clusters and judge their quality and similarity. The authors reviewed all clusters independently, then iteratively refined the clusters and high-level clustering rules, reaching consensus after regular discussion. The resulting high-level rules closely aligned with prior work on the ranking of perceptual tasks [116].

Next, we split the initial broad clusters into smaller clusters based on our judgments

of the effectiveness and expressiveness of the visualizations. For example, a scatterplot with both necessary quantitative variables on positional axes is more effective for judging correlation than a plot with both variables mapped to color and size channels (FTC.2). Thus, we split the broad clusters based on properties of answers that could impact the visual form of the visualization: mark type, what data type is mapped to the encoding channels, and (for stacked bar or area) the orientation of the chart (orientation of a stacked bar can affect whether a subgroup of interest is aligned to the baseline, and therefore the difficulty of estimating values from that subgroup).

While forming the *clusters*, we further identified visualization answers that make it *practically impossible* or extremely difficult to answer the question statement and eliminated them based on item-specific rules. For example, it would be practically impossible to determine the correlation between two quantitative variables if the mark type is *bar* and one of the necessary quantitative variables is mapped to the *size* channel instead of the positional axis (FTC.2).

The resulting *clusters* for each item represent sets of visualization answers that we would expect to receive similar scores. However, we wanted to both validate our clustering and receive broader expert input to determine final scores. Therefore, we formed an expert panel to collect experts' scores on answers sampled from these clusters.

5.5 Expert Panel

The goal of the expert panel is threefold: (1) collect expert ratings of relevance to compute the content validity index (CVI) of each item (CVI is a measure of how well an item on a test measures what it is intended to measure [131]); (2) gather expert criteria for assessing the

quality of the visualization answers in order to validate our clustering scheme; and (3) use experts' scores of visualization answers sampled from our clusters to build autograders. We recruited 17 researchers in Information Visualization who have obtained ($N = 13$) or are obtaining ($N = 4$) a doctorate.

The expert panel was conducted on Qualtrics. For each item, we generated 5 representative sets of visualization answers. This was to have the expert panel rate a variety of samples. Each representative set contained one answer from each of its item's clusters, selected at random (see Section 5.4.3 and the breakdown of clusters by item in Figure 5.3). Each expert was randomly assigned 3 items from the set of 9, then shown one representative set of answers for each item they were assigned. We ensured that each item was seen by at least 5 experts, which is important for the purpose of computing the content validity index.

We asked each expert to (1) rate the relevance of each question statement to assessing a test taker's ability to create a visualization that serves that item's corresponding task, (2) score each visualization answer on a scale from 0 (poorly designed and unable to answer the question) to 100 (well designed and able to answer the question); this yields a holistic judgment of answer quality, which we later use to build our autograder, and (3) explain their grading criteria.

5.5.1 Content Validity of AVEC

The content validity index (CVI) is a metric that uses expert ratings to assess how well test items measure what they are intended to measure. The CVI is typically calculated from relevance ratings on a 4-point scale and is the proportion of experts who rated an item 3 (quite relevant) or 4 (highly relevant) [131]. Items with a CVI below 78% are candidates for

revision [131]. We use CVI in conjunction with IRT analysis in Section 5.6.6 to finalize the item bank.

In our expert panel, if an expert rated the relevance of an item below 3, they were asked to provide free-text comments to explain their reasoning. We used their comments to evaluate whether the revision for the item would be minor; if minor, the item would not require a second round of expert review [131]. Two items were candidates for revision:

- DD.3: “Total yearly spending did not exceed \$10,000 in 2020” (CVI: 0.33, task: *Describe Distributions*). Several experts mentioned that DD.3 only asks for one specific aggregated value, which would not need a distribution of values. Thus, we made a minor wording revision to better align the statement to its intended task: “Total yearly spending rarely exceeded \$10,000”. This way the question does not focus on one particular year.
- MPC.2: “The number of times that daily spending was around \$50 increased from 2020 to 2023” (CVI: 0.67, task: *Make Pattern Comparisons*). Experts’ comments did not suggest particular wording changes to the T/F statement, and 4 out of 6 experts rated it a 3 or 4, so we kept it as a candidate for revision and defer to the results from test tryout (Section 5.6.5) for the final verdict.

The remaining seven items have a CVI of at least 80%.

5.5.2 Qualitative Analysis of Grading Criteria

To better understand the experts’ scoring rules, and to validate (or if needed, refine) our initial clustering of answers, we asked the experts to provide free-text responses to explain their grading criteria for the visualization answers. We coded their responses and

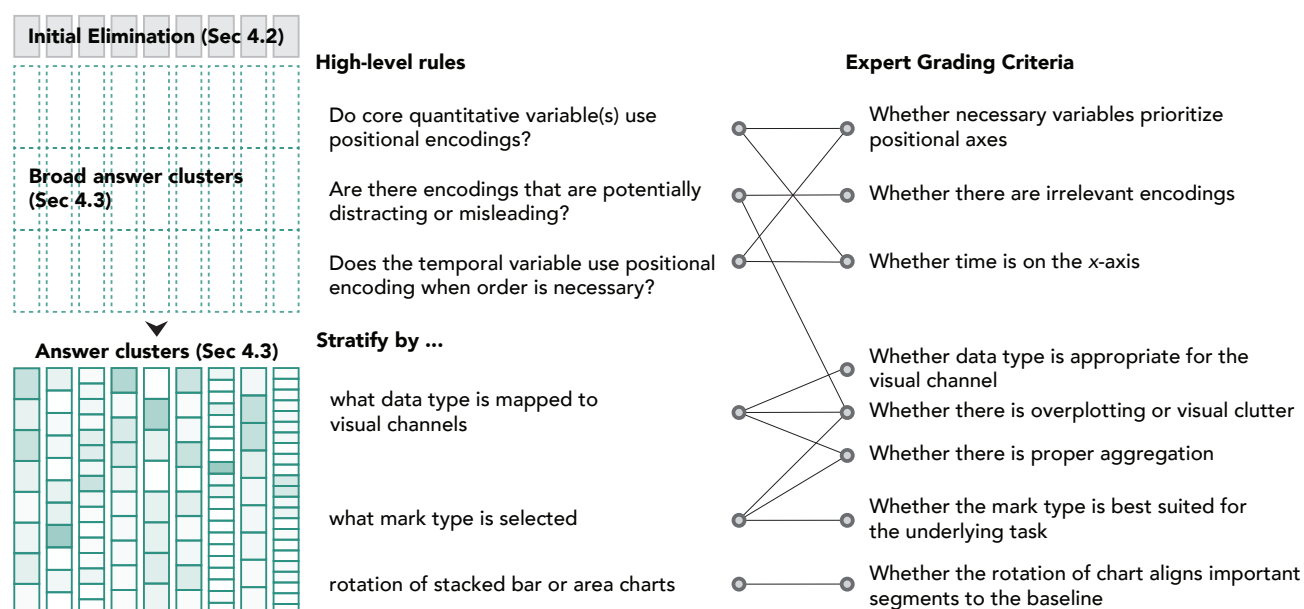


Figure 5.4: The mapping between our clustering rules and expert grading criteria. Each grading criterion could be mapped to one or more clustering rules we applied to cluster visualization answers of similar quality. This mapping suggests that our clusters capture the meaningful differences between answers that experts were using to form their holistic grades.

compared them against our clustering rules. If each expert criterion can be mapped to at least one of the clustering rules we used, then it suggests that our clusters were reasonably partitioned and captured the meaningful differences between visualization answers.

Coding Process

Two of the authors independently examined these comments and assigned codes to summarize each expert's grading criteria. The two coders then discussed their codes for each comment, resolved differences and ambiguities, and created a final consensus codebook and consensus labels for experts' grading criteria.⁹ Using the grading criteria codebook, we categorized the codes at the item level as item-specific rules, which we then use to assess the rules we applied for clustering visualization answers of similar quality. This ensures that our clusters are not missing some meaningful differences between answers

⁹The codebook is in supplemental materials: <https://osf.io/hg7kx>.

that experts are using to form their holistic grades.

Results

We found a total of eight recurring grading criteria in the experts' comments across all items, and each criterion can be matched to one or more rules that we first used to divide the visualization answers into clusters (see Figure 5.4). While some of these mappings are straightforward (e.g., *whether the rotation of chart aligns important segments to the baseline* is captured by the clustering rule *rotation of stacked bar or area charts*), a few depend on properties of our data or our tool. For example, in our datasets *visual clutter* most often occurred when a categorical variable is mapped to bar size, leading to overlapping bars of varying sizes, so the grading criterion *whether there is overplotting or visual clutter* is captured by the clustering rules *what data type is mapped to visual channels*, *what mark type is selected*, and *are there encodings that are potentially distracting or misleading*. As explained in Section 5.3.3, we apply aggregation based on the selection of mark types and data variable types, so the grading criterion *whether there is proper aggregation* is captured by the clustering rules *what data type is mapped to visual channels* and *what mark type is selected*. The expert grading criteria corroborated the rules we used for clustering answers of similar quality, which suggests that our clusters do capture the meaningful differences between answers. Thus, we proceed to use these clusters and expert scores as a basis for developing an autograder to grade visualization answers (Section 5.5.3) and further refine the autograder with participants' data from test tryout (Section 5.6.2).

5.5.3 Development of Initial Autograder

We develop an initial version of the autograder based on scores from the expert panel. We later refine and finalize the autograder using test tryout data in Section 5.6.

5.5.3.1 Holistic Grading

Existing approaches for grading constructed-response questions can be separated into two classes: *holistic* and *analytical* [111]. Although both approaches are based on guidelines (e.g. a rubric), the holistic approach is more top-down and aims to make a single judgement about the answer as a whole, while the analytic approach is more bottom-up, accumulating points based on specific features of the answer [111]. We used the holistic approach for grading the visualization answers, because it can better capture the *qualitative differences* between visualizations (e.g., two visualizations that both pass the same *analytical* rule of mapping necessary variables may still differ qualitatively depending on the chart type or the specific orientation of the chart).¹⁰ Thus, we asked experts to score answers on a single scale from 0 (poorly designed and unable to answer the question) to 100 (well designed and able to answer the question).

5.5.3.2 Grading Mechanism: Developing a Z-Score Lookup Table

To remove idiosyncrasies in experts' scoring (e.g., tendencies to use only one end of the scale), within each expert we standardized their scores into Z-scores using the mean and standard deviation of their scores. To compare the scores across items, we further standardized the expert-level Z-scores by transforming them into item-level Z-scores using the mean and standard deviation of the experts' Z-scores for each item.

We then examined each item-specific cluster of answers that experts scored. For each cluster, if the variance of the Z-scores for the answers was relatively low (meaning there are not much disagreements between experts' scores), then we assigned the mean

¹⁰We originally planned to grade each visualization answer holistically and analytically, but opted to just holistically because we reviewed the cluster-specific scores (Section 5.5.3.2) and found the holistic scores experts gave already took the important properties of the question, underlying task, and data into consideration.

Z-score as the final Z-score for all visualization answers in that cluster. If the variance was high (meaning there are some disagreements between experts' scores), we referred to the experts' text responses on their grading criteria as well as our own judgment to determine if the disagreement warrants (1) further dividing the cluster, (2) merging the cluster with another similar and relatively lower-variance cluster, or (3) keeping the cluster as-is and assigning the mean Z-score as the final Z-scores for the cluster. We also looked through the clusters ordered by their mean Z-scores to see if the relative order of scores reflects a reasonable ordering of the quality of the visualizations.

5.5.3.3 Results

After examining all 116 clusters across the 9 items, we identified and merged 2 pairs of clusters where one of the clusters within the pair had higher variance but the other cluster of the pair contained visualization answers of similar quality and had lower variance. We noticed that item MPC.2 had an especially high variance within several clusters that we believe reflected substantive disagreements between experts' scoring schemes, and we identified several instances of mean Z-scores for clusters on that item that would imply a relative ordering of answer quality that we disagreed with. We decided to keep MPC.2 provisionally for test tryout and re-examine this item after IRT analysis in Section 5.6.

Our initial autograder contains 114 clusters,¹¹ each with its own corresponding Z-score, which becomes the lookup table for automatically assigning scores to visualization answers. We further refine the autograder after collecting test tryout data, described next.

¹¹After we conducted item analysis using test tryout data, we decided not to include MPC.2 in the final bank. Therefore, there is a total of 103 clusters with their associated Z-scores in the lookup table after removing clusters corresponding to MPC.2.

5.6 Test Tryout

We tried out our 9 test items to refine the autograder and conduct IRT analysis. We then used the item analysis results along with evidence from Section 5.5 to finalize the item bank.

5.6.1 Participants and Procedure

We recruited 100 participants from a Prolific pool with people who are aged 18 to 65, are fluent in English, and have normal or corrected-to-normal vision. The study was expected to take 30 minutes, and participants were compensated 6 USD for successfully completing the study. We filtered out three participants who did not complete the study and two participants who failed all of the attention check questions. The 95 remaining participants consisted of 50 males and 45 females aged 18 to 54.

At the beginning of the survey, we presented the consent form to all participants, which described what to expect in the study. The study contained two sections: the training section and the main section.¹² The training section included five training questions which asked participants to reproduce charts using the assessment tool. Participants must correctly answer all training questions in order to proceed. Afterwards, participants were asked to fill out the System Usability Scale (SUS) [28]. We used the SUS to determine if the assessment tool is easy to use (**DC.2**), which helps ensure that differences in test-takers' scores are due to their ability and not the usability of the tool.

The main section included all 9 items from our item bank and 2 attention check questions. Item order was randomized. For each item, participants were asked to create

¹²The demo video of the assessment tool participants used and a live demo link are both documented here: <https://osf.io/hg7kx/>.

the best chart for determining whether a T/F statement is true. Then, they were asked to use the chart they created to select an answer for the T/F question and provide their reasoning. The participants were instructed that they could not go back once they proceeded to the next item. We designed the attention checks to ensure there was no room for misinterpretation if the participant had paid attention, following the format of Instructional Manipulation Checks (IMCs) [125]. Each attention check question explicitly instructed participants to select a specific mark type, drop-down value(s), and T/F answer.¹³

5.6.2 Autograder Refinement and Final Scores

We assigned initial grades to participants' answers using the Z-score lookup table from Section 5.5.3.3. That is, our autograder assigned each answer the Z-score from its corresponding cluster. If a visualization answer was eliminated during the initial elimination phase (Section 5.4.2), the autograder assigned it a score of -2, which is less than the minimum autograder score (-1.22) and is (by definition) two standard deviations below the mean, indicating a visualization of poor quality. Using -2 as the lowest autograder score also helps make the score distribution roughly symmetrical (the highest autograder score is 1.81).

Next, to refine the autograder and generate final grades for participants' answers, we examined each visualization answer that received a Z-score of -2 to identify any question-specific edge cases that should instead be merged into one of the existing clusters. As a result, we merged 8 types of edge cases into another cluster:

- One edge case was associated with FTC.1, which involved a rule that governed mapping quantitative variables to the color channel with mark *area*. This usually caused

¹³See supplemental materials for the exact attention check questions: <https://osf.io/hg7kx>.

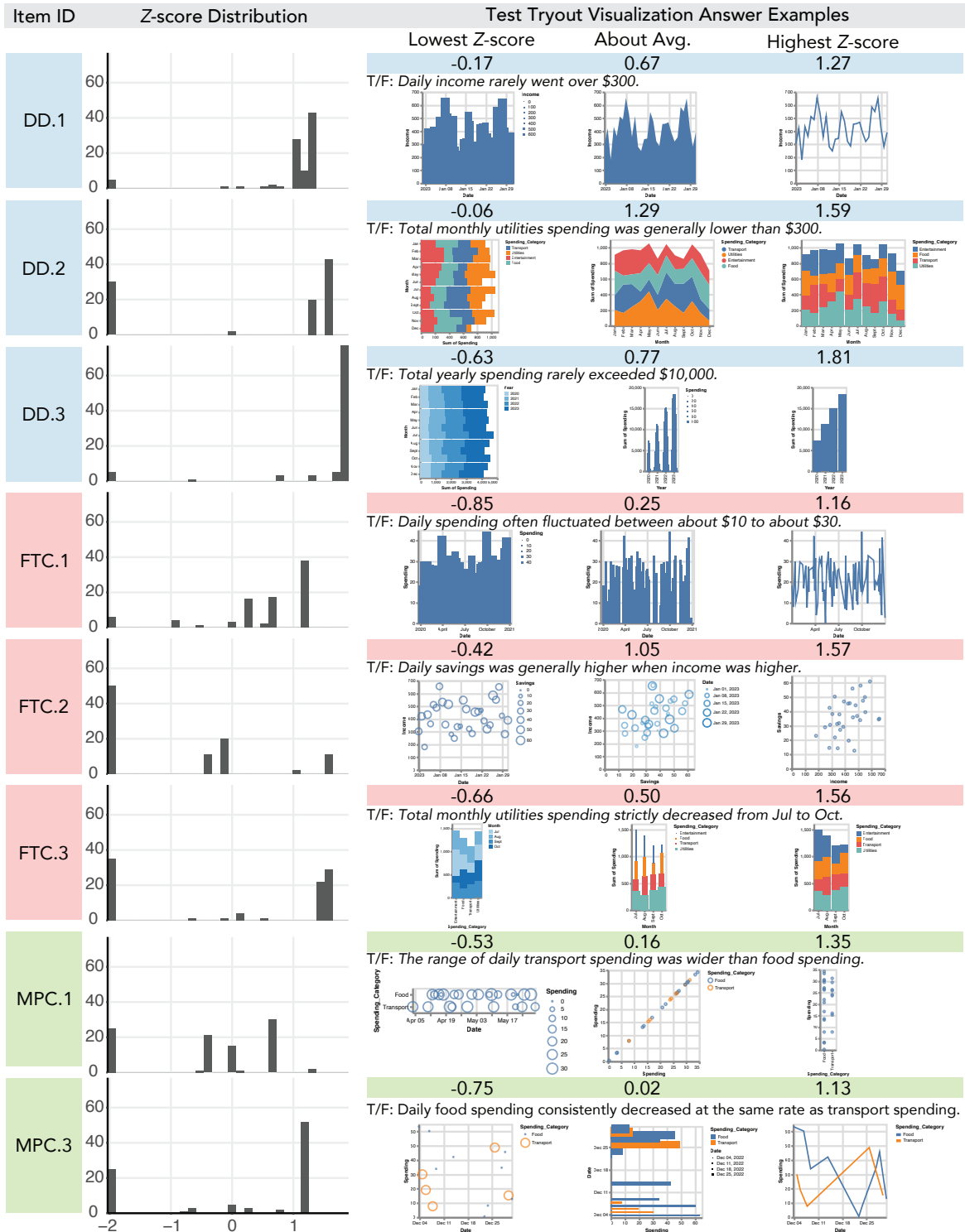


Figure 5.5: Distributions of Z-scores of participants' answers alongside examples of the lowest-scoring, about average, and highest-scoring visualization answers from test tryout for each item in the final bank. The lowest-scoring examples are the lowest test tryout Z-scores greater than -2.

the visualization specifications to be unrenderable on the interface due to having too many colors. FTC.1, however, used a smaller dataset that made it an exception to this rule.

- Two cases were associated with MPC.1. In one case, the necessity of data variables varied depending on mark type (“Date” became a necessary variable along with “Spending” and “Spending Category” when the mark type is *line*, because without “Date” the line would inappropriately connect the data points in each spending category). In the other case, a visualization answer had been eliminated based on the linter rule of *Use different fields for x-axis and y-axis*, which actually did not result in a poor visualization because the task for MPC.1 (making spending range comparisons) is still doable if “Spending” was mapped to both positional axes with mark *point*. This suggests that visualization linters must be used with caution and the relevance of the linter rules may vary depending on the context and question of interest. See supplemental materials¹⁴ for the visualization examples.

The other 5 out of 8 edge cases are associated with MPC.2.¹⁵ After the autograder refinement, we reran the autograder to assign final Z-score grades to the visualization answers from test tryout. Figure 5.5 shows each item in the final bank alongside distributions of Z-scores of participants’ answers and example visualization answers of varying quality from test tryout.

¹⁴<https://osf.io/hg7kx>

¹⁵Because we ultimately remove MPC.2 from the final item bank during test revision (Section 5.6.6), we defer details on its edge cases to supplemental materials: <https://osf.io/hg7kx>.

5.6.3 Bayesian IRT Model

Model Specifications

The 2-parameter IRT model has been used by visualization literacy test developers to infer item easiness (i.e., how easy it is to correctly answer the item) and item discrimination (i.e., how well the item can differentiate people of different abilities) of test items [72, 55]. The traditional 2-parameter IRT model uses logistic regression, because it assumes that an answer is graded in a binary way: correct or incorrect. Since our items are graded on a continuous scale, we modified the traditional IRT model to instead use linear regression by changing the response distribution to a Gaussian. We implemented a Bayesian IRT model with the brms R package [29].

The outputs of this model are posterior distributions of item easiness, item standard deviation, and the ability of each test-taker. In the context of this model, item discrimination can be understood as the inverse of the item standard deviation: when the standard deviation of an item is high, then it cannot easily differentiate people with different abilities, because there is a lot of variance in people's scores on that item. Thus, the higher the standard deviation, the lower the discriminating power of an item. We set the priors for the intercept to $\mathcal{N}(0, 1)$, standard deviations of ability, easiness, and responses to $\mathcal{N}^+(0, 1)$, and the correlation matrix to LKJ(2).¹⁶

Sample Size

To determine the sample size for test tryout, we conducted a pilot study ($N = 10$) to fit our modified IRT model, and used the model to simulate various sample sizes. We found a sample size of 100 was sufficient for model convergence, had reasonable posterior pre-

¹⁶This model was preregistered on OSF: <https://osf.io/hg7kx/>.

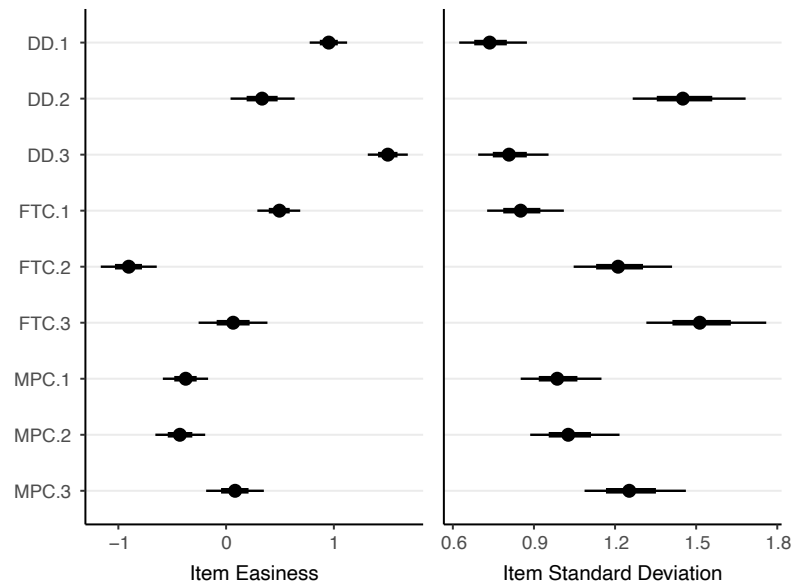


Figure 5.6: The item easiness estimates and item standard deviation estimates for each item. The dots in the item easiness and standard deviation coefficient plots represent the median estimates.

dictive checks, and yielded intervals for the item easiness and standard deviation parameters from the model output that are of similar width to prior work that has also applied IRT [72].¹⁷

5.6.4 Descriptive Statistics and SUS Score

The total completion time in minutes ranged from 8.50 to 64.76 ($M = 21.70$, $SD = 12.61$), suggesting that 30 minutes are reasonable to complete the test. The item-level correctness of the T/F questions ranged from 0.67 to 0.96 ($M = 0.82$, $SD = 0.10$). The participant-level correctness of the T/F questions ranged from 0.44 to 1 ($M = 0.82$, $SD = 0.15$).

We computed the SUS scores for each participant following the established procedure [28]. The average SUS score of participants is 77.76 ($SD = 19.80$), which is considered to be in the acceptable range [12]. This provides some evidence that we were able to create

¹⁷See OSF preregistration: <https://osf.io/hg7kx/>.

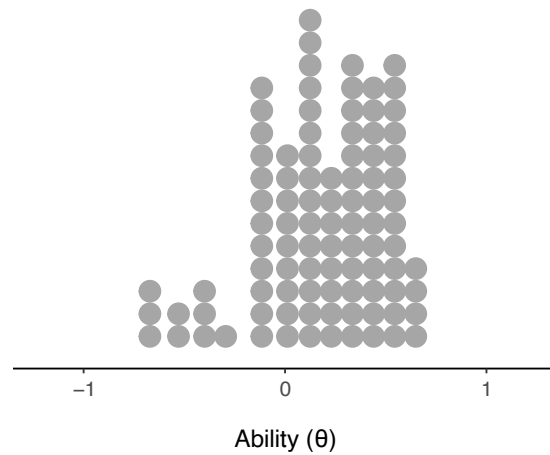


Figure 5.7: The distribution of the estimated abilities of participants from test tryout.

an interface that is easy to use (**DC.2**), such that unfamiliarity with the interface should be less likely to interfere with measurement of participants' visualization encoding abilities.

5.6.5 IRT Analysis Results

We ran our IRT model on the test tryout data with final grades of participants' answers for the items. The minimum bulk effective sample size is 6,530 and the minimum tail effective sample size is 6,796, and all $\hat{R}$ values are approximately 1.

Figure 5.6 shows the coefficient plots displaying both item easiness and standard deviation (higher values correspond to lower discriminating power), each with 95% and 66% credible intervals (CI) and a dot indicating the median estimate. The median estimates of the item easiness parameter ranged from -0.90 to 1.50, with an average of 0.19. The median estimates of the item standard deviation parameter ranged from 0.74 to 1.51, with an average of 1.09. Figure 5.7 shows the distribution of the estimated abilities of participants.

5.6.6 Final Item Bank

We made a holistic judgment when finalizing the item bank by considering many factors [44], including the content validity of an item, how difficult an item is, its standard deviation compared to other items in the bank, and preserving the diversity of the bank. We re-evaluated the two items identified as revision candidates ($CVI < 78\%$) in Section 5.5.1:

- DD.3: we retained the revised item because it has a median easiness of around 1.50, which contributes to the coverage of a wide range of abilities.
- MPC.2: we removed MPC.2 because it had a low CVI of 0.67, high variance of expert scores within answer clusters, and there is one item in the same task category (MPC.1) that has similar easiness and standard deviation, so removing MPC.2 would not necessarily affect the diversity of the item bank.

The remaining 7 items were retained because they possess at least one of the following qualities: (1) high CVI, (2) low item standard deviation, or (3) they contribute to the diversity of the item bank. We also calculated McDonald's Omega [54, 62, 135], a reliability measure widely used to assess the internal consistency of a test, and found the overall reliability ω_t to be 0.73, which is considered to be acceptable [123]. As a result, our final bank consists of 8 items, which can serve as a valid assessment of the visual encoding ability in visualization construction, and the items have varying easiness and can differentiate participants of different levels of ability.

5.7 Discussion

5.7.1 Usage Recommendations

Based on our test tryout data, the mean completion time was around 21 minutes (Section 5.6.4). If test administrators prefer a shorter test, they could pick from the item bank based on their desired item easiness and standard deviation estimates. For instance, if the target test population is people with lower abilities, administrators could select a set of items with low item standard deviation at lower-ability ranges. Or if they would like to focus on a specific visualization task, such as *Describe Distributions*, then they could include only the items pertaining to that task. Items with low standard deviation (corresponding to high discriminating power in traditional IRT) should generally be preferred if one were to select a customized subset, as with prior work on assessment development using IRT [72].

Our procedure of developing AVEC could also be extended for test administrators interested in having a larger item bank. The design space we used to generate the T/F statements contains redundancies that we removed during item generation, such as T/F statements that only differ in the quantitative variable of interest (Section 5.3.2). One could start with the validated items in AVEC then identify other items that might have similar answer clusters as a way to expand to a larger bank. If answers for the new item fit into existing clusters, the autograder can automatically grade the answer. Otherwise, test administrators should reassess the set of visualization answers for the new item(s) and devise a scoring scheme most appropriate for the context of use.

5.7.2 Applications of AVEC in Future Research

One obvious area of application for AVEC is facilitating the study of the abilities involved in visualization literacy, such as the interplay between visualization interpretation and construction. For instance, AVEC could be used to investigate whether people who can appropriately map data to visual channels are also better at interpreting visualizations, and vice versa. Additionally, future work can study how much the ability to reason about potentially misleading visualizations [72] might contribute to construction-related abilities such as the visual encoding ability, which can further inform visualization teaching curricula and the design of interventions for improving a person's overall visualization literacy.

Beyond assessing a person's visual encoding ability, AVEC can also be used to evaluate the effectiveness of interventions designed to improve this ability, such as different curricula in visualization classes, through pre-post testing. These studies might help teachers identify areas that need special attention during instruction. Future work could also explore the potential for AVEC as an educational intervention itself and investigate whether providing automated and personalized feedback to users through the web-based assessment tool could improve their visual encoding ability.

5.7.3 Alternative Scoring Strategies

Unlike multiple-choice items, answer correctness on a constructed-response item is often not binary, and the degree of answer correctness may vary depending on the rubric or scoring scheme. We used item-specific scores from an expert panel to create a lookup table of Z-scores. The lookup table offers a flexible way of scoring the possible visualization answers in AVEC and could potentially be adjusted based on test administrators'

needs. For instance, test administrators could reduce the granularity of the grades by merging clusters and their mean Z -scores, or could transform the Z -scores into simpler rankings (e.g., whole number scores from 0 to 4).

We used a holistic grading approach, as explained in Section 5.5.3.1, but another approach, such as analytical grading, may lead to alternate rubrics and interpretation of the scores. Alternative ways of grading visualization answers might start with the set of expert grading criteria we summarized from the expert panel and assign points based on the number of criteria a visualization answer satisfied. Or one could take a mixed approach to grading that involves both holistically evaluating the visualization answer then applying analytical points, or vice versa. Our answer clusters and the Z -score lookup table can be easily adapted to support such alternative scoring schemes. For example, a test administrator who adopts a partially automated scoring system might use the Z -score lookup table to first filter out visualization answers that did not pass a predetermined threshold (e.g., filtering out answers that are below average), then apply analytical rules to grade the remaining visualization answers, which should greatly reduce the time for grading. Alternatively, future work could further refine the autograder using the data we collected from the expert panel. One could adapt existing methods of ranking visualizations (e.g., Draco [122, 166], MultiVision [164]) to incorporate our item-specific expert knowledge, e.g. by tailoring the knowledge base and fine-tuning the constraint weights [166] or learning weights for soft constraints from ranked pairs of visualizations (as in Draco-Learn [122]). If test administrators alter grading schemes, they should re-run the IRT analysis with the new grades, because the grades of the visualization answers would affect the item parameter estimates.

5.7.4 Limitations and Future Work

Individual Differences in Visualization Skills

In the test tryout study, we collected participants' basic demographics (Section 5.6.1) but did not collect their visualization expertise levels or familiarity with different chart types. We therefore did not explore individual differences in visualization skills or the relationships between participants' prior expertise levels and their visual encoding ability. Similar to prior work on visualization literacy assessments (e.g., Boy et al. [26], VLAT [106], CALVI [72]), this limitation does not necessarily affect AVEC's validity, which we demonstrated through having (1) items with high CVI ratings from experts (Section 5.5.1) and (2) an assessment tool that preserves the combinatorial nature of selecting appropriate visual encodings while still being easy to use (Section 5.6.4). Exploring the relationships between different visualization skills (e.g., reading, constructing) would be interesting directions for future research.

Visualization Construction Beyond Visual Mappings

Visualization is a complex, multi-step process. Although AVEC focuses on the core of the visualization reference model (visual mappings) [34] and includes two types of aggregations (count and sum), there are many other facets of visualization construction that are important to study and measure. For instance, effectively visualizing data also requires creators to correctly handle raw data and perform appropriate data transformations (e.g., avoiding cherry picking [72, 109, 75, 113], normalizing data correctly [75, 121, 113]) for the resulting visualization to convey a clear—and correct—message. Creating effective, engaging visual narratives often requires creativity and extensive iteration (e.g., [150, 19, 146]). Expert designers may make use of interactivity or animation (e.g., [138]) to pull

readers in or better convey their intended message; these techniques require time and skill to master. Future work should expand our efforts in assessing construction-related abilities and deepen the study of the complex skills related to visualization construction and beyond.

Chapter 6

An Autoethnography on Visualization Literacy: A Wicked Measurement Problem

This chapter is based on work published in IEEE VIS 2025 [73] and is done in collaboration with Anne-Flore Cabouat, Karen Bonilla, Yuan Cui, Yiren Ding, Noëlle Rakotondravony, Mackenzie Michael Creamer, Jasmine Tan Otto, Maryam Hedayati, Bum Chul Kwon, Angela Locoro, Lane Harrison, Petra Isenberg, Michael Correll, and Matthew Kay.¹

6.1 Introduction

A growing number of empirical studies in the data visualization community now rely on measures of *visualization literacy* (i.e., the ability to interpret and construct visualizations) as a basis for designing more effective visualizations (e.g., [70]) and identifying a baseline for skill improvement in participants (e.g., [91]). Prior work has also developed targeted interventions to improve visualization literacy (e.g., [5, 100, 20]), which in turn requires valid measurements for effective evaluation. To meet the demands of such measurement-reliant studies, researchers are thus incentivized to develop—and have already developed—various quantitative assessments of this construct (e.g., [26, 106, 23, 72, 2, 71]).

However, the complex, fuzzy, and multifaceted nature of visualization literacy makes it hard to pinpoint precisely what we are measuring, or indeed *want* to measure. Numerous definitions and frameworks for conceptualizing visualization literacy have been proposed [26, 23, 24, 106, 72, 147, 85]. Different conceptualizations have led to different measurement scales tailored to different visualization literacy components [26, 106, 72]. Initiatives such as meetups, panels, and workshops have tried to distill what it means to be literate in visualizations [94, 95, 74, 132]. Despite these longstanding efforts, debates around what visualization literacy even *is*—let alone how to define it—persist: we still do not comprehensively understand *visualization literacy* as a construct.

Even as we are unable to precisely define it, we must forge ahead, making the compromises necessary to study visualization literacy as both test users and test designers. We view visualization literacy measurement as a *wicked problem*, the kind of problems defined in social planning and public policy as “complex, intractable, open-ended, unpredictable” [4], and which “do not have an enumerable (or an exhaustively describable) set of potential solutions” [137]. The tension between not fully understanding visualization literacy as a construct while still needing to measure it is what makes the study of it so challenging.

We¹ gathered at the ACM CHI 2024 workshop on visualization literacy [74] to begin to tackle this wicked problem. We engaged in critical debates about how to better define, study, and improve visualization literacy. Questions central to our discussions included *what are the assumptions underlying visualization literacy?* and *what worked / didn’t work in measurement?* Conversations continued after the workshop, where we reflected in depth on our professional and practical experiences. These reflections exposed our tacit

¹A subset of the authors were organizers of the workshop.

thoughts on visualization literacy—thoughts which often exist only in between polished works and are left unreported in individual research papers.

To bring these tacit thoughts together, we contribute an autoethnography detailing the *fluidity* inherent to visualization literacy—fluidity that makes the construct an ever-moving target across domains, cultures, and contexts of application. As test users and test developers with extensive experience studying visualization literacy, we documented our opinions and reflections in a structured way through internal interviews. Based on a thematic and diffractive analysis of our interview data (Section 6.3), we discuss how the complexity and fluidity of visualization literacy leads to misalignment between what the construct is and how measurements of it are used or designed. From conceptual (Section 6.4), operational (Section 6.5), and methodological (Section 6.6) perspectives, we discuss challenges that threaten measurement validity. From these challenges, we provide several calls to action (Section 6.4.4, Section 6.5.4, Section 6.6.3) that pertain to each of the three perspectives, such as expanding test scopes, making it easier to use tests, broadening test modalities, improving test ecological validity, and taking an iterative approach to test design incorporating interdisciplinary expertise. We hope that by embracing the complexity and fluidity of the construct, the community can better navigate the wicked problem of visualization literacy measurement.

6.2 Related Work

We outline some of the existing visualization literacy measurements and describe collaborative autoethnography and diffractive analysis as methods to surface tacit thoughts.

6.2.1 Visualization Literacy Measurements

Visualization literacy is not a straightforward construct to measure. The various definitions (e.g., [9, 23, 24]) of it proposed through prior papers, past meetups, workshops, and panels suggest the inherent complexity of the construct (e.g., [94, 95, 74, 132]). Researchers have made many attempts at developing frameworks to conceptualize visualization literacy, such as mapping out the landscape of the construct [147] or reconceptualizing it with visualization competencies, comprehension processes, and practices [85]. There are also individual definitions of visualization literacy focusing on different components of the construct and each leading to corresponding assessment(s) (e.g., [26, 106, 72, 71]). Other works improved the time efficiency of these tests [55, 130]. Other assessments have been designed specifically to evaluate particular interventions [2] or to assess literacy on specific chart types, such as treemaps [66] and parallel coordinates plots [67]. These measurements often exist somewhat in isolation from each other and from efforts in other fields [101, 124, 70].

The theory of *multiliteracies* [36, 46] sheds some light on the complexity and diversity in visualization literacy measurement. That theory describes how literacies can be both multimodal (involving media beyond text [46]) and multicontextual (involving “different cultural, social or domain-specific contexts” [46]). While visualization literacy is clearly multimodal, its multicontextual variability may be what makes defining and measuring it exceptionally hard, leading to the diversity of existing attempts described above.² It is in the face of this wicked measurement problem that we looked for a way to both understand the current state of the problem and search for concrete next steps.

²We return to multimodality and multicontextuality in Section 6.4.1 and Section 6.4.2.

6.2.2 Collaborative Autoethnography

As experts in the area, autoethnography offers a way to document our reflections and tacit thoughts about the state (and future) of visualization literacy—thoughts that are often unreported in individual papers and invisible to literature reviews. Autoethnography is a qualitative research method that focuses on the personal experiences of the authors (*auto*) by describing and systematically analyzing (*graphy*) these reflections for the purpose of understanding cultural experiences (*ethno*) [63, 93]. This method “acknowledges and accommodates subjectivity, emotionality, and the researcher’s influence on research, rather than hiding from these matters or assuming they don’t exist” [64]. The main source of reliability in autoethnographical work comes from the person’s credibility (e.g., could they have had the experiences described?) [64], and a valid autoethnography evokes a feeling that the described experience is “lifelike, believable, and possible” [64].

Prior work in the visualization community has used autoethnography to surface details of complex design processes [112, 57]. However, a weakness of individual autoethnography is that the researcher may be too close to the experience to see it in a nuanced way [103, 104]. *Collaborative autoethnography* builds on top of traditional autoethnography to address this weakness. It involves two or more researchers “pooling their stories to find some commonalities and differences and then wrestling with these stories to discover the meanings of the stories in relation to their sociocultural contexts” [38], and has the potential to create a more rigorous, multi-perspective analysis [103, 77]. Thus we adopt a collaborative autoethnographic approach to gather and analyze our tacit thoughts about visualization literacy.

6.2.3 Diffractive Analysis

To help us find differences in our pooled stories, we also employed *diffractive analysis*. Diffractive analysis has parallels in feminist and interpretivist approaches that value the differences in researchers' perspectives and view them as important factors in shaping the research analysis [105, 13, 14]. Prior work [3, 105] in HCI and visualization has applied variations of this approach to incorporate analysts' differing opinions into a qualitative analysis. Because one driving motivation for our collaborative autoethnography is to distill the various in-depth experiences that we have accumulated about visualization literacy and its assessment, we sought to avoid passive agreement by employing diffractive analysis to increase the possibility of differences in opinion and encourage constructive disagreements during our analysis.

6.3 Methods

We lean on our expertise and experiences to seek answers to these motivating questions: *what makes visualization literacy so hard to measure?* and *are existing tests being used as they were intended?*

6.3.1 Expert Backgrounds

The participants of this autoethnography are researchers in information visualization who are either pursuing a doctorate ($N = 8$) or have obtained a doctorate ($N = 7$). We started regular conversations and meetings after the CHI 2024 workshop on visualization literacy [74] that a subset of us organized. While we have all used or considered using assessments of visualization literacy, 9 of us have had experience in developing assessments [106, 72, 31, 84, 71, 55, 56, 115] (Table 6.1).

Table 6.1: The authors’ initials in alphabetical order along with self-reported (at the time of the interviews) years of experience in visualization literacy, visualization research and professional experience, assessment design experience, and other prior work related to visualization literacy.

Author	VisLit (yrs)	Vis (yrs)	Designed Assessment(s)	Other VisLit Work
AFC	1.5	10	[31]	[30]
AL	5	8	[115]	[17, 16]
BCK	9	17	[106]	[94, 74, 107, 100]
JO	1.5	10		[127, 126]
KB	2	2		[58, 74]
LH	9	17	[55, 56]	[80, 117, 20, 58, 74]
LG	3.5	3.5	[72, 55, 56, 71]	[74]
MC	7	15		[48, 47, 74]
MH	4.5	4.5		[86, 85, 74]
MK	7	12	[72, 55, 56, 71]	[58, 86, 85, 74]
MMC	2.5	2.5		[53]
NR	2.5	4		[8, 20]
PI	1.5	20	[31, 84]	[30]
YC	3	3	[72, 55, 56, 71]	[74]
YD	5	5	[55, 56]	[20, 58, 74]

6.3.2 Internal Interviews

To collect the autoethnographic data in a systematic way, we conducted internal semi-structured interviews with 14 authors.³ Four of the authors served as interviewers, and each interview was conducted by a pair of interviewers. We coordinated the interview sessions so that the interviewers and interviewees do not have supervisor-supervisee relationships and are not regular co-authors. We asked each interviewee to reflect on how visualization literacy assessments were integrated into their research, their past experiences on the topic, and thoughts on measurements of visualization literacy more generally. The questions we asked during the semi-structured interviews included *what tests do you use/not use and why?*, *what are the issues you’ve experienced using (e.g., implementing, administering, analyzing) the tests?*, *what do you think visualization literacy even is?*, and

³One author was not interviewed due to time constraints, but they contributed their tacit thoughts during group discussions, analysis, and paper writing.

*what is your wish list for what you'd like to see covered in tests?*⁴ Asking interviewees to elaborate on the reasoning behind their decisions (e.g., why and how they decided to use certain tests) and to reflect on the challenges they might have encountered along the way allowed us to tap into valuable tacit thoughts on design and use of tests. Each interview session lasted about 45 minutes. We used transcribed audio recordings for analysis.

6.3.3 Thematic Analysis and Diffraction

We used thematic analysis [43] to analyze our interview transcripts and supplemented it with diffractive analysis [14]. We employed a series of analysis steps to also embrace differences in opinion and enrich our conversations instead of relying on methods that prioritize agreement among data collectors (e.g., inter-rater reliability [120]). Disagreements among us can offer diverse perspectives to help us form a more holistic view of the many factors that make measuring visualization literacy a wicked problem.

Independent Coding In total, 9 of the authors served as analysts and followed the same analysis procedure for the 14 interview transcripts (Figure 6.1). Each interview transcript received 2 analysts. We started with independent coding to allow the analysts to arrive at their own set of interpretations and codes. To further encourage differences in opinion, we aimed to distance the analysts from the transcripts and made assignments while ensuring the following criteria were met: (1) each analyst did not analyze their own interview transcript, (2) each transcript was assigned at least one analyst that was not one of the interviewers that conducted the interview, (3) each pair of analysts was not from the same research lab, and (4) each analyst did not analyze a transcript from someone from their research lab. Each analyst independently coded emerging themes from their assigned transcripts.

⁴See supplemental materials for the interview protocol: <https://osf.io/xwr4c/>.

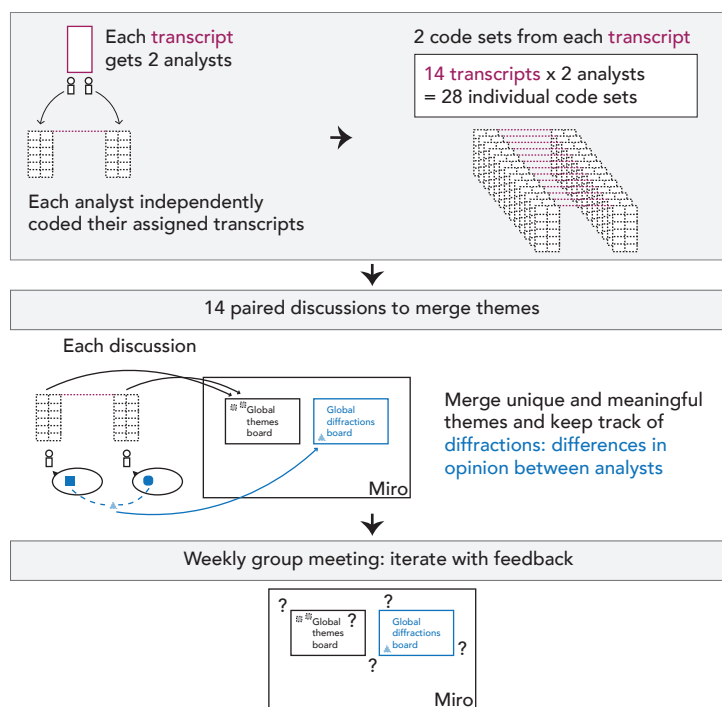


Figure 6.1: A flow chart demonstrating our analysis procedure, which involved each analyst individually coding their assigned transcript, pairs of analysts merging their codes and themes on Miro while keeping track of diffractions, and the large group discussing the themes and diffractions.

Paired Discussions Once both analysts from one transcript finished their independent coding, they met to discuss their codes. During their discussion, they merged their independent codes into theme sets while keeping track of differences in opinion between them (i.e., diffractions). This process allowed analysts to reflect on their interpretations, compare and contrast their understandings, and take note of any diffractions that emerged. The analysts contributed to two Miro boards during their discussions.⁵ One Miro board documented the emerging themes and sub-themes (the global themes board), and another Miro board was dedicated to any diffractions that surfaced during the analyst discussions (the global diffractions board). All of the analysts edited and / or added to the same two Miro boards, meaning that the themes from different analyst pairs were merged as different paired discussions happened. Two types of diffractions emerged: (1) dis-

⁵See supplemental materials for the Miro boards: <https://osf.io/xwr4c/>.

agreements within a pair of analysts on interpretation of particular quotes, and (2) disagreements with another pair of analysts on their code(s) or theme(s).

Large Group Meetings We held weekly group meetings that included authors who were not analysts to discuss the global themes board and global diffractions board. This allowed us to incorporate non-analysts' perspectives and to surface differing opinions and refine themes. Discussions of diffractions led to (1) resolving disagreements on interpretation of particular quotes by asking the interviewee of the quote to clarify and (2) large-group discussions about any new themes and how to better structure existing themes. The diffractions served as the main discussion points during these meetings, guiding the analysis.

Themes emerging from our conversations revolved around conceptual (Section 6.4), operational (Section 6.5), and methodological (Section 6.6) challenges. We detail these challenges and how they may lead to misalignment between what the construct of visualization literacy is and how measures of it are used and designed.

6.4 Conceptual: What Exactly are We Measuring?

A challenge we raised repeatedly throughout our interviews was the difficulty of pinpointing exactly what visualization literacy is. Having a broad definition, such as “the ability to interpret and construct visualizations”, may be helpful in communicating a general idea about the topic, but it is less useful when the goal is to operationalize it and design a precise measure of it. Here, we describe the factors that contribute to the complexity and fluidity of the construct, thereby making it difficult to define and conceptualize. We also discuss the term *visualization literacy* itself and our diverging opinions on its usefulness.

6.4.1 Sources of Definitional Complexity

Visualization literacy is a complex construct, yet definitions of it often seem overly simplified or do not fully capture what we think might constitute being “literate” in visualizations. This feeling of inadequacy is partly due to the many implicit and often hard-to-describe skills that contribute to making someone proficient in visualization. **Visualization abilities draw on knowledge from a variety of disciplines**, such as mathematics, data science, statistics, computer science, geography, design, art, cognitive science (to name a few), as many of us pointed out when asked *what do you think visualization literacy even is?* This complexity echoed the sentiment towards visualization from prior literature [9] and multimodality within the theory of multiliteracies [36, 46]. As BCK stated:

Visualization literacy [...] consists of multiple skills, multiple skills in different levels, in different depths. 'cause I mean, math I think is involved in it. Some sort of sense-making in the actual literacy of understanding human language is also part of it. So everything is intertwined—graphicacy, understanding shapes, geometries, altogether. [...] To be able to use the visualization to solve problems, I think you need to understand the construction skills, understanding the mapping between data and visualizations. (BCK)

Indeed, cognitive skills, such as spatial reasoning and verbal processing, along with sensory abilities like color vision, also interact and play a crucial role in applying visualization skills [110], making the construct even more difficult to isolate. But we should embrace its interconnectedness:

The community probably should go [toward] understanding where visualization literacy fits along with other cognitive characteristics, traits, skills, and how much they are influencing each other. (BCK)

In addition to skills from related domains, there are also knowledge and skills specific to data visualization, such as visualization grammar, graphical conventions, and visual metaphors. Research efforts to categorize visualization skills in varying granularity have led to a vast collection of visualization task taxonomies (e.g., [7, 27]). However, **we were broadly dissatisfied with tasks typically used to define visualization literacy**, such as low-level tasks like value retrieval or data comparisons:

Things like retrieve value is one of the common tasks in these visualization literacy assessments. But [...] if I want to retrieve a value, I go to a table. (YC)

Like a Cleveland and McGill [task]: estimate the ratio between two values in a chart. Like, that's just not a thing that people seem to need to do. (MC)

While there is value in understanding whether people can read data from a given chart, there are other higher-level (and perhaps fuzzier) visualization skills that are just as important (if not more) to assess:

A more complete idea of what visualization literacy should be [...] more on the side of, what do we actually use visualization for? And not just can you read a bar chart. (MK)

We need to improve the ecological validity of visualization literacy assessments by focusing on skills that are representative of real-world use cases, such as decision-making, the ability to infer new knowledge from a visualization, or the creation of effective⁶ visualizations. Although there is nascent work assessing such skills [2, 71], there is still an inadequate representation of these higher-level tasks in most widely-used tests. If this imbalance persists, it will limit our ability to reach a more comprehensive understanding of

⁶Though what “effective” means also varies across contexts (Section 6.4.2), making these higher-level skills even fuzzier.

visualization literacy.

The complexity of the different skills and knowledge that make up this construct makes it difficult to precisely define and conceptualize. **It is difficult to measure what we cannot fully describe.** One source of this definitional difficulty may come from differences in how people learn about visualization, which may be through exposure (e.g., implicit internalization of visualization rules by viewing real-world examples) or through more formal settings like classrooms. Thus, different people may have different mental models of how visualizations work. By analogy to language, MK explained:

I think most people, when they're speaking their first language, are probably not as good at the technical grammar of that language as people who are second language speakers sometimes. Because as a second language speaker, you often learn those specific rules of grammar, right? And as a first language speaker, you internalize them, but unless you actually sit down and learn them, you often can't articulate them. (MK)

Two people who are equally able to read, write, and speak a given language may differ in their explicit knowledge of lower-level skills, such as grammar. Similarly, the specific skills underlying a proficiency in data visualization and the explicitness of those skills might differ from one person to another, depending on how they learned. These differences contribute to the difficulty of identifying a precise set of skills that can define what it means for someone to be literate in visualizations.

6.4.2 It Changes Across Domains, Cultures, and Time

Not only is visualization literacy complex, it is also quite context-dependent, a property we refer to as its *fluidity*.⁷ For instance, many examples from our conversations highlighted that **visualization literacy is domain-specific**. The skills required to understand visualizations in a medical or healthcare context may be vastly different than what is necessary in business intelligence. Early work in visualization literacy assessment development also pointed out that domain knowledge could play a role when applying visualization skills [26]. This raises the questions of if it makes sense to attempt to define *visualization literacy* as an immutable construct across domains. Perhaps this domain-dependent characteristic resembles language *dialects*:

You have academic visualization literacy or [...] business intelligence visualization literacy or whatever. I think that those are different dialects. (MK)

If someone is “literate” in one specific visualization domain (e.g., business intelligence), that does not imply that they are immediately “literate” in another (e.g., data journalism). JO recalled:

I’ve heard these fun anecdotes about [...] “oh, when I make charts for these business people, I have to use this particular kind of chart. If I present the same information in a different chart type, they don’t care, right? It doesn’t look right.” (JO)

We considered if perhaps, amid different visualization dialects, it would be useful to look for a *minimum standard* for what constitutes someone being “literate” in visualizations, or what a minimum standard might mean in particular contexts. There were many diverg-

⁷cf. the multi-contextual dimension of the theory of multiliteracies [36, 46].

ing opinions: a minimum standard might include basic skills to get through life (e.g., map navigation skills), the skills needed to ensure the quality of information consumed (e.g., critical thinking), or even an understanding of fundamental syntactic elements of visualizations (e.g., axes and scales). While we could not form consensus, these might serve as starting points for identifying what we need to measure within different contexts.

The fluidity of visualization literacy does not stop at the domain-specific level. As alluded to in Section 6.4.1, how people develop visualization skills may differ due to their cultural background and experiences: **visualization literacy is also culturally-dependent**. This dependency leads to differences in what people consider to be important in a visualization or what they use visualizations for. A simple visualization reading task implicitly requires culture-dependent knowledge such as its language, local maps, or region-specific political information. As MC put it, visualization literacy “*is also this very contingent and culturally-bound thing.*” By analogy to language literacy:

You can be very literate in English without having read a Shakespeare play[...So]it’s gonna be hard to make a universalizable principle[about visualization literacy]because so much is bound up with sort of the visual and graphical cultures that are associated with the types of visualizations that people will need to interpret. (MC)

Any attempt to define visualization literacy as a static construct that is absolute across cultural contexts would be an oversimplification, potentially compromising future assessments and interventions.

The fluidity of visualization literacy is amplified if we step back and consider it on a larger time scale. A few of us (BCK, LH, MC, MK) who have been studying visualization literacy for a longer period of time reflected on how attitudes in the community have changed

over the years, suggesting that **visualization literacy also evolves over time**. Since the early conversations about the construct, the conceptualization of visualization skills has shifted from a definite set of skills, stemming from precisely-delimited visualization taxonomies, to a more complex understanding of visualization literacy involving skills and cognitive abilities from other disciplines, such as numeracy and spatial reasoning. BCK reflected on years of experience working in visualization literacy:

My understanding of visualization literacy has evolved. After a series of follow-up studies, conversations, and studies conducted by other people, and also by just pure observations in the field [...] After 10 years, I think the visualization literacy [...] [skills are] all combined in interesting ways [...] using [visualizations] to solve different problems require different sets of skills. And it's also highly contextual, I think. So in certain cases, for a certain context, you may have a higher chance to solve the problem with a given chart. Maybe in other context you don't necessarily. (BCK)

Moreover, some of us anticipate technological advancements changing the way we interact with or create visualizations, which could affect the conceptualization of the construct and subsequent assessments. KB wondered, “**how do you even disentangle technical skill from actual visualization skill?**” While technical skills play a part in creating visualizations, several of us do not see technical skills as a necessary part of visualization literacy for a general public, as MK expressed:

If you couldn't make [visualizations] in like ggplot or matplotlib or D3, that's fine, because most people don't have to do that. (MK)

In that respect, perhaps technological advancements would not have a large impact on what constitutes visualization literacy. YD also felt visualization literacy will not become obso-

lete:

Right now the AI probably can summarize the information from the charts. [...But] I think people with good data visualization literacy can definitely retrieve the information they want from the charts much faster than [by asking] the AI. (YD)

As researchers in visualization, it is in our best interest to be attentive to how visualization literacy differs across different contexts (e.g., domain, culture, and time) and to stay open to the possibility of change.

6.4.3 Differing Opinions on “Visualization Literacy”

The term “visualization literacy” itself is often the source of first impressions about the construct. This is especially important when we are communicating within our own research community but also when disseminating research output to other fields. We had diverging opinions on the utility of the term “visualization literacy”: while some of us saw it as a useful metaphor, others advocated for finding other terms.

“Visualization literacy” draws an explicit parallel to language literacy, which has some advantages from an empirical standpoint, such as how we might conceptualize the construct in a way that is useful for measurement:

Maybe we should be borrowing more from the literacy metaphor in terms of thinking [of] it as this longitudinal lifelong thing with multiple levels that you’re doing as opposed to a psych assay with some low to high point scale. (MC)

Other than using it as a metaphor to learn from the language literacy domain, the term was less attractive for some of us. MH, for example, is hesitant to use the term when talking

about specific skills:

I try not to use the word “literacy” because I feel like literacy has so many connotations that come with it in terms of societal implications and in terms of when you were calling someone literate, you were also calling someone else illiterate. (MH)

This echoes a prior sentiment from MC [48], which he elaborated on during his interview when asked, *have you encountered any challenges communicating visualization literacy as a concept?*

The main challenge as mentioned is that I don’t think it’s a meaningful concept. But I’ve definitely already lost that battle with respect to the vis community, right? [...] I feel like [the visualization literacy concept] is often used as an excuse for not doing a good job as a designer, or often as a way to maybe downplay effects that we see [...] that challenge conventional wisdom about data visualization. Like, they’ll go [...] “well, you see this thing in graphical perception, but that’s probably with, you know, low graphical literacy crowd workers, of course doesn’t apply to us.” (MC)

When visualization literacy guides how we design and evaluate visualizations, the “literate” versus “illiterate” split that the term imposes may not serve the goal of improving design for a wider audience [48]. There are also inconsistencies associated with the usage, because visualization literacy can encompass many different complex skills (see Section 6.4.1):

If you say the word “literacy” to 10 different people, there’s gonna be 10 different kinds of internal understandings of what that is. (MH)

Some of us worry that **the broadness of the term and its inconsistent usage reduces clarity**. Those who expressed this sentiment proposed alternative terms such as “graphical com-

prehension”, “visualization ability”, “visualization competency”. LH differed:

I think people will know what you mean when you say “visualization literacy”, but if the term were changed to, I don’t know, just “visualization ability”, I think it would potentially fall flat and maybe fail to connect with the audiences outside of visualization that we wish to engage with. So I’m a little comfortable with ambiguity and impreciseness in service of connection. (LH)

A more broadly-familiar term might be an acceptable compromise to ease conversations with other related domains (e.g., statistics, cognitive science). Still, our differing points of view on the term itself further reflect the complexity of visualization literacy and our varying priorities.

6.4.4 Calls to Action

It is difficult to pinpoint exactly what we are measuring, as visualization literacy is an inherently complex, context-dependent construct. This poses definitional challenges, which inevitably lead to measurement challenges. To progress, we must acknowledge we are tackling a wicked problem and be ready to make compromises in definition and measurement.

Expand the Scope of Visualization Skills We Assess

We generally expressed concerns about the limited scope of visualization tasks measured in existing assessment tests and a need to improve the ecological validity of visualization assessments (Section 6.4.1). Future work should focus on assessing higher-level activities with data visualization, such as decision-making or the ability to create data visualizations. We should **seek ecological validity by focusing tests on what people actually need to use visualizations for**. Although I have made efforts to shift the focus of visualization literacy assessments towards high-level tasks like visualization construction (Chapter 5) or the detection of misleading representations of data (Chapter 3), there is still much to do to expand the scope of higher-level skills we assess. Other opportunities to broaden the scope of assessments include measures of skills needed to make sense of different types of data displays (e.g., interaction, animations). Indeed, existing tests mostly rely on static representations, while many visual displays of data are interactive.

Current tests also typically only cover common, basic chart types (e.g., cartesian planes and maps). We should **expand the scope of assessments to less common—but useful—chart types and idioms**, such as icon arrays, which deviate from what we can expect in a K–12 curriculum. Given the difficulty of precisely describing the underlying skills involved in visualization proficiency (Section 6.4.1), we may also lean more on qualitative methods to surface implicit visualization skills before attempting to assess them. Broadening the visualization skills we assess, however, will require careful consideration of potentially-complex relationships between those skills in order to ensure valid measurement. For example, a person’s ability to understand interactive visual data displays might partially depend on their ability to interact with digital interfaces at large. While we offer some ideas for broadening test scopes, it would be a useful exercise to speculate about what the world of visualization literacy measurement could look like, as a way to inform how it should be.

6.5 Operational: How Do We Use Existing Tests?

Despite the difficulties with conceptualizing visualization literacy, researchers have already attempted to measure it to better understand their target population's visualization literacy levels [70] or evaluate the effectiveness of interventions to improve visualization skills [5, 21, 69]. As test users, we share the considerations that go into deciding when and how to use a particular test, such as our research goals and community norms. We also reflect on the assumptions that often come with these tests, which can influence our decisions on test usage in practice.

6.5.1 In Pursuit of Standard and Objective Measures

Researchers use existing visualization literacy assessments for a variety of reasons. One reason that emerged is **to measure visualization literacy in an objective way**. We note that early assessments answered a then-emerging need in the community for more objective measures:

Before we had the assessment test, we might ask people to self-report their comfort with things like statistics or visualizations. And I remember distinctly from that that some of the times people greatly overestimated or they put like a really high score on their self report, and yet the performance was pretty low in comparison to other people that we tested. (LH)

The desire for more objective tests in the community led to the wide adoption of some of these early assessments, as LH reflected: “they were kind of widely adopted, because people were literally searching for this sort of thing and it was available” . Because VLAT, as one of the earliest tests in the community, offered a more objective way to assess visualiza-

tion skills than what was previously possible, it made sense that most of us who used tests in experimental contexts have used or considered using VLAT. LH explains that part of this test's success stems from its foundation in existing community standards:

VLAT is very logically constructed in that it kind of builds on well-accepted task taxonomies [...] So, there's like a logic to it that's very easy to explain and very easily could result in like a series of questions. And then the scoring is very logical as well. (LH)

Across our interviews and conversations, there were many instances where we described choosing the latest community standard when making decisions about how we conduct our studies. Reflecting on these matters, we note that **community norms also exert influence on study design choices**. These observations reflect a potential desire in the community to have a “universal” test to produce “generalizable” results. Having a standardized measure becomes an advantage in research contexts when the goal is to compare results across multiple studies. In such cases, we might choose the same measurement scale as in existing work. However, MC pointed out the tension between efforts to obtain generalizable results and the need for scales tailored to specific study interests:

If everyone's using their own metric, then it's hard to make these generalizable or universal claims, but if really what we want to get out of those scales are very-specific-skills-based, or are people reliably pulling the right things out of this [specific] chart, then that sounds to be much more contingent and less universal. (MC)

We as researchers should carefully consider such trade-offs and be clear on the goal of our studies when selecting a measurement scale, as **the most appropriate measure might not be the community standard**. We acknowledge that there is no easy solution to resolving this tension. On one hand, an appropriate test that is well-suited to a particular research

question may not always exist, often due to the difficulty of conceptualizing what we want to measure (Section 6.4). On the other hand, when an appropriate test is not one that is a standard in the community, community norms can impose pressure on researchers to not use it.

Our discussion also revealed a need for assessments that extend outside of the research community. For example, some of us have received outreach from industry about using VLAT “as a part of a manual or tutorial for their company workshop or orientation” (BCK), or A-VLAT [55] (adaptive VLAT) to assess the visualization literacy of a company’s employees. This reflects a desire in industry to use these assessments as “[part] of their onboarding process, onboarding materials” (YD).

Interest outside of academia demonstrates the potential for increased impact of visualization literacy assessments, yet also highlights further limitations in the current assessment landscape. For example, current assessments are typically designed with no feedback, and assign a single number (score) only at the end. This makes individual or group learning difficult or impossible in some cases, and highlights barriers that would need to be addressed before organizations could make effective use of the current set of visualization literacy assessments.

6.5.2 Implementation Challenges Burdening Researchers

Reflecting on our practical experiences using tests, **it was a lot easier to adopt tests with readily available materials**, which also potentially contributed to the wide adoption of VLAT: the straightforwardness of implementing and scoring VLAT made it easy for researchers to use. However, we also surfaced several practical challenges during use that influenced our decisions in choosing tests. One of the most often mentioned barriers was that the **test**

length may be too long for some study designs, which could introduce an unnecessary burden to participants. Researchers often evaluate how a test might fit into their overall study design:

If you are just trying to check their experience levels, it feels very costly for the users 'cause [it] takes an hour or 30 minutes to an hour overhead cost and your main test might take three minutes. (BCK)

The time requirement for some of these tests led to efforts to reduce the length, such as shortening it [130] or adaptively and more efficiently assessing the ability [55]. **Some of these enhanced methods require more technical knowledge to implement**, making them harder to adopt:

I think there are a lot of potential technical hurdles, like little engineering things that don't seem like a big problem, but could really affect the choices that a research team makes, especially under time pressure. [...] Running the psychometric or statistical models to analyze participants' data on the assessments requires some knowledge, first of all, theoretical knowledge about how those models work, but also more practical implementation knowledge. [...] We were lucky in that we received a lot of help and from people with these expertise. But generally speaking, I think it's not a trivial process. (YC)

6.5.3 Non-replication of Tests Across Contexts

Because each test was created within a certain context, it likely comes with assumptions about the test population. Such assumptions relate to some baseline knowledge that the test-takers are expected to possess regardless of their visualization skills, such as understanding a language (e.g., English) or being familiar with the topic of a visualization (e.g.,

geospatial maps, politics), or knowledge specific to their educational level (e.g., high school, college). **Cultural assumptions impact how researchers adopt the tests in differing contexts.** As existing tests were originally developed in English, with Western-centered cultural assumptions, there is a need for translating and adapting these tools to other languages. BCK recalls that “there were at least two or three people asking me for permission to translate [VLAT] into their own language.” However, adapting a measuring instrument across cultures is a time-consuming process involving terminology challenges and revalidation steps such as back-translation. NR, who was motivated by the cultural dependency of the visualization literacy tests like VLAT and worked on cross-cultural adaptation, recalled many trade-offs:

We absolutely did not know how do we say “heat map” in [Arabic] and we just made a compromise of something that closely sounds like a heat map if someone heard about it. So that was a challenge to find the right translation, the taxonomy, in the other language.
(NR)

Not only did they need to translate the language used in the tests, NR also had to modify the topics of visualizations and “replace the question with local election systems”. This was necessary to be mindful of sensitive topics across different cultures. Sometimes, the topic of a visual representation could even derail test administration and defeat the purpose of evaluating visualization skills. NR recalled:

When we tried to ask them like, here is a chart about transportation, I think it was about the metro [...] and we asked them like, tell us what you learned from this chart. And they just ended up [...] talking about how messy transportation is in Madagascar [...]. If the culture of reading a chart is not yet established, then people can see the chart as a trigger for parallel discussions, [which] is not necessarily relevant for the evaluation at the moment.

(NR)

This *non-replication* across cultures challenges our rationale in Section 6.5.1 to use the same test to compare results across studies, suggesting that it may only be possible to have “standardized” measuring tools for comparisons within a particular context of use. Using *dialects* as a metaphor for visualization contexts, MK noted: “A lot of quote unquote rules of visualization are actually rules that are rooted in specific dialects.” Such differing contexts are not limited to cultural differences, but may also concern the application domain, such as statistical or spatial data. For example, MMC, who has worked with astronomy visualizations, found a lot of the current assessments to be not particularly useful:

A lot of what we look at for visualization tests measures one’s ability to infer statistical data, which maybe perhaps isn’t necessarily as helpful as like scientific data with seeing like a galaxy formation or trying to come up with a nebula or anything like that. (MMC)

Current tests can only represent a small snapshot of the construct under a very specific context, and may not be built to accommodate the fluid nature of visualization literacy. We should therefore communicate these domain-specific and culturally-dependent test assumptions more clearly to facilitate future test usage.

6.5.4 Calls to Action

How we choose to use a particular visualization literacy test is influenced by factors like research goals, community norms, and ease of implementation. The unspoken test assumptions regarding population characteristics can also lead to usage barriers. We call on test designers to make tests easier to use and the community to build centralized repositories.

Develop How-to Guides for Customization and Adaption

Many of us wished to adapt existing tests while maintaining construct validity. Because of the rigorous test design and validation processes (Section 6.6.2), later test administrators might feel reluctant to adapt tests in fear of affecting the test validity. Yet in some cases, alterations of questions may be more desirable in a particular study context (e.g., cross-cultural adaption in Section 6.5.3 or shorter studies in Section 6.5.2) that is different than that of the original test questions. **It is crucial to clearly communicate test assumptions and the knowledge necessary for adaptation**, so users can better adjust the test in different contexts. LH entertained the idea of having a “how-to” guide for customization:

I think that the main limiting factor [is the absence of] a widely accepted way of paring down the questions or slicing and dicing, and what that means if you do that. I would like to have the paper that says “this is justified”, or “this is how you construct something for a particular context” with justification [...] I would follow that process potentially, if it wasn't too onerous.(LH)

This could help reduce the burden researchers might feel when they need to make alterations, and could improve the alignment between study designs and measurement.

Form Centralized Assessments and Implementation Materials

The availability of materials is a huge factor in implementing and administering tests (Section 6.5.2). **It is crucial to make tests and the necessary materials readily available.** We can also benefit from a collective effort toward a centralized repository that contains the different assessments, including a modular approach to cover the multidimensionality of the construct. Such a centralized approach could preserve the benefit of having individual tests for isolating particular components of visualization literacy for measurement and, at the same time, offer easily accessible ways for users to find and implement their desired tests. Existing efforts in the community are already taking steps toward better assisting users in implementing and administering tests, such as the reVISit framework [60]. We can build upon and extend prior efforts in centralizing these tests to lower the barrier for using assessments of visualization literacy.

6.6 Methodological: How Do We Design Tests?

Designing a valid test is time- and resource-consuming. The subset of us who have designed and validated existing tests all acknowledged that it was no easy feat. Many considerations went into designing valid tests, such as what we want to (or can) measure and how we can measure it. Depending on what we are measuring, the way we assess the validity of a test also differs. There are many challenges during test development from a test content and format perspective, which lead to the compromises we often have to make as designers of visualization literacy assessments.

6.6.1 Test Modality Limits What We Can Measure

As a community, we took very logical first steps in studying and measuring visualization literacy by seeking theoretical and methodological insights from the psychology and educa-

tion sciences, such as Item Response Theory [26] and systematic test development frameworks [44, 106]. These processes remain important sources of structure for test design. However, **areas of test design such as test format and scoring have stayed relatively constant over time**, despite the fluidity of visualization literacy as a construct. This misalignment impacts what we can measure with existing test formats.

We observed a tendency for test designers to lean more on quantitative measures. This tendency then favors test modalities that are easy to implement and summarize quantitative results from, such as multiple choice questions.

There is a general assumption that if you're gonna build an assessment, you want to build something where you don't have to have a human sit down and grade it. Which, you know, may or may not be a good baseline assumption. (MK)

While multiple choice questions can provide a range of benefits from an implementation perspective, such as being easy to grade, there are still challenges during test design. One challenge is **designing meaningful distractors in addition to not-so-obvious correct answers**. The quality of distractors in multiple choice questions has a direct impact on the question difficulty and test validity at large. For instance, if the correct answer is too obvious, then the question may not capture the visualization skill it aims to assess. Depending on the purpose of the test, sometimes distractors need to be designed in a very specific way, as LG recalled:

For CALVI, [...] there is definitely an answer that is considered correct, but also [a distractor] that is seemingly correct if they didn't realize the issue. So that makes it a little bit iterative because we needed to make sure that the visualization for that question can actually mislead the person, if they didn't notice that issue, to choose that answer. We had

to think about [the visualization and the distractor] together when designing the question, the visualization, and the misleading answer. So that was a lot to consider at once, but for each item, that was necessary because we needed something to catch whether they identified the issue or not. (LG)

The challenge of designing meaningful distractors may be easier to overcome for measurements of visualization interpretation skills compared to construction-related skills, as the *combinatorial* nature of visualization construction is less likely to be supported by the multiple choice format [71]. **The multiple choice modality thus limits what visualization skills we can assess.**

Following the multiple choice questions format, the outcomes of these tests naturally become aggregated scores—sometimes broken down by chart type. This straightforwardness in scoring also contributed to the wide adoption of some tests, as discussed in Section 6.5.1. However, because the results output, or scoring, influences the amount of information researchers can retrieve from using assessments, **the simplicity of the resulting score may lead to less-refined analyses of visualization literacy as a construct.** Noticing this limitation, several of us voiced a desire for more insight on test-takers' visualization abilities and suggested leaning more on qualitative methods.

I wish we could learn more about how people get things wrong when they take the assessments [...]. Right now it's mostly like a number, so that's less meaningful. So maybe more open-ended questions on these assessments. But of course that introduces additional complications such as grading and not being objective. But that is the trade-off. (YC)

I would love to see scoring rubrics for [...] open-ended answers, which are constructed responses [in learning sciences]. They come with this “rubric” that helps you score analytically or holistically. (AFC)

Offering a different reason on why we should adopt more qualitative rather than quantitative ways of assessment, MC expressed:

There’s a lot of things that we seem to assess quantitatively in the vis field that I don’t know if we’ve done the necessary qualitative observational stuff that seems to be a prerequisite for some of those things yet. [...] I just don’t know how people read bar charts, for instance [...] Are they anchoring at the tip of that line and then projecting it to that y-axis and then converting that to a number in their head, and then doing comparisons of that number? Are they doing [...] very quick area estimations? [...] So when you get from there to “our design was better because they completed task one, like 0.5 seconds faster on average”, there’s a huge gap between those empirical claims to me and how useful those empirical claims are, or quantitative empirical claims. (MC)

This relates to the implicit knowledge that we might internalize through experience (Section 6.4.1), which could be key to interpreting some of the results from quantitative assessments. Surfacing this implicit knowledge, such as through qualitative methods, might be necessary to lay a solid foundation for building quantitative assessments. However, the very reason that contributed to the popularity of these tests—logical test structure and scoring—may now be driving people away from alternative test formats that contradict those properties. LG wondered:

Are we just using multiple choice as a default without thinking about the more complex ways of measuring stuff? And if we are, then we are maybe missing some opportunities

to be able to measure these higher-level skills. (LG)

While we were able to rely on certain test formats like multiple choice questions to collectively gain valuable insights on visualization literacy, staying too comfortable with current attempts of assessing the construct may lead to missed opportunities for eliciting and understanding all of the exciting aspects of visualization literacy, as choices on test modality also directly influence what we can effectively measure.

6.6.2 The Need to Reduce Noise

When designing test questions, it is difficult to determine the correct level of abstraction to operate on. In other words, the granularity of the categorization of visualization skills directly impacts what each test question will focus on. Test designers often have to determine the taxonomy of visualization tasks to base their test questions on, such as lower-level or high-level categorizations of visualization skills. A more granular categorization could result in redundant test questions and risk the diversity of the question bank. A less granular categorization, however, might lead to difficulties in translating abstract categorizations into realistic and concrete test questions. As MK reflected,

You can create this whole abstraction, this nice taxonomy, you think that it works well and then you start making the actual charts and you realize [...] it's more difficult when you actually see the real thing. (MK)

This creates a challenge that test designers often have to battle with, which first starts with some abstractions (e.g., visualization tasks), and “the question then becomes, okay, we have to design a bunch of visualizations that seem somewhat realistic[...and not something that] you never really would see in the real world” (LG).

On top of the compromises we often have to make on the level of abstraction we can operate on while still staying somewhat realistic, we are faced with the the need to reduce noise when designing the questions in the test. For instance, as we pointed out in Section 6.4, visualization skills are difficult to isolate from other skills and knowledge.

There will be a lot of noise coming from knowledge about this topic, knowledge about the dataset. Because when you ask a question about something, they don't necessarily have to use the chart. They can just get something out of their memory or something. So we have to be careful in that aspect. (BCK)

As a result, the design process for individual questions is often very iterative and involves qualitative work, which was evident in prior work in test development and evaluation [106, 72, 55, 56, 31, 71, 84]. Such careful iterations are necessary, because isolating the relevant visualization skill and reducing noise in the questions is crucial for the validity of a test. These challenges of test design can only increase as we shift to assessing higher-level visualization skills, such as visualization construction, visual decision-making, or visual metaphors, because the sources of noise inevitably increase as complexity increases (Section 6.4).

6.6.3 Calls to Action

Test designers face many challenges during the development process and often have to make compromises due to the complexity of the construct as mentioned in Section 6.4.1. The test modality can also constrain what the test is able to measure, thereby influencing the ecological validity of a test, as oftentimes what we *can* measure is not aligned with what we *need* to use visualizations for in the real world. We call on test designers to seek ecological validity, take an iterative test design approach over time, and lean on specialized knowledge to support test development and validation.

Improve Ecological Validity of Tests

When we create visualizations and questions for assessments, we must face difficult trade-offs between abstraction and realism. Translating tasks from theoretical taxonomies into concrete visual representations and instructions is not a trivial process and requires careful, iterative design processes. Our efforts to minimize noise from individuals' knowledge or contextual factors in tests may also come at the cost of reducing the ecological validity:

I would like [...] higher-level and then also more realistic use cases, like the ways that people would use visualizations in their lives versus just kind of completely decontextualized. (MH)

[In] health literacy, there's a decision making element, right? Because if you have a chart in that field, usually it's to make some diagnosis[...]or] life changes to be healthier. And then versus something like Cleveland and McGill, there's not any decision really to be made. It's just making a judgment, but that's not teaching you how to apply that information in what you do day to day. So maybe that could mean having more realistic examples in these assessments. (KB)

To improve ecological validity, **we should situate measurements in real-world contexts**. Such measures could better reflect people's true abilities in visualizations. Informal learning environments such as the educational game from Chapter 4 could be one of those contexts.

Expand Assessment Modalities

Following our call to expand test scopes (Section 6.4.4), the format we use to assess these higher-level skills would have to change as well, as multiple choice questions are less likely to offer adequate support (Section 6.6.1). AVEC from Chapter 5 demonstrated one alternative format other than multiple choice. We should make more efforts to **embrace alternative assessment formats**, such as using more open-ended questions. We expect the difficulty of isolating visualization skills to generally increase as the complexity of the ability we assess increases, due to the increasing sources of noise (e.g., complex assessment tools, technical skills). It is important, however, to not overly bind ourselves with the properties (e.g., logical, straightforwardness) that contributed to our progress in visualization literacy, as the nature of the construct itself (e.g., complex, fluid) often contradicts those properties. We must expand how we approach visualization literacy assessments, because “measurement is in a way influencing people’s perceptions on the construct” (LG).

Seek Interdisciplinary Collaborations to Enhance Measurements

As an interdisciplinary field, visualization literacy can benefit a lot from inputs from related domains, such as education and psychology, to enrich our conceptualization of what visualization literacy is and to improve how we measure it. KB expressed:

We need to work more closely with psychologists to actually get this empirical backing to all the stuff we do because they do this all the time [...] We can bring the CS side and the other things, and they can bring the actual psychology measurements. (KB)

Formal training in methods such as statistics and psychometrics testing that are necessary for analyzing and validating tests is still lacking in our community. **We should embrace opportunities to lean on specialized knowledge to support test development.** For example, psychologists have proposed theoretical and methodological work on ensuring cross-cultural validity [76, 82] when translating measurement instruments (something we are already attempting, as discussed in Section 6.5.3). Such work, however, might not be easily accessible to researchers in our field, as it requires specialized knowledge that is often outside our typical scope. **Cross-domain collaborations could address the many intricacies of psychometric measurements.** It is good that we are already making these connections with meetups, panels, and workshops, which we have already seen benefits from, such as this position paper. We could leverage these connections and continue to spur interdisciplinary conversations, which would be important for fostering collaboration across these domains.

6.7 Discussion

6.7.1 Calls for Mindset Changes in the Community

In this autoethnography we identified several *calls to action* from the conceptual, operational, and methodological perspectives. Here, we call for additional *mindset* and *attitude* changes that are more cross-cutting and applicable to the community at large.

Be ready to compromise between what you *need* to measure and you actually *can* measure, because the fluidity of visualization literacy makes it an ever-moving target. A single test that covers every dimension is difficult to achieve. AL recalled: “My initial study was [...] to touch different dimensions [...] [in one study,] but things were messy, because the test was not so reliable.” We may thus need to develop tests specific to a particular domain or context. However, too individualized an approach to test development may make it difficult for test users to get a holistic understanding of participants’ visualization literacy. Regardless of the context or literacy component one focuses on, **authors should explicitly define which components of visualization literacy they study** to facilitate clear communication within the field and outside of it. At a certain point, we may need to simply acknowledge that there is a limit to what we can precisely measure.

Resist the urge to choose a measure just to conform to community standards. Due to the complexity of visualization literacy and the small set of existing assessments, it is easy to end up studying aspects of visualization literacy not covered by existing assessments. In such cases, **it is not appropriate to use a widely-adopted test just to check the “measure visualization literacy” box.** Visualization literacy tests should not be used mindlessly to conform to a perceived standard, just as user studies should not be mindlessly used to

evaluate a visualization system if another evaluation is more appropriate[79]. While we acknowledge that it may be difficult to resist these normative expectations, particularly during peer review, we are hopeful that the community will be receptive to well-reasoned deviations from what might be considered standard.

Design and re-design tests on a longer time scale, iteratively. As we are far from solving the visualization literacy measurement problem, existing tests should not be viewed as “complete” products. **We should scrutinize existing tests**, especially when we use them in practice. Some areas for improvement can only be discovered through actual use, not via theorizing about how we should measure visualization literacy. Test users should report what worked and did not work whenever possible, which will inform test designers on areas for iterative improvement.

Take outward-facing responsibility for how tests are disseminated and maintained, because diverse user groups outside of academia are also using these tests (Section 6.5.1). Reflecting on the web version of VLAT and its role in connecting different audiences, BCK recalled:

I still get, from time to time, emails [...] about VLAT, trying to correct some errors in the website, or just asking questions or just sending me a really good vibe about “Hey, thanks for making this” . [...] That means people do need this. (BCK)

Potential test users who are outside of our field might not be as familiar with the nuances of visualization literacy as we are, so they are trusting us to develop valid assessments that they can adopt easily. One way of taking greater responsibility is to talk to users from other domains who are adopting our tests to learn how we can further improve how we measure visualization literacy.

6.7.2 A Reflection on Our Methodological Approach

Our approach to surface tacit thoughts relied on a combination of collaborative autoethnography, thematic analysis, and diffractive analysis. Although conducting autoethnography with a group adds rigor, it has known logistical challenges, including “establishing shared goals, agreeing how the group will work together, and meeting timelines” [103]. These issues could have been exacerbated in a paper with over a dozen authors across multiple continents on a tight timeline. We addressed these challenges through (1) a shared spreadsheet documenting assigned tasks and progress for all analysts, coordinated by the first author and visible to all—creating social accountability; (2) a group meeting where analysts practiced the analysis process on a short interview excerpt; and (3) a flow diagram analysts followed to keep procedures consistent.⁸ These procedures helped us be transparent about progress and kept everyone on the same page about shared goals and timelines.

While we leaned on thematic analysis (Section 6.3.3) to formulate our tacit thoughts into a coherent narrative, we intentionally employed diffractive analysis to avoid passive agreement among ourselves. We explicitly added some distance between the analysts and the transcripts, which was our attempt at further reducing bias and increasing nuance to address the weakness of potentially lacking nuance in individual autoethnography [103]. Our approach to collaborative autoethnography created opportunities for a clash of perspectives and helped us distill the many factors that contribute to the wicked problem of visualization literacy measurement. However, we acknowledge that the majority of our perspectives stem from a Western-centric visualization philosophy. Future work could engage in more cross-cultural reflections, which would be especially valuable as visualization literacy itself is fluid across cultures.

⁸Figure 6.1 is a compact version of the diagram we used during analysis.

6.8 Next Steps

In this autoethnographic reflection, we leaned on our professional and practical experiences as users and designers of visualization literacy tests to surface the tacit thoughts often left unreported in individual research papers. Despite our formulation of visualization literacy measurement as a wicked problem—one that is complex and without an exhaustive set of potential solutions [137]—we identified some concrete next steps to study visualization literacy more comprehensively. For example, test designers should expand the scope of their tests, both in terms of content and modality, and work towards improving the ecological validity of tests. Test users should report both what worked and what did not to help us collectively iterate on test design and how-to guides. We should develop centralized repositories for implementing tests. Most importantly, we should shift our mindsets to recognize the construct’s fluidity and expect change in both our understanding of the construct and its measurements. As suggested by the formulation of this as a wicked problem, we do not claim to have an exhaustive set of rules or procedures for readers to follow. Instead, we advocate embracing the wickedness of this measurement problem, being willing to make compromises, and continually discussing—and updating—our conceptualizations of visualization literacy.

Chapter 7

Discussion

7.1 Assessment Tools for Complex Visualization Skills

Ultimately we can only improve what we can measure, so precise, valid measurements of visualization literacy are widely needed in order to work toward a more comprehensive understanding of the construct. However, an immediate challenge facing studies of more complex skills in visualization literacy, such as visualization construction, is how the assessment tool might affect the measurements of such abilities. Unless people are hand-drawing visualizations, their experience with the tool they use to construct a visualization inevitably affects the kinds of visualizations they are able to create. This presents a challenge when the goal is to measure their ability to create visualizations, not their familiarity with a specific tool. For example, a tool with a steeper learning curve that is familiar only to some users (e.g., Tableau, PowerBI) might artificially deflate unfamiliar participants' scores, even if they do have high visualization construction abilities. Thus, part of the validity of a test of construction-related abilities is dependent on demonstrating that the usability of the assessment tool does not adversely affect measurement of the ability. At a minimum, assessment tools should adhere to acceptable standards of usability (e.g., [28]) and provide adequate training to ensure participants familiarize themselves with the tool prior to as-

essment.

The assessment tool used to administer AVEC could be expanded to create a more versatile tool for measuring other construction-related abilities. However, arriving at a simple-to-use tool that can also measure complex construction-related abilities is a challenging design problem: higher-level skills often require more functionality, but a more complex tool runs the risk of confounding measurements of a person's ability with measurements of their familiarity with the assessment tool (or other tools like it). Future work for assessing and studying more complex skills should especially pay attention to the tool-building component and might also investigate people's ability to construct different types of visualizations with a minimal usage of computerized tools (e.g., drawing). This type of comparison study could offer insights on just how much a tool might contribute to a person's overall ability of constructing visualizations.

In the context of visualization interpretation, where there are much more established measurements, the assessment tool is essentially the multiple-choice questionnaire. This question of assessment tools then echoes the concern of test modality discussed in Section 6.6.1: because the assessment tool affects the validity of the measurement, test administrators must consider whether the multiple-choice format aligns with the underlying ability they are trying to measure.

7.2 How Are People Using CALVI?

Since CALVI's publication in CHI 2023, it has received 127 citations on Google Scholar¹. To better understand the impacts of CALVI, I conducted a forward citation analysis. After excluding theses and papers that are not written in English, 99 papers remained in the cor-

¹As of June 16, 2026.

pus. Out of the 99 papers, 17 of them actually used CALVI beyond simply citing it as related work. There were three major categories of use: psychometric reuse in human studies, as a design and teaching resource, and as a benchmark for AI / LLM Evaluation. These applications of CALVI surfaced several interesting findings that I discuss below.

7.2.1 What can CALVI inform us about visualization literacy?

One of the papers that used CALVI in human studies investigated the role of visual attention and found that attention patterns can distinguish between experts and novices of visualizations [39]. As discussed in CALVI, some misleaders are misleading because of visual flaws that require the viewer's attention to the right part of the visualization [72]. Experts may be well-trained to check the visual properties of visualizations compared to novices, so it makes sense that experts would show a strong attentional focus [38]. Certain CALVI questions (e.g., easier ones) could also reinforce this by using explicit annotations that mark the misleader in a question. Other findings from this corpus strengthen the need for investigating visualization literacy. For example, Verma and Fan [155] employed visualization literacy test questions to identify factors that could predict the questions' difficulty. However, they found that chart type and task were not sufficient in explaining the variation in question difficulty [155]. This further suggests that more factors are affecting visualization understanding than what we can currently name. Reflecting on the development of CALVI, other dimensions of characterizing items could be the length of the multiple-choice options; longer texts in the options may be harder to parse, and therefore could contribute to the difficulty level of an item. A better (perhaps more granular ways) of characterizing visualization test questions may be a useful step toward a more clearly defined construct.

7.2.2 Who Are the Misleaders Misleading?

Out of the three categories of use identified from the forward citation analysis described in Section 7.2, AI evaluation was the largest category ($N = 8$). While this is not necessarily surprising, considering the surge in AI /LLM research in recent years, it serves as an opportunity to reflect on *what it means for a psychometric instrument to travel beyond its validation context*. CALVI was designed and validated with a human population, so its psychometric parameters such as item difficulty is based on human characteristics. For example, what makes a misleader misleading to humans may not be what challenges a model, due to the absence of the perceptual biases and social expectations that shape human chart reading. Several papers that used CALVI to evaluate a model’s visualization understanding broke down model performance by misleader categories. It was interesting to see that models are performing better on misleaders that are a lot harder for humans, such as missing normalization (e.g., [153, 129, 41]). This suggests that models have a fundamentally different way of interpreting data visualizations compared to humans, making results from a test designed to capture human misinterpretations hard to interpret when used in a context that functions in a different way. Because model visualization understanding cannot be assumed to mirror its human counterpart, it would be necessary to first conceptually define what model visualization understanding actually is. This would yield more productive and meaningful benchmarks that would put meaning into accuracy numbers.

7.3 Measurements for Understanding Versus Measurements for Evaluation

Reflecting on my work in both measurements and intervention, I am realizing a distinction between measurements for *understanding* and measurements for *evaluation*. A measurement that works in one context may not be well suited in another, even when the underlying ability is the same. For example, standardized tests like VLAT [106] and CALVI [72] work well in human empirical studies that aim to learn something about the population being tested, such as their ability to read visualizations and perform visual tasks. But standardized tests in the context of interventions could introduce unnecessary challenges and noise for the evaluation of its effectiveness.

One such challenge is the time requirement, which was also discussed in Chapter 6. Standardized tests are usually time-consuming, so it is sometimes impractical for researchers to administer. Another challenge relates to how well the assessment aligns with the goal of the intervention. Through the development of the education game described in Chapter 4, I have observed that it would be out of place to administer CALVI in an informal, gamified setting. This suggests a need to take a step back and consider measurements and interventions as a whole.

The current workflow for intervention studies is generally that measurement development is separate from intervention development, and when it comes to the evaluation, the researcher would then apply a measure to test the effectiveness of their intervention. However, this separation could be the source of threats to validity in intervention studies, because the measurement format may not be the best fit for evaluation in the context of the intervention. For instance, a measurement in a standardized test format would not be best

suited for evaluating the effectiveness of a museum exhibit. To advance both visualization measurements and interventions for improving visualization skills, we need not only to design measurements that can precisely test a certain ability, but also to consider whether the context of use provides the necessary conditions that make a measurement valid. Conditions could be the attention span required of the test taker or what the user expects out of the intervention. An interesting direction for future work is to design measurements in the context of interventions and streamline the process to enable researchers running intervention studies to adapt, customize, or create measurements better suited for their intervention context.

7.4 Reflections and Future Directions

I took a bottom-up approach to studying visualization literacy by identifying and studying one component at a time. The development process for CALVI and AVEC both started with defining precisely what the component is, then constructing a design space for generating a bank of questions. Not only have the tests contributed new questions to assess visualization literacy, but they have also offered a model for test development that has been replicated and demonstrated to work well. Researchers and educators could follow the same procedure to systematically study other components of visualization literacy. I think measurement design itself is a useful exercise for defining and better understanding visualization literacy, and more work should be done on the measurement front to collectively build a more comprehensive understanding of the construct.

My work have also opened up new avenues for future work in the space of visualization literacy. For instance, building on the systematic procedure for test development, future work could develop a system to help researchers and educators who are new to test

development to design and validate assessments. Not only would this create tests that are valid in a specific study context, it would also help the community to collectively build a repository of validated assessments of visualization literacy. Such a repository would allow future studies of large scale studies of visualization literacy that scattered tests cannot support.

Looking back, the assumptions of these tests offer new opportunities for advancement. For example, test questions are often generated from a two-dimensional design space (i.e., visualization tasks and chart types). This inevitably limits the scope of the questions and sometimes oversimplifies the construct they are trying to measure. Some components of visualization literacy may not be able to be simplified into a set of tasks, such as visualization design related abilities. This may have contributed to the focus on simplified versions of the problem. For instance, all of the tests in the community overlook the iterative process that exist in both visualization interpretation and creation. Reading a visualization often require more than one look of the chart and creating visualization often requires an iterative process. Yet, none of the assessments can capture the intermediate steps that necessarily exists in visualization understanding. Developing ways to capture these intermediate and iterative (often messier) steps can better inform us of precisely where the challenges exist in visualization literacy. More informative assessments could support the development of more effective interventions for improving visualization literacy.

Chapter 8

Conclusion

Over the course of this thesis, I broadened the conceptualization of visualization literacy and operationalized both the ability to critically interpret and create well-formed visualizations. I also reflect on the process of assessment design and usage in the community to outline a research agenda toward tackling the measurement challenges facing visualization literacy. We need more efforts from the research community in order to build a comprehensive understanding of the construct. I also demonstrate the development of an educational intervention targeting critical thinking, building on the broadened understanding of visualization literacy. Together, my dissertation highlights the need for better understanding visualization literacy, an approach for doing so through the development of enhanced measurements and interventions, and a path forward for continuing to build a more complete understanding of the construct.

References

- [1] Eytan Adar and Elsie Lee-Robbins. “Roboviz: A Game-Centered Project for Information Visualization Education”. In: *IEEE Transactions on Visualization and Computer Graphics* 29.1 (2023), pp. 268–277. DOI: 10.1109/TVCG.2022.3209402.
- [2] Patrick Adelberger et al. “Iguanodon: A Code-Breaking Game for Improving Visualization Construction Literacy”. In: *IEEE Trans. Visual Comput. Graphics* 31.9 (2025), pp. 5713–5725. DOI: 10.1109/TVCG.2024.3468948.
- [3] Derya Akbaba et al. “Troubling Collaboration: Matters of Care for Visualization Design Study”. In: *Proc. CHI*. CHI ’23. Hamburg, Germany: ACM, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581168. URL: <https://doi.org/10.1145/3544548.3581168>.
- [4] John Alford and Brian W Head. “Wicked and less wicked problems: a typology and a contingency framework”. In: *Policy and Society* 36.3 (2017), pp. 397–413. ISSN: 1449-4035. DOI: 10.1080/14494035.2017.1361634. eprint: <https://academic.oup.com/policyandsociety/article-pdf/36/3/397/42564575/14494035.2017.1361634.pdf>. URL: <https://doi.org/10.1080/14494035.2017.1361634>.
- [5] Basak Alper et al. “Visualization Literacy at Elementary School”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 5485–5497. ISBN: 9781450346559. DOI: 10.1145/3025453.3025877. URL: <https://doi.org/10.1145/3025453.3025877>.
- [6] Lorenzo Amabili, Kuhu Gupta, and Renata Georgia Raidou. “A Taxonomy-Driven Model for Designing Educational Games in Visualization”. In: *IEEE Computer Graphics and Applications* 41.6 (2021), pp. 71–79. DOI: 10.1109/MCG.2021.3115446.

- [7] Robert Amar, James Eagan, and John Stasko. “Low-Level Components of Analytic Activity in Information Visualization”. In: *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*. INFOVIS’05. USA: IEEE Computer Society, 2005, p. 15. ISBN: 078039464x. DOI: 10.1109/INFOVIS.2005.24. URL: <https://doi.org/10.1109/INFOVIS.2005.24>.
- [8] Henintsoa S. Andrianarivony et al. “Investigating the Use of Native and Secondary Language with Data Visualization in Madagascar”. In: *IEEE VIS Posters*. Poster paper. Oklahoma City, 2022. URL: <https://osf.io/e4cwz>.
- [9] Benjamin Bach et al. “Challenges and Opportunities in Data Visualization Education: A Call to Action”. In: *IEEE Trans. Visual Comput. Graphics* 30.1 (2024), pp. 649–660. DOI: 10.1109/TVCG.2023.3327378.
- [10] Benjamin Bach et al. “Special Issue on Visualization Teaching and Literacy”. In: *IEEE Computer Graphics and Applications* 41.6 (2021), pp. 13–14. DOI: 10.1109/MCG.2021.3117412.
- [11] S. Sandra Bae et al. “Cultivating Visualization Literacy for Children Through Curiosity and Play”. In: *IEEE Trans Vis Comput Graph* 29.1 (2023), pp. 257–267. DOI: 10.1109/TVCG.2022.3209442.
- [12] Aaron Bangor, Philip Kortum, and James Miller. “Determining what individual SUS scores mean: adding an adjective rating scale”. In: *J. Usability Studies* 4.3 (May 2009), pp. 114–123.
- [13] Karen Barad. “Diffracting Diffraction: Cutting Together-Apart”. In: *Parallax* 20.3 (2014), pp. 168–187. DOI: 10.1080/13534645.2014.927623. eprint: <https://doi.org/10.1080/13534645.2014.927623>. URL: <https://doi.org/10.1080/13534645.2014.927623>.
- [14] Karen Barad. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press, 2006. ISBN: 978-0-8223-3901-4. DOI: 10.1215/9780822388128. URL: <https://doi.org/10.1215/9780822388128>.
- [15] Randy Elliot Bennett. “ON THE MEANINGS OF CONSTRUCTED RESPONSE”. In: *ETS Research Report Series* 1991.2 (1991), pp. i–46. DOI: <https://doi.org/10.1002/j.2333-8504.1991.tb01429.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1991.tb01429.x>.

- [16] Sara Beschi et al. “Characterizing Data Visualization Literacy for Standardization: A Systematic Literature Review”. In: *IEEE Access* 13 (2025), pp. 65704–65725. DOI: 10.1109/ACCESS.2025.3559298.
- [17] Sara Beschi et al. *Envisioning Visual Information Literacy: some reflection on conceptual and practical stances*. <https://osf.io/kbaes>. 2024.
- [18] Ryan Best et al. *The 40 Weirdest (And Best) Charts We Made In This Long, Strange Year*. Dec. 2020. URL: <https://fivethirtyeight.com/features/the-40-weirdest-and-best-charts-we-made-in-2020/>.
- [19] Alex Bigelow et al. “Iterating between Tools to Create and Edit Visualizations”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 481–490. DOI: 10.1109/TVCG.2016.2598609.
- [20] Ryan Birchfield et al. “VisQuiz: Exploring Feedback Mechanisms to Improve Graphical Perception”. In: *2022 IEEE Visualization and Visual Analytics (VIS)*. Oklahoma City, OK, USA: IEEE, 2022, pp. 95–99. DOI: 10.1109/VIS54862.2022.00028.
- [21] Fearn Bishop et al. “Construct-A-Vis: Exploring the Free-Form Visualization Processes of Children”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 451–460. DOI: 10.1109/TVCG.2019.2934804.
- [22] R Darrell Bock and Robert D Gibbons. *Item response theory*. Hoboken, NJ: John Wiley & Sons, 2021.
- [23] Katy Börner, Andreas Bueckle, and Michael Ginda. “Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments”. In: *Proceedings of the National Academy of Sciences* 116.6 (2019), pp. 1857–1864. DOI: 10.1073/pnas.1807180116. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.1807180116>.
- [24] Katy Börner et al. “Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors”. In: *Information Visualization* 15.3 (2016), pp. 198–213. DOI: 10.1177/1473871615594652. eprint: <https://doi.org/10.1177/1473871615594652>. URL: <https://doi.org/10.1177/1473871615594652>.
- [25] Denny Borsboom, Gideon J Mellenbergh, and Jaap Van Heerden. “The concept of validity.” In: *Psychological review* 111.4 (2004), p. 1061.

- [26] Jeremy Boy et al. “A Principled Way of Assessing Visualization Literacy”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1963–1972. DOI: 10.1109/TVCG.2014.2346984.
- [27] Matthew Brehmer and Tamara Munzner. “A Multi-Level Typology of Abstract Visualization Tasks”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2376–2385. DOI: 10.1109/TVCG.2013.124.
- [28] John Brooke et al. “SUS-A quick and dirty usability scale”. In: *Usability evaluation in industry* 189.194 (1996), pp. 4–7.
- [29] Paul-Christian Bürkner. “Bayesian Item Response Modeling in R with brms and Stan”. In: *Journal of Statistical Software* 100.5 (2021), pp. 1–54. DOI: 10.18637/jss.v100.i05. URL: <https://www.jstatsoft.org/index.php/jss/article/view/v100i05>.
- [30] Anne-Flore Cabouat et al. *Position paper: A case to study the relationship between data visualization readability and visualization literacy*. <https://hal.science/hal-04523790>. O’ahu (Honolulu), United States, May 2024. URL: <https://hal.science/hal-04523790>.
- [31] Anne-Flore Cabouat et al. “PREVis: Perceived Readability Evaluation for Visualizations”. In: *IEEE Trans. Visual Comput. Graphics* 31.1 (2025), pp. 1083–1093. DOI: 10.1109/TVCG.2024.3456318.
- [32] Kelly Caine. “Local Standards for Sample Size at CHI”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. San Jose, California, USA: Association for Computing Machinery, 2016, pp. 981–992. ISBN: 9781450333627. DOI: 10.1145/2858036.2858498. URL: <https://doi.org/10.1145/2858036.2858498>.
- [33] Jorge D. Camba, Pedro Company, and Vetria L. Byrd. “Identifying Deception as a Critical Component of Visualization Literacy”. In: *IEEE Computer Graphics and Applications* 42.1 (2022), pp. 116–122. DOI: 10.1109/MCG.2021.3132004.
- [34] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, eds. *Readings in information visualization: using vision to think*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. ISBN: 1558605339.
- [35] Marianne Sheelagh Therese Carpendale. “A framework for elastic presentation space”. AAINQ51848. PhD thesis. CAN: Simon Fraser University, 1999. ISBN: 0612518485.

- [36] Courtney Cazden et al. “A Pedagogy of Multiliteracies: Designing Social Futures”. In: *Harvard educational review* 66.1 (1996), pp. 60–92.
- [37] CDC. *U.S. COVID-19 Cases in the Last 7 Days by State and Territory*. <https://twitter.com/CDCgov/status/1387509465013006336>. Apr. 2021.
- [38] Heewon Chang, Faith Ngunjiri, and Kathy-Ann C Hernandez. *Collaborative Autoethnography*. New York: Routledge, 2016. DOI: 10.4324/9781315432137.
- [39] Minsuk Chang et al. “Tell Me Without Telling Me: Two-Way Prediction of Visualization Literacy and Visual Attention”. In: *IEEE Transactions on Visualization and Computer Graphics* 32.1 (2026), pp. 1405–1415. DOI: 10.1109/TVCG.2025.3634815.
- [40] Qing Chen et al. “VizLinter: A Linter and Fixer Framework for Data Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2022), pp. 206–216. DOI: 10.1109/TVCG.2021.3114804.
- [41] Zixin Chen et al. “Unmasking Deceptive Visuals: Benchmarking Multimodal Large Language Models on Misleading Chart Question Answering”. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by Christos Christodoulopoulos et al. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 13756–13789. ISBN: 979-8-89176-332-6. DOI: 10.18653/v1/2025.emnlp-main.695. URL: <https://aclanthology.org/2025.emnlp-main.695/>.
- [42] Ed Huai-Hsin Chi and J.T. Riedl. “An operator interaction framework for visualization systems”. In: *Proceedings IEEE Symposium on Information Visualization (Cat. No.98TB100258)*. Research Triangle, CA, USA: IEEE, 1998, pp. 63–70. DOI: 10.1109/INFVIS.1998.729560.
- [43] Victoria Clarke and Virginia Braun. “Thematic analysis”. In: *The Journal of Positive Psychology* 12.3 (2017), pp. 297–298. DOI: 10.1080/17439760.2016.1262613. eprint: <https://doi.org/10.1080/17439760.2016.1262613>. URL: <https://doi.org/10.1080/17439760.2016.1262613>.
- [44] Ronald Jay Cohen, W. Joel Schneider, and Renée Tobin. *Psychological Testing and Assessment*. 10th. New York, NY: McGraw Hill LLC, 2022. ISBN: 1260837025.
- [45] John Cook et al. “The cranky uncle game—combining humor and gamification to build student resilience against climate misinformation”. In: *Environmental Education Research* 29.4 (2023), pp. 607–623. DOI: 10.1080/13504622.2022.2085671. eprint: <https://doi.org/10.1080/13504622.2022.2085671>. URL: <https://doi.org/10.1080/13504622.2022.2085671>.

- [46] Bill Cope and Mary Kalantzis. “The Things You Do to Know: An Introduction to the Pedagogy of Multiliteracies”. In: *A Pedagogy of Multiliteracies: Learning by Design*. London: Palgrave Macmillan UK, 2015, pp. 1–36. ISBN: 978-1-137-53972-4. DOI: 10.1057/9781137539724_1. URL: https://doi.org/10.1057/9781137539724_1.
- [47] Michael Correll. *The Mythos of Visualization Literacy*. <https://mcorrell.medium.com/the-mythos-of-visualization-literacy-5dd02db56254>. 2018.
- [48] Michael Correll. *What Does “Visualization Literacy” Mean, Anyway?* <https://medium.com/multiple-views-visualization-research-explained/what-does-visualization-literacy-mean-anyway-22f3b3badc0>. 2018.
- [49] Michael Correll, Enrico Bertini, and Steven Franconeri. “Truncating the Y-Axis: Threat or Menace?” In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–12. ISBN: 9781450367080. DOI: 10.1145/3313831.3376222. URL: <https://doi.org/10.1145/3313831.3376222>.
- [50] Michael Correll and Jeffrey Heer. “Black hat visualization”. In: *Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVE), IEEE VIS*. Vol. 1. Phoenix, Arizona, USA: IEEE, 2017, p. 10.
- [51] Michael Correll and Jeffrey Heer. “Surprise! Bayesian Weighting for De-Biasing Thematic Maps”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 651–660. DOI: 10.1109/TVCG.2016.2598618.
- [52] Niall McCarthy. *Confirmed COVID-19 Cases in the U.S.* <https://www.statista.com/chart/20978/coronavirus-cases-us-map/>. Apr. 2021.
- [53] Mackenzie Creamer, Lace Padilla, and Michelle Borkin. *Finding Gaps in Modern Visualization Literacy*. 2024. DOI: 10.31219/osf.io/jy9v2. URL: osf.io/jy9v2_v1.
- [54] Lee J Cronbach. “Coefficient alpha and the internal structure of tests”. In: *psychometrika* 16.3 (1951), pp. 297–334.
- [55] Yuan Cui et al. “Adaptive Assessment of Visualization Literacy”. In: *IEEE Transactions on Visualization and Computer Graphics* 30.1 (2024), pp. 628–637. DOI: 10.1109/TVCG.2023.3327165.
- [56] Yuan Cui et al. “Promises and Pitfalls: Using Large Language Models to Generate Visualization Items”. In: *IEEE Trans. Visual Comput. Graphics* 31.1 (2025), pp. 1094–1104. DOI: 10.1109/TVCG.2024.3456309.

- [57] Foroozan Daneshzand, Charles Perin, and Sheelagh Carpendale. “Design Process of ‘Shredded Lives’ : An Illustrated Exploration”. In: *2024 IEEE VIS Arts Program (VISAP)*. 2024, pp. 39–52. DOI: 10.1109/VISAP64569.2024.00008.
- [58] Russell Davis et al. “The Risks of Ranking: Revisiting Graphical Perception to Model Individual Differences in Visualization Performance”. In: *IEEE Trans. Visual Comput. Graphics* 30.3 (2024), pp. 1756–1771. DOI: 10.1109/TVCG.2022.3226463.
- [59] Christine DeMars. *Item Response Theory*. Oxford, United Kingdom: Oxford University Press, Apr. 2010. ISBN: 9780195377033. DOI: 10.1093/acprof:oso/9780195377033.001.0001. URL: <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>.
- [60] Yiren Ding et al. “reVISit: Supporting Scalable Evaluation of Interactive Visualizations”. In: *2023 IEEE Visualization and Visual Analytics (VIS)*. 2023, pp. 31–35. DOI: 10.1109/VIS54172.2023.00015.
- [61] Damien Djaouti, Julian Alvarez, and Jean-Pierre Jessel. “Classifying serious games: the G/P/S model”. In: *Handbook of research on improving learning and motivation through educational games: Multidisciplinary approaches*. IGI Global Scientific Publishing, 2011, pp. 118–136. DOI: 10.4018/978-1-60960-495-0.ch006.
- [62] Thomas J. Dunn, Thom Baguley, and Vivienne Brunsden. “From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation”. In: *British Journal of Psychology* 105.3 (2014), pp. 399–412. DOI: <https://doi.org/10.1111/bjop.12046>. eprint: <https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/bjop.12046>. URL: <https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/bjop.12046>.
- [63] Carolyn Ellis. *The Ethnographic I: A Methodological Novel about Autoethnography*. AltaMira Press, 2004.
- [64] Carolyn Ellis, Tony E. Adams, and Arthur P. Bochner. “Autoethnography: An Overview”. In: *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 12.1 (2010). DOI: 10.17169/fqs-12.1.1589. URL: <https://www.qualitative-research.net/index.php/fqs/article/view/1589>.
- [65] Arlen Fan et al. “Annotating Line Charts for Addressing Deception”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3502138. URL: <https://doi.org/10.1145/3491102.3502138>.

- [66] Elif E. Firat, Alena Denisova, and Robert S. Laramee. “Treemap Literacy: A Classroom-Based Investigation”. In: *Eurographics 2020 - Education Papers*. Ed. by Mario Romero and Beatrice Sousa Santos. The Eurographics Association, 2020. ISBN: 978-3-03868-102-1. DOI: 10.2312/eged.20201032.
- [67] Elif E. Firat et al. “P-Lite: A study of parallel coordinate plot literacy”. In: *Visual Informatics* 6.3 (2022), pp. 81–99. ISSN: 2468-502X. DOI: <https://doi.org/10.1016/j.visinf.2022.05.002>. URL: <https://www.sciencedirect.com/science/article/pii/S2468502X22000377>.
- [68] FiveThirtyEight. *Data Visualization – FiveThirtyEight*. <https://fivethirtyeight.com/tag/data-visualization/>. 2022.
- [69] Johannes Gäbler et al. “Diagram Safari: A Visualization Literacy Game for Young Children”. In: *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. CHI PLAY ’19 Extended Abstracts. Barcelona, Spain: Association for Computing Machinery, 2019, pp. 389–396. ISBN: 9781450368711. DOI: 10.1145/3341215.3356283. URL: <https://doi.org/10.1145/3341215.3356283>.
- [70] Mirta Galesic and Rocio Garcia-Retamero. “Graph Literacy: A Cross-Cultural Comparison”. In: *Medical Decision Making* 31.3 (2011), pp. 444–457. DOI: 10.1177/0272989X10373805. eprint: <https://doi.org/10.1177/0272989X10373805>. URL: <https://doi.org/10.1177/0272989X10373805>.
- [71] Lily W. Ge, Yuan Cui, and Matthew Kay. “AVEC: An Assessment of Visual Encoding Ability in Visualization Construction”. In: *Proc. CHI*. CHI ’25. ACM, 2025. ISBN: 9798400713941. DOI: 10.1145/3706598.3713364. URL: <https://doi.org/10.1145/3706598.3713364>.
- [72] Lily W. Ge, Yuan Cui, and Matthew Kay. “CALVI: Critical Thinking Assessment for Literacy in Visualizations”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581406. URL: <https://doi.org/10.1145/3544548.3581406>.
- [73] Lily W. Ge et al. “An Autoethnography on Visualization Literacy: A Wicked Measurement Problem”. In: *IEEE Transactions on Visualization and Computer Graphics* 32.1 (2026), pp. 1394–1404. DOI: 10.1109/TVCG.2025.3634792.
- [74] Lily W. Ge et al. “Toward a More Comprehensive Understanding of Visualization Literacy”. In: *CHI EA*. CHI EA ’24. Honolulu, HI, USA: ACM, 2024. ISBN: 9798400703317.

- DOI: 10 . 1145 / 3613905 . 3636289. URL: <https://doi.org/10.1145/3613905.3636289>.
- [75] Lily W. Ge et al. “V-FRAMER: Visualization Framework for Mitigating Reasoning Errors in Public Policy”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10 . 1145 / 3613904 . 3642750. URL: <https://doi.org/10.1145/3613904.3642750>.
 - [76] Kurt F. Geisinger. “Cross-Cultural Normative Assessment: Translation and Adaptation Issues Influencing the Normative Interpretation of Assessment Instruments”. In: *Psychological Assessment* 6.4 (1994), pp. 304–312. ISSN: 1939-134X. DOI: 10 . 1037 / 1040-3590 . 6 . 4 . 304.
 - [77] Patricia Geist-Martin et al. “Exemplifying Collaborative Autoethnographic Practice via Shared Stories of Mothering”. In: *Journal of Research Practice* 6.1 (2010), pp. M8–M8.
 - [78] *What is Go Viral?* <https://www.who.int/news/item/23-09-2021-what-is-go-viral>. 2021.
 - [79] Saul Greenberg and Bill Buxton. “Usability evaluation considered harmful (some of the time)”. In: *Proc. CHI*. CHI ’08. Florence, Italy: ACM, 2008, pp. 111–120. ISBN: 9781605580111. DOI: 10 . 1145 / 1357054 . 1357074. URL: <https://doi.org/10.1145/1357054.1357074>.
 - [80] Anzu Hakone et al. “PROACT: Iterative Design of a Patient-Centered Visualization for Effective Prostate Cancer Health Risk Communication”. In: *IEEE Trans. Visual Comput. Graphics* 23.1 (2017), pp. 601–610. DOI: 10 . 1109 / TVCG . 2016 . 2598588.
 - [81] Thomas M. Haladyna. *Developing and validating multiple-choice test items*. eng. 2nd ed. Mahwah, N.J: L. Erlbaum Associates, 1999. ISBN: 0585121303.
 - [82] Ronald K. Hambleton and Anil Kanjee. “Increasing the Validity of Cross-Cultural Assessments: Use of Improved Methods for Test Adaptations”. In: *European Journal of Psychological Assessment* 11.3 (1995), pp. 147–157. ISSN: 1015-5759. DOI: 10 . 1027 / 1015-5759 . 11 . 3 . 147.
 - [83] Shiqing He and Eytan Adar. “VizItCards: A Card-Based Toolkit for Infovis Design Education”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 561–570. DOI: 10 . 1109 / TVCG . 2016 . 2599338.

- [84] Tingying He et al. “BeauVis: A Validated Scale for Measuring the Aesthetic Pleasure of Visual Representations”. In: *IEEE Trans. Visual Comput. Graphics* 29.1 (2023), pp. 363–373. ISSN: 1941-0506. DOI: 10.1109/TVCG.2022.3209390.
- [85] Maryam Hedayati, Ayse Hunt, and Matthew Kay. *From Pixels to Practices: Reconceptualizing Visualization Literacy*. 2024. DOI: 10.31219/osf.io/6mq42.
- [86] Maryam Hedayati and Matthew Kay. “What University Students Learn In Visualization Classes”. In: *IEEE Trans. Visual Comput. Graphics* 31.1 (2025), pp. 1072–1082. DOI: 10.1109/TVCG.2024.3456291.
- [87] Aspen K. Hopkins, Michael Correll, and Arvind Satyanarayan. “VisuaLint: Sketchy In Situ Annotations of Chart Construction Errors”. In: *Computer Graphics Forum* 39.3 (2020), pp. 219–228. DOI: <https://doi.org/10.1111/cgf.13975>.
- [88] Dandan Huang et al. “Personal Visualization and Personal Visual Analytics”. In: *IEEE Transactions on Visualization and Computer Graphics* 21.3 (2015), pp. 420–433. DOI: 10.1109/TVCG.2014.2359887.
- [89] Samuel Huron, Yvonne Jansen, and Sheelagh Carpendale. “Constructing Visual Representations: Investigating the Use of Tangible Tokens”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2102–2111. DOI: 10.1109/TVCG.2014.2346292.
- [90] Samuel Huron et al. “Constructive visualization”. In: *Proceedings of the 2014 Conference on Designing Interactive Systems*. DIS ’14. Vancouver, BC, Canada: Association for Computing Machinery, 2014, pp. 433–442. ISBN: 9781450329026. DOI: 10.1145/2598510.2598566. URL: <https://doi.org/10.1145/2598510.2598566>.
- [91] Elaine Huynh et al. “Designing Narrative-Focused Role-Playing Games for Visualization Literacy in Young Children”. In: *IEEE Trans. Visual Comput. Graphics* 27.2 (2021), pp. 924–934. DOI: 10.1109/TVCG.2020.3030464.
- [92] Yvonne Jansen and Pierre Dragicevic. “An Interaction Model for Visualizations Beyond The Desktop”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2396–2405. DOI: 10.1109/TVCG.2013.134.
- [93] S.H. Jones. “The Sage Handbook of Qualitative Research”. In: ed. by Norman K. Denzin and Yvonna S. Lincoln. Sage, 2005. Chap. Autoethnography: Making the personal political.

- [94] Alark Joshi et al. *Visualization literacy for general audiences-can we make a difference?* IEEE VIS 2021 Panel.
- [95] Mandy Keck et al. *EduVis: Workshop on Visualization Education, Literacy, and Activities*. <https://arxiv.org/abs/2303.10708>. 2023.
- [96] Won Kim et al. "A Taxonomy of Dirty Data". In: *Data Mining and Knowledge Discovery* 7.1 (Jan. 2003), pp. 81–99. ISSN: 1573-756X. DOI: 10 . 1023 / A : 1021564703268. URL: <https://doi.org/10.1023/A:1021564703268>.
- [97] Gordon Kindlmann and Carlos Scheidegger. "An Algebraic Process for Visualization Design". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2181–2190. DOI: 10.1109/TVCG.2014.2346325.
- [98] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. "Trust and Recall of Information across Varying Degrees of Title-Visualization Misalignment". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300576. URL: <https://doi.org/10.1145/3290605.3300576>.
- [99] David R Krathwohl. "A Revision of Bloom's Taxonomy: An Overview". In: *Theory into practice* 41.4 (2002), pp. 212–218. DOI: 10.1207/s15430421tip4104_2.
- [100] Bum Chul Kwon and Bongshin Lee. "A Comparative Evaluation on Online Learning Approaches Using Parallel Coordinate Visualization". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: ACM, 2016, pp. 993–997. ISBN: 9781450333627. DOI: 10 . 1145 / 2858036 . 2858101. URL: <https://doi.org/10.1145/2858036.2858101>.
- [101] Kevin Lai et al. "Measuring Graph Comprehension, Critique, and Construction in Science". In: *Journal of Science Education and Technology* 25 (2016), pp. 665–681. DOI: 10.1007/s10956-016-9621-9.
- [102] Marc Lallanilla. *Misleading Gun-Death Chart Draws Fire*. Apr. 2014. URL: <https://www.livescience.com/45083-misleading-gun-death-chart.html>.
- [103] Judith C. Lapadat. "Ethics in Autoethnography and Collaborative Autoethnography". In: *Qualitative Inquiry* 23.8 (2017), pp. 589–603. DOI: 10 . 1177 / 1077800417704462. eprint: <https://doi.org/10.1177/1077800417704462>. URL: <https://doi.org/10.1177/1077800417704462>.

- [104] Judith C. Lapadat et al. “Life Challenge Memory Work: Using Collaborative Autobiography to Understand Ourselves”. In: *International Journal of Qualitative Methods* 9.1 (2010), pp. 77–104. DOI: 10 . 1177 / 160940691000900108. eprint: <https://doi.org/10.1177/160940691000900108>. URL: <https://doi.org/10.1177/160940691000900108>.
- [105] Amanda Lazar et al. “Adopting Diffractive Reading to Advance HCI Research: A Case Study on Technology for Aging”. In: *ACM Trans. Comput.-Hum. Interact.* 28.5 (2021). ISSN: 1073-0516. DOI: 10 . 1145/3462326. URL: <https://doi.org/10.1145/3462326>.
- [106] Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. “VLAT: Development of a Visualization Literacy Assessment Test”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 551–560. DOI: 10 . 1109/TVCG.2016.2598920.
- [107] Sukwon Lee et al. “The Correlation between Users’ Cognitive Characteristics and Visualization Literacy”. In: *Applied Sciences* 9.3 (2019). ISSN: 2076-3417. DOI: 10 . 3390/app9030488. URL: <http://www.mdpi.com/2076-3417/9/3/488>.
- [108] Stephan Lewandowsky et al. “Misinformation and Its Correction: Continued Influence and Successful Debiasing”. In: *Psychological Science in the Public Interest* 13.3 (2012). PMID: 26173286, pp. 106–131. DOI: 10 . 1177 / 1529100612451018. eprint: <https://doi.org/10.1177/1529100612451018>. URL: <https://doi.org/10.1177/1529100612451018>.
- [109] Maxim Lisnic et al. “Misleading Beyond Visual Tricks: How People Actually Lie with Charts”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10 . 1145/3544548.3580910. URL: <https://doi.org/10.1145/3544548.3580910>.
- [110] Zhengliang Liu, R. Jordan Crouser, and Alvitta Ottley. “Survey on Individual Differences in Visualization”. In: *Computer Graphics Forum* 39.3 (2020), pp. 693–712. DOI: <https://doi.org/10.1111/cgf.14033>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14033>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14033>.
- [111] Samuel A Livingston. “Constructed-Response Test Questions: Why We Use Them; How We Score Them. R&D Connections. Number 11.” In: *Educational Testing Service* (2009).
- [112] David Lloyd and Jason Dykes. “Human-Centered Approaches in Geovisualization Design: Investigating Multiple Methods Through a Long-Term Case Study”. In: *IEEE*

- Trans. Visual Comput. Graphics* 17.12 (2011), pp. 2498–2507. DOI: 10 . 1109 / TVCG . 2011 . 209.
- [113] Leo Yu-Ho Lo et al. “Misinformed by Visualization: What Do We Learn From Misinformative Visualizations?” In: *Computer Graphics Forum* 41.3 (2022), pp. 515–525. DOI: <https://doi.org/10.1111/cgf.14559>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14559>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14559>.
 - [114] Leo Yu-Ho Lo et al. “Why Change My Design: Explaining Poorly Constructed Visualization Designs with Explorable Explanations”. In: *IEEE Transactions on Visualization and Computer Graphics* 30.1 (2024), pp. 955–964. DOI: 10 . 1109 / TVCG . 2023 . 3327155.
 - [115] Angela Locoro, William P. Fisher, and Luca Mari. “Visual Information Literacy: Definition, Construct Modeling and Assessment”. In: *IEEE Access* 9 (2021), pp. 71053–71071. DOI: 10 . 1109 / ACCESS . 2021 . 3078429.
 - [116] Jock Mackinlay. “Automating the design of graphical presentations of relational information”. In: *ACM Trans. Graph.* 5.2 (Apr. 1986), pp. 110–141. ISSN: 0730-0301. DOI: 10.1145/22949.22950. URL: <https://doi.org/10.1145/22949.22950>.
 - [117] Hamid Mansoor and Lane Harrison. “Data Visualization Literacy and Visualization Biases: Cases for Merging Parallel Threads”. In: *Cognitive Biases in Visualizations*. Ed. by Geoffrey Ellis. Cham: Springer International Publishing, 2018, pp. 87–96. ISBN: 978-3-319-95831-6. DOI: 10 . 1007 / 978 - 3 - 319 - 95831 - 6 _ 7. URL: https://doi.org/10.1007/978-3-319-95831-6_7.
 - [118] Morgan McFall-Johnsen. *A ‘cuckoo’ graph with no sense of time or place shows how Georgia bungled coronavirus data as it reopens*. May 2020. URL: <https://www.businessinsider.com/graph-shows-georgia-bungling-coronavirus-data-2020-5>.
 - [119] William J McGuire. “Vaccine for brainwash”. In: *Psychology today* 3.9 (1970), p. 36.
 - [120] Mary L McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia medica* 22.3 (2012), pp. 276–282.
 - [121] Andrew McNutt, Gordon Kindlmann, and Michael Correll. “Surfacing Visualization Mirages”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020,

- pp. 1–16. ISBN: 9781450367080. DOI: 10.1145/3313831.3376420. URL: <https://doi.org/10.1145/3313831.3376420>.
- [122] Dominik Moritz et al. “Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 438–448. DOI: 10.1109/TVCG.2018.2865240.
 - [123] J.C. Nunnally and I.H. Bernstein. *Psychometric Theory*. McGraw-Hill series in psychology no. 972. New York: McGraw-Hill Companies, Incorporated, 1994. ISBN: 9780070478497. URL: <https://books.google.com/books?id=r0fuAAAAAAAJ>.
 - [124] Yasmina Okan et al. “Using the Short Graph Literacy Scale to Predict Precursors of Health Behavior Change”. In: *Medical Decision Making* 39.3 (2019), pp. 183–195. DOI: 10.1177/0272989X19829728. eprint: <https://doi.org/10.1177/0272989X19829728>. URL: <https://doi.org/10.1177/0272989X19829728>.
 - [125] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. “Instructional manipulation checks: Detecting satisficing to increase statistical power”. In: *Journal of Experimental Social Psychology* 45.4 (2009), pp. 867–872. ISSN: 0022-1031. DOI: <https://doi.org/10.1016/j.jesp.2009.03.009>. URL: <https://www.sciencedirect.com/science/article/pii/S0022103109000766>.
 - [126] Jasmine Otto. *Visualization Notebooks are Paper Tools*. <https://osf.io/z2u3w>. 2024.
 - [127] Jasmine T. Otto and Scott Davidoff. “Visualization Artifacts are Boundary Objects”. In: *ProcBELIV*. 2024, pp. 81–88. DOI: 10.1109/BELIV64461.2024.00014.
 - [128] Anshul Vikram Pandey et al. “How Deceptive Are Deceptive Visualizations? An Empirical Analysis of Common Distortion Techniques”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI ’15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, pp. 1469–1478. ISBN: 9781450331456. DOI: 10.1145/2702123.2702608. URL: <https://doi.org/10.1145/2702123.2702608>.
 - [129] Saugat Pandey and Alvitta Ottley. “Benchmarking Visual Language Models on Standardized Visualization Literacy Tests”. In: *Computer Graphics Forum* 44.3 (2025), e70137. DOI: <https://doi.org/10.1111/cgf.70137>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.70137>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.70137>.
 - [130] Saugat Pandey and Alvitta Ottley. “Mini-VLAT: A Short and Effective Measure of Visualization Literacy”. In: *Computer Graphics Forum* 42.3 (2023), pp. 1–11. DOI: <https://doi.org/10.1111/cgf.70137>.

//doi.org/10.1111/cgf.14809. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14809>.

- [131] Denise F. Polit, Cheryl Tatano Beck, and Steven V. Owen. “Is the CVI an acceptable indicator of content validity? Appraisal and recommendations”. In: *Research in Nursing & Health* 30.4 (2007), pp. 459–467. DOI: <https://doi.org/10.1002/nur.20199>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/nur.20199>.
- [132] Fateme Rajabiyazdi et al. *EduVis: 2nd IEEE VIS Workshop on Visualization Education, Literacy, and Activities*. <https://ieee-eduvis.github.io/IEEE-EduVis-Workshop-2024.pdf>.
- [133] Claudia Ramly et al. “Digitally training graph viewers against misleading bar charts”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. 2021.
- [134] Robyn L. Raschke and Paul John Steinbart. “Mitigating the Effects of Misleading Graphs on Decisions by Educating Users about the Principles of Graph Design”. In: *Journal of Information Systems* 22.2 (Sept. 2008), pp. 23–52. ISSN: 0888-7985. DOI: 10.2308/jis.2008.22.2.23. URL: <https://doi.org/10.2308/jis.2008.22.2.23>.
- [135] William Revelle and David M. Condon. “Reliability from α to ω : A tutorial.” In: *Psychological Assessment* 31.12 (2019). Place: US Publisher: American Psychological Association, pp. 1395–1411. ISSN: 1939-134X(Electronic),1040-3590(Print). DOI: 10.1037/pas0000754.
- [136] Jihyun Rho and Martina A Rau. “Exploring educational approaches to addressing misleading visualizations”. In: *Educational Psychology Review* 37.1 (2025), p. 14. DOI: 10.1007/s10648-025-09988-0.
- [137] Horst WJ Rittel and Melvin M Webber. “Dilemmas in a general theory of planning”. In: *Policy sciences* 4.2 (1973), pp. 155–169. DOI: 10.1007/BF01405730.
- [138] George Robertson et al. “Effectiveness of Animation in Trend Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008), pp. 1325–1332. DOI: 10.1109/TVCG.2008.125.
- [139] Rickman Roedavan et al. “Serious Game Development Model Based on the Game-Based Learning Foundation.” In: *Journal of ICT Research & Applications* 15.3 (2021).
- [140] Jon Roozenbeek and Sander van der Linden. “Breaking Harmony Square: A game that “inoculates” against political misinformation”. In: *The Harvard Kennedy School Misinformation Review* (2020).

- [141] Jon Roozenbeek and Sander Van der Linden. “Fake news game confers psychological resistance against online misinformation”. In: *Palgrave Communications* 5.1 (2019), pp. 1–10. DOI: 10.1057/s41599-019-0279-9.
- [142] Puripant Ruchikachorn and Klaus Mueller. “Learning Visualizations by Analogy: Promoting Visual Literacy through Visualization Morphing”. In: *IEEE Trans Vis Comput Graph* 21.9 (2015), pp. 1028–1044. DOI: 10.1109/TVCG.2015.2413786.
- [143] William H. Ryan and Ellen R. K. Evers. “Graphs with logarithmic axes distort lay judgments”. In: *Behavioral Science & Policy* 6.2 (2020), pp. 13–23. DOI: 10.1177/237946152000600203. eprint: <https://doi.org/10.1177/237946152000600203>. URL: <https://doi.org/10.1177/237946152000600203>.
- [144] Marc Satkowski et al. “Who benefits from Visualization Adaptations? Towards a better Understanding of the Influence of Visualization Literacy”. In: *2022 IEEE Visualization and Visual Analytics (VIS)*. 2022, pp. 90–94. DOI: 10.1109/VIS54862.2022.00027.
- [145] Arvind Satyanarayan et al. “Vega-Lite: A Grammar of Interactive Graphics”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 341–350. DOI: 10.1109/TVCG.2016.2599030.
- [146] Edward Segel and Jeffrey Heer. “Narrative Visualization: Telling Stories with Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 1139–1148. DOI: 10.1109/TVCG.2010.179.
- [147] Mara Solen. *Scoping the Future of Visualization Literacy: A Review*. <https://osf.io/eypgm>. 2022.
- [148] Chase Stokes, Cindy Xiong Bearfield, and Marti A. Hearst. “The Role of Text in Visualizations: How Annotations Shape Perceptions of Bias and Influence Predictions”. In: *IEEE Transactions on Visualization and Computer Graphics* 30.10 (2024), pp. 6787–6800. DOI: 10.1109/TVCG.2023.3338451.
- [149] Li Qian Tay et al. “A comparison of prebunking and debunking interventions for implied versus explicit misinformation”. In: *British Journal of Psychology* 113.3 (2022), pp. 591–607. DOI: <https://doi.org/10.1111/bjop.12551>. eprint: <https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/bjop.12551>. URL: <https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/bjop.12551>.
- [150] Randall Teal. “Developing a (Non-linear) Practice of Design Thinking”. In: *International Journal of Art & Design Education* 29.3 (2010), pp. 294–302. DOI: <https://doi.org/10.1111/j.1469-7580.2010.00411.x>.

- org/10.1111/j.1476-8070.2010.01663.x. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1476-8070.2010.01663.x>.
- [151] The New York Times. *Graphics - The New York Times*. 2022. URL: <https://www.nytimes.com/spotlight/graphics>.
 - [152] Matthew Tobiasz, Petra Isenberg, and Sheelagh Carpendale. “Lark: Coordinating Co-located Collaboration with Information Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 1065–1072. DOI: 10.1109/TVCG.2009.162.
 - [153] Jonathan Tonglet et al. “Protecting multimodal large language models against misleading visualizations”. In: *arXiv preprint arXiv:2502.20503* (2025).
 - [154] Eliane R. A. Valiati, Marcelo S. Pimenta, and Carla M. D. S. Freitas. “A Taxonomy of Tasks for Guiding the Evaluation of Multidimensional Visualizations”. In: *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*. BELIV ’06. Venice, Italy: Association for Computing Machinery, 2006, pp. 1–6. ISBN: 1595935622. DOI: 10.1145/1168149.1168169. URL: <https://doi.org/10.1145/1168149.1168169>.
 - [155] Arnav Verma and Judith E Fan. “Measuring and predicting variation in the difficulty of questions about data visualizations”. In: *arXiv preprint arXiv:2505.08031* (2025).
 - [156] Fernanda B. Viegas et al. “ManyEyes: a Site for Visualization at Internet Scale”. In: *IEEE Transactions on Visualization and Computer Graphics* 13.6 (2007), pp. 1121–1128. DOI: 10.1109/TVCG.2007.70577.
 - [157] *Workshop on Visualization Play, Games, and Activities*. <https://visgames2025.netlify.app/>. 2025.
 - [158] VisLies. *VisLies 2020*. 2020. URL: <http://www.vislies.org/2020/>.
 - [159] VisLies. *VisLies 2021*. 2021. URL: <https://www.vislies.org/2021/>.
 - [160] VisLies. *Vis Lies 2017 Gallery*. 2017. URL: <https://www.vislies.org/2017/gallery/>.
 - [161] Washington Post. *Washington Post | FlowingData*. 2022. URL: <https://flowingdata.com/tag/washington-post/>.

- [162] Zhen Wen and Michelle Zhou. “Evaluating the Use of Data Transformation for Information Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008), pp. 1309–1316. DOI: 10.1109/TVCG.2008.129.
- [163] Leland Wilkinson. *The grammar of graphics*. New York, NY: Springer, 2012. DOI: <https://doi.org/10.1007/0-387-28695-0>.
- [164] Aoyu Wu et al. “MultiVision: Designing Analytical Dashboards with Deep Learning Based Recommendation”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2022), pp. 162–172. DOI: 10.1109/TVCG.2021.3114826.
- [165] Brenda W. Yang et al. “Truncating Bar Graphs Persistently Misleads Viewers”. In: *Journal of Applied Research in Memory and Cognition* 10.2 (2021), pp. 298–311. ISSN: 2211-3681. DOI: <https://doi.org/10.1016/j.jarmac.2020.10.002>. URL: <https://www.sciencedirect.com/science/article/pii/S2211368120300978>.
- [166] Junran Yang et al. “Draco 2: An Extensible Platform to Model Visualization Design”. In: *2023 IEEE Visualization and Visual Analytics (VIS)*. Melbourne, Australia: IEEE, 2023, pp. 166–170. DOI: 10.1109/VIS54172.2023.00042.
- [167] Caroline Ziemkiewicz and Robert Kosara. “The Shaping of Information by Visual Metaphors”. In: *IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008), pp. 1269–1276. DOI: 10.1109/TVCG.2008.171.