CALVI: Critical Thinking Assessment for Literacy in Visualizations

Lily W. Ge
Northwestern University
Evanston, Illinois, USA
wanqian.ge@northwestern.edu

Yuan Cui
Northwestern University
Evanston, Illinois, USA
yuancui2025@u.northwestern.edu

Matthew Kay
Northwestern University
Evanston, Illinois, USA
mjskay@northwestern.edu

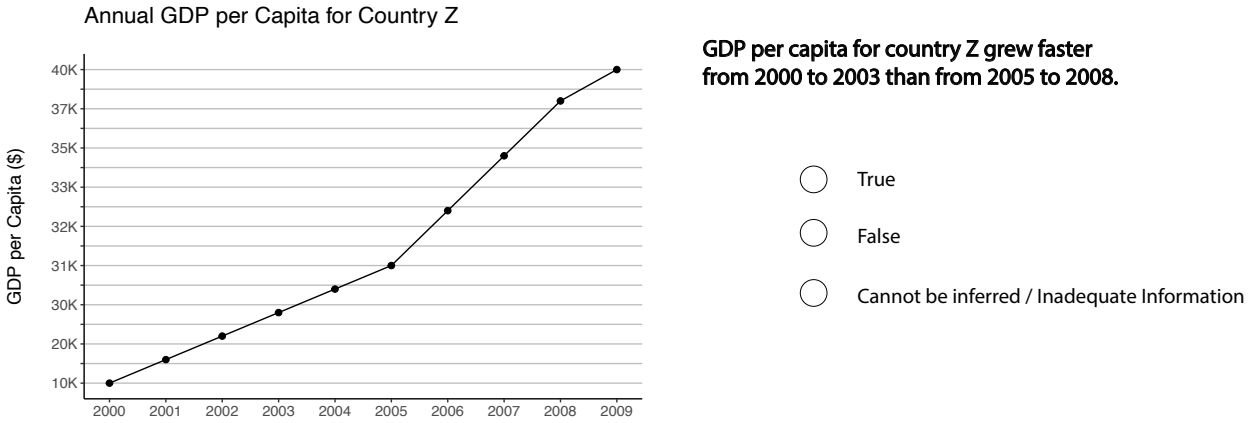


Figure 1: An example question from CALVI with *Manipulation of Scales - Inappropriate Use of Scale Functions* (misleader) on a line chart. At first glance, the answer may seem to be False. However, as shown on the y-axis, GDP per capita increased about 20K from 2000 to 2003, while it only increased about 6K from 2005 to 2008. The inconsistent scale (number labels on the axis do not match the tick mark spacings) on the y-axis gives a misleading visual impression.

ABSTRACT

Visualization misinformation is a prevalent problem, and combating it requires understanding people's *ability to read, interpret, and reason about erroneous or potentially misleading visualizations*, which lacks a reliable measurement: existing visualization literacy tests focus on well-formed visualizations. We systematically develop an assessment for 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. Our final bank of 45 items shows high reliability, and we provide item bank usage recommendations for future tests and different use cases. Related materials are available at: https://osf.io/pv67z/.[1]

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; **Empirical studies in visualization**; **Information visualization**.

KEYWORDS

Information visualization, Visualization literacy, Visualization misinformation, Measurement, Psychometrics

ACM Reference Format:
Lily W. Ge, Yuan Cui, and Matthew Kay. 2023. *CALVI*: Critical Thinking Assessment for Literacy in Visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581406

1 INTRODUCTION

Visualizations are constantly in the public eye: for example, news and media outlets often use visualizations to illustrate and support

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581406

[1]This is the authors' version of the work. It is posted here for your personal use. Not for redistribution. The definitive version will be published in ACM CHI 2023, https://doi.org/10.1145/3544548.3581406.

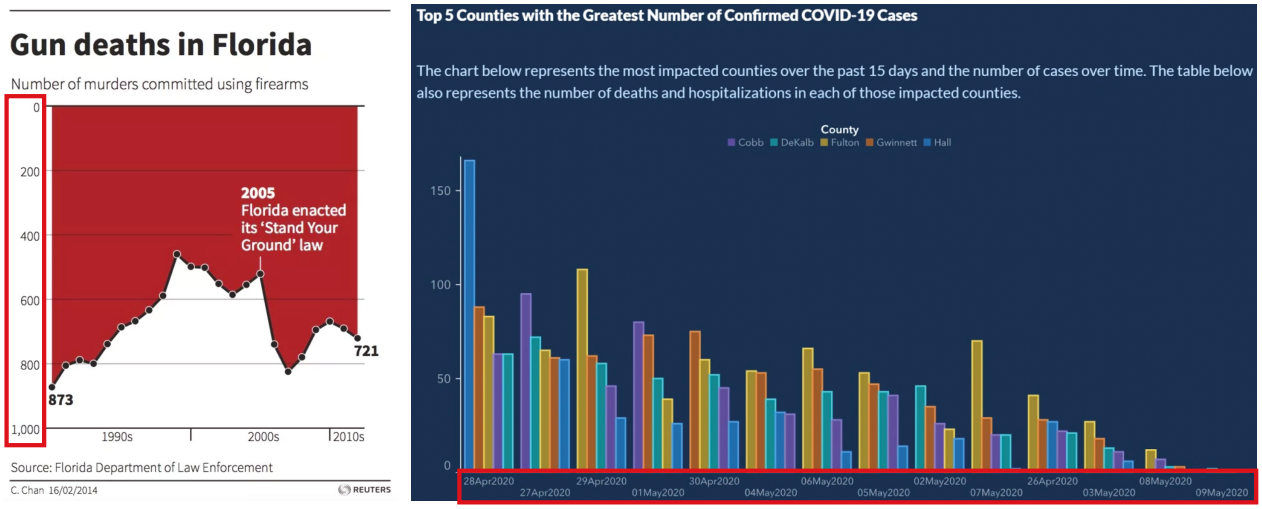


Figure 2: Real-world examples of visualization misinformation. The misleader in these examples is *Manipulation of Scales - Unconventional Scale Directions*. (A) The inverted y-axis gives a visual impression that gun deaths went down drastically after the "Stand Your Ground" law in 2005, when they actually increased [28]. (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 [31].

their claims [4, 36, 38]. 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 2.A) [28]. 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 2.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 [31]. Unfortunately, there is an abundance of real-world examples of this kind of *visualization misinformation*, which occurs when charts send misleading messages to the audience.[2] 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 [7, 29] 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 identify 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 [13], which we consider a form of visualization misinformation.[3] However, they considered only one type of deception (inappropriate y-axis range) [13]. 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.

[2]Many such examples have been discussed at the VisLies meetups held in conjunction with the *IEEE VIS* conference [40–42].

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

2 RELATED WORK

2.1 Visualization Misinformation

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 [22]. Others have studied misalignment between a visualization and its title and how this impacts trust and recall of information [27]. 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) [26]. 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 [32]. 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 [32]. More recently, Lo et al. developed a taxonomy of misinformative visualizations that included errors or issues that could be present on poorly designed visualizations [30]. 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 first step in developing effective interventions to combat visualization misinformation.

2.2 Visualization Literacy

Previously, researchers have studied visualization literacy under varying definitions. Boy et al. focused on line charts, bar charts, and scatterplots and defined visualization literacy to be "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" [7]. Börner et al. designed a study specifically looking at the visualization literacy of a target audience that is also interested in science and museums [10]. Although this audience group encounters visualizations in their everyday lives and have a higher interest in math and science, Börner et al. found that most of the participants still struggled to interpret the visualizations presented in the study [10]. Differing from Boy et al., Börner et al. defined visualization literacy to be "the ability to make meaning from and interpret patterns, trends, and correlations in visual representations of data" [10]. Their study results also showed an urgent need in the educational space to better teach students about visualizations. In a later work, Börner et al. proposed a framework aiming to assess visualization literacy and teach visualizations; the ability to construct visualizations was also included in their definition of visualization literacy [9]. In order to provide a more tailored instruction on visualizations, it is important to better understand the students' current abilities in visualization interpretation. Lee et al. developed a Visualization Literacy Assessment Test (VLAT) that measures people's visualization literacy, which they defined as "the ability and skill to read and interpret visually represented data in and to extract information from data visualizations" [29]. However, every visualization in VLAT is assumed to be correct, and we argue that this assumption does not transfer to the real world because the public will likely encounter erroneous visualizations as well.

As discussed above, not every visualization is guaranteed to be correctly or effectively designed. Thus, only measuring visualization literacy with correctly designed visualizations is not comprehensive enough. To be able to identify visualization misinformation, one needs the ability to *critically* think about what is given, which could be referred to as the ability to judge the accuracy of the information presented [21]. This aspect related to critical thinking has been studied in other domains such as statistical literacy [23, 43], where researchers have also looked at graph interpretation related to statistical literacy, some specifically targeting the critical aspect [2, 3, 33]. However, this critical thinking aspect has not been studied extensively in the context of visualization literacy. Camba et al., focusing on misleading or deceptive visualizations in the education space, identified deception recognition as an important part of visualization literacy [13], but studied only one type of deception (i.e., inappropriate y-axis range) out of the many ways a chart can mislead. This presents an opportunity to put the existing taxonomies of misleading visualizations we discussed in Section 2.1 into practice to more comprehensively understand people's ability to identify visualization misinformation.

2.3 Systematic Procedure for Test Development

We will follow the procedure outlined in Psychological Testing and Assessment [14] to systematically develop our test. Here, we describe each phase 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 [14]. 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 [14]. 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 [14]. 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. [29] used CTT; Boy et al. [7] used IRT). Depending on the IRT model, different samples sizes are recommended [19]. We opted for IRT because it can be extended further to develop computer adaptive testing (CAT) [5, 19]. 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 [14].

3 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 2, 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.

3.1 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 [24]. 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 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).

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. [30, 32]. 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 of misleader as they are not part of the visualization construction process[4]), and further categorized them into higher-level or lower-level categories (Ⓐ in Figure 3). The same process is repeated with categorizations from Lo et al.[5] 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. (Ⓑ in Figure 3). 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 (Ⓒ in Figure 3).

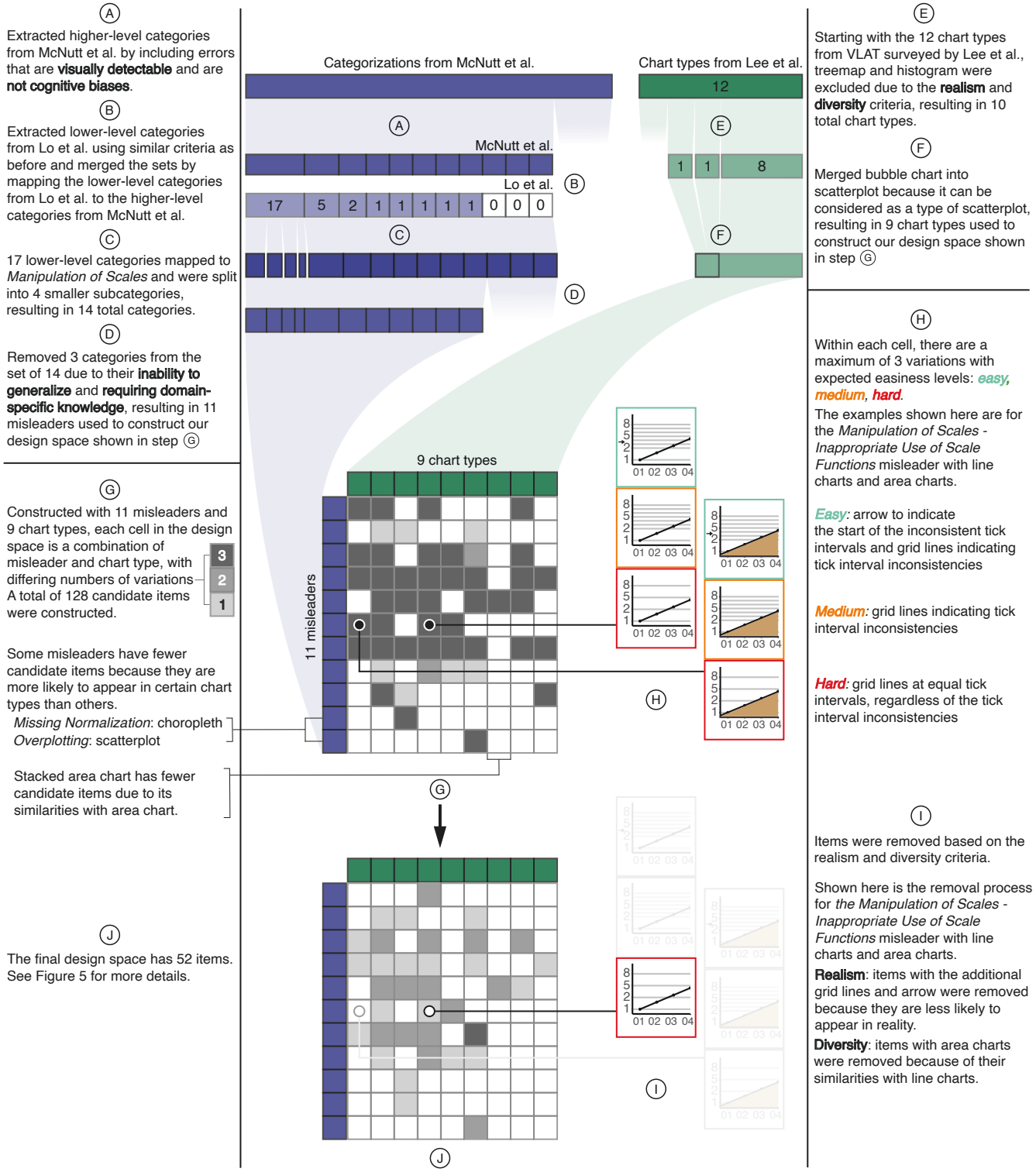
In order to systemically apply the misleaders to different chart types, the misleaders have to be generalizable across chart types. Thus, misleaders with an **inability to generalize** cannot be applied to a variety of chart types to populate the design space. Additionally, the items need to be self-contained, so they **cannot require domain-specific knowledge**. Three high-level categories from McNutt et al. [32] were removed from the set of 14 based on these two criteria (Ⓓ in Figure 3). 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 1.

Chart Types. We started with the 12 chart types from VLAT surveyed by Lee et al. [29] (Ⓔ in Figure 3). 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. [29]. 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 (Ⓕ in Figure 3). 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

[4]However, 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.

[5]Details from the derivation process can be found in supplemental materials.



Ⓐ
Extracted higher-level categories from McNutt et al. by including errors that are **visually detectable** and are **not cognitive biases**.

Ⓑ
Extracted lower-level categories from Lo et al. using similar criteria as before and merged the sets by mapping the lower-level categories from Lo et al. to the higher-level categories from McNutt et al.

Ⓒ
17 lower-level categories mapped to *Manipulation of Scales* and were split into 4 smaller subcategories, resulting in 14 total categories.

Ⓓ
Removed 3 categories from the set of 14 due to their **inability to generalize** and **requiring domain-specific knowledge**, resulting in 11 misleaders used to construct our design space shown in step Ⓖ

Ⓔ
Starting with the 12 chart types from VLAT surveyed by Lee et al., treemap and histogram were excluded due to the **realism** and **diversity** criteria, resulting in 10 total chart types.

Ⓕ
Merged bubble chart into scatterplot because it can be considered as a type of scatterplot, resulting in 9 chart types used to construct our design space shown in step Ⓖ

Ⓖ
Constructed with 11 misleaders and 9 chart types, each cell in the design space is a combination of misleader and chart type, with differing numbers of variations A total of 128 candidate items were constructed.

Some misleaders have fewer candidate items because they are more likely to appear in certain chart types than others.
Missing Normalization: choropleth
Overplotting: scatterplot

Stacked area chart has fewer candidate items due to its similarities with area chart.

Ⓗ
Within each cell, there are a maximum of 3 variations with expected easiness levels: ***easy***, ***medium***, ***hard***.
The examples shown here are for the *Manipulation of Scales - Inappropriate Use of Scale Functions* misleader with line charts and area charts.

Easy: arrow to indicate the start of the inconsistent tick intervals and grid lines indicating tick interval inconsistencies

Medium: grid lines indicating tick interval inconsistencies

Hard: grid lines at equal tick intervals, regardless of the tick interval inconsistencies

Ⓘ
Items were removed based on the realism and diversity criteria.

Shown here is the removal process for *the Manipulation of Scales - Inappropriate Use of Scale Functions* misleader with line charts and area charts.
Realism: items with the additional grid lines and arrow were removed because they are less likely to appear in reality.
Diversity: items with area charts were removed because of their similarities with line charts.

Ⓙ
The final design space has 52 items. See Figure 5 for more details.

Figure 3: 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.

Table 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. [30, 32, 40]
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. [32, 40]
Inappropriate Aggregation	Aggregating data in an improper way that leads to inaccurate conclusions. [30, 32, 42]
Manipulation of Scales - Inappropriate Order	The axis labels or legends appear to be in a random order due to manipulation of data ordering. [30]
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. [15, 16, 30, 32, 34, 41, 42]
Manipulation of Scales - Inappropriate Use of Scale Functions	Applying arbitrary non-linear functions to scales. [32]
Manipulation of Scales - Unconventional Scale Directions	The direction of scales of axes or legends is created against convention such as inverting axes or scales. [32, 34]
Misleading Annotations	Annotations that contradict or make it harder to read the visualization. [27]
Missing Data	A visual representation implies data exist but the data is actually missing. [25, 32]
Missing Normalization	Displaying unnormalized data in absolute quantity when normalized data in relative quantity is of interest. [17, 30, 32, 42]
Overplotting	Displaying too many things on a plot can obscure parts of the data. [30, 32]

types as columns (Ⓖ in Figure 3).[6] 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 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). This is only our attempt to have items with varying levels of easiness, a property of items ultimately measured by IRT (see Section 5). We show another example alteration in Figure 4, 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.

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). 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 in reality for the *Manipulation of Scales - Inappropriate Use of Scale Functions* misleader, so such items were removed (Ⓘ in Figure 3). Per the diversity criterion, we removed item types that are redundant. One

[6]The initial design space can be found in supplemental materials.

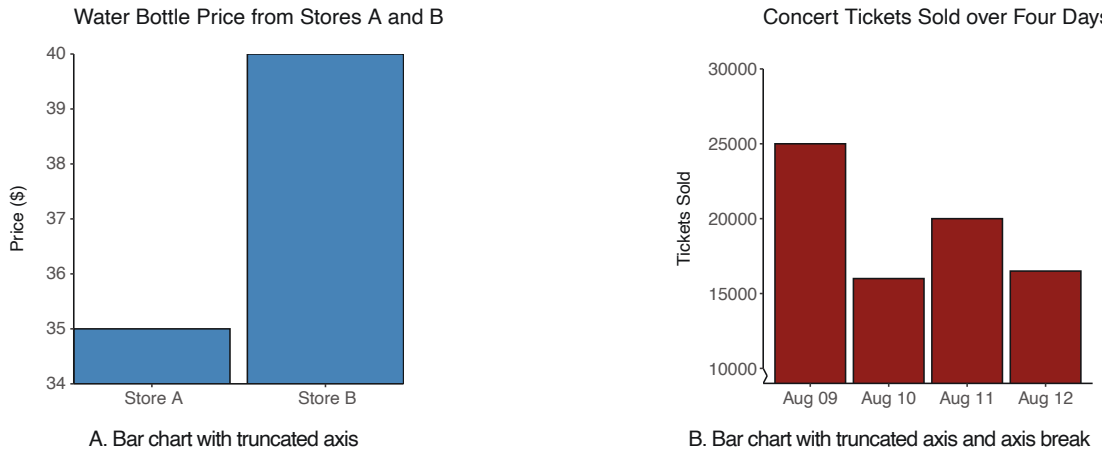


Figure 4: Two variations of the same item type: bar chart with misleader *Manipulation of Scales - Inappropriate Scale Range*.

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 (Ⓙ in Figure 3). An overview of our final design space is in Figure 5.

3.3 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 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

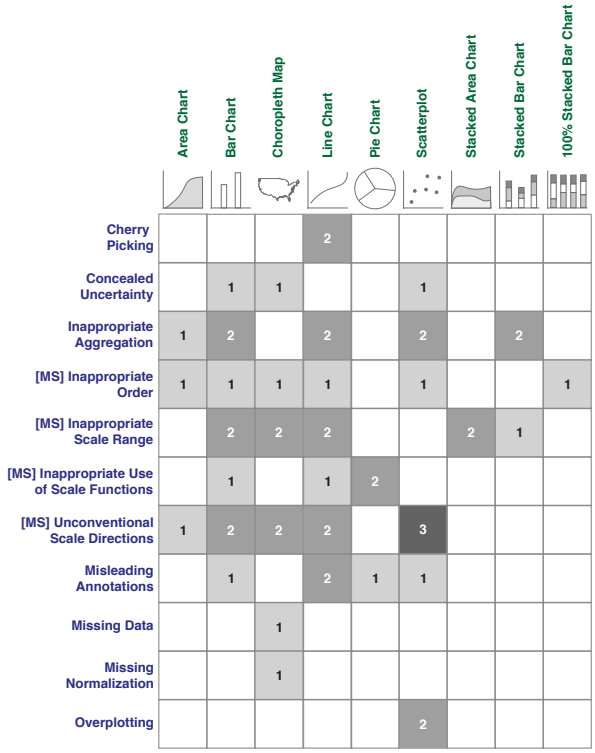


Figure 5: The design space with 11 misleaders and 9 chart types. The numbers in the cells indicate how many variations are associated with the combination of misleader and chart type. Four misleaders directly relate to the manipulation of scales (MS).

aggregate values, all of which are visualization tasks rooted in literature [1, 8, 39].[7] The first four tasks were taken from VLAT [29],

[7]The specific task(s) of each item in the bank can be found in supplemental materials.

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 9.5.

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 6. 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 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 6).

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.

3.4 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.

3.5 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 tryout 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 participants[8] 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.[9] 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.

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

[8] A sample size greater than common standards at CHI [12].
[9] The consent form, instructions, test items, and open-ended questions can be found in supplemental materials.

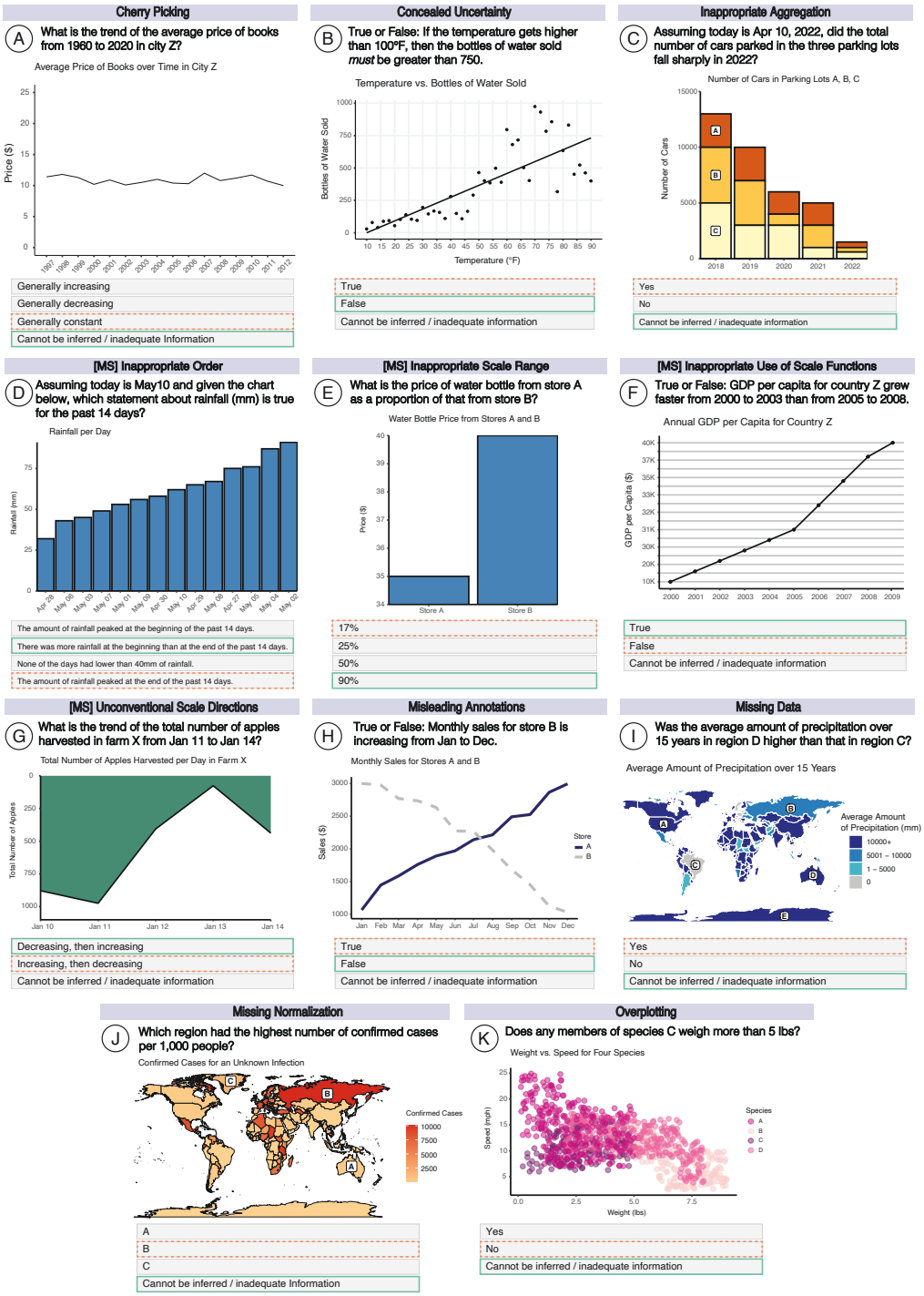


Figure 6: Representative items of the 11 misleaders from the bank. Correct answers are indicated with solid outlines, and wrong-due-to-misleader answers are indicated with dashed outlines. Arriving at the correct answers generally requires the test taker to reason about and reflect on the construction 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.

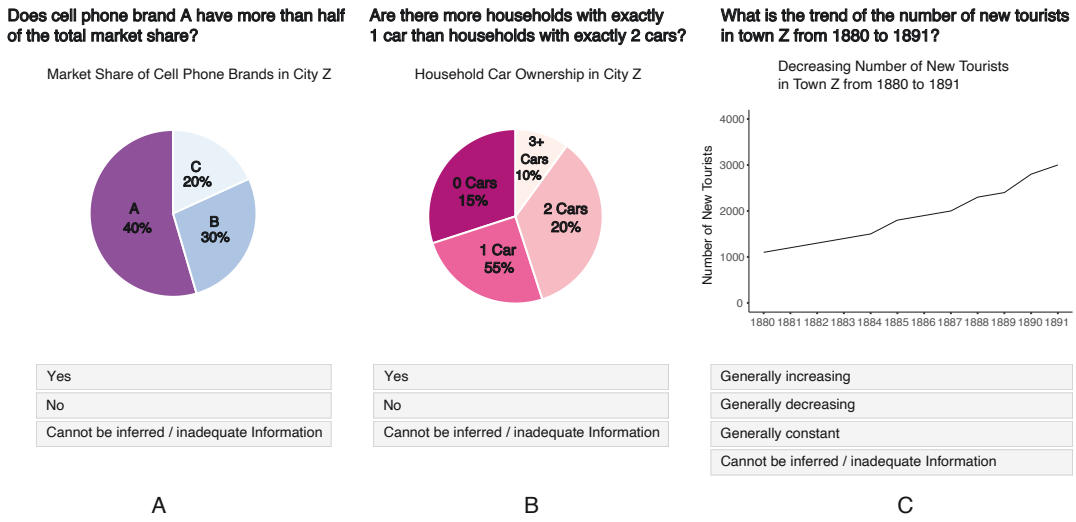


Figure 7: The set of items the participants may receive with visualizations that contain conflicting information and have a corresponding question in the open-ended section during test tryout (see Supplemental Materials). In A and B, the percentage labels do not match the sizes of the pie slices. In C, the title says decreasing trend while the trend of the line is actually increasing.

visualizations when there is conflicting information present, so we decided to implement 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 7).

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 distinct 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 [19]. We fit the simulated 500 participants to our preregistered model (details in Section 5.2) 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.[10] Thus, we chose 500 as our sample size for test tryout.

4 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 (Section 3.5), the open-ended section includes questions for three items that contained conflicting information (shown in Figure 7) to further understand why some preferred one choice over the other.

[10]The script for this simulation and model check analysis can be found in supplemental materials.

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.

4.1 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 materials) ranged from 4.62 to 89.60 (*M = 19.80, SD = 10.25*). This suggests that 30 minutes is reasonable to complete the test.

5 ITEM ANALYSIS

5.1 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 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 1 is often referred to as the item response function, and the curve of this function is called the item characteristic curve (ICC).

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

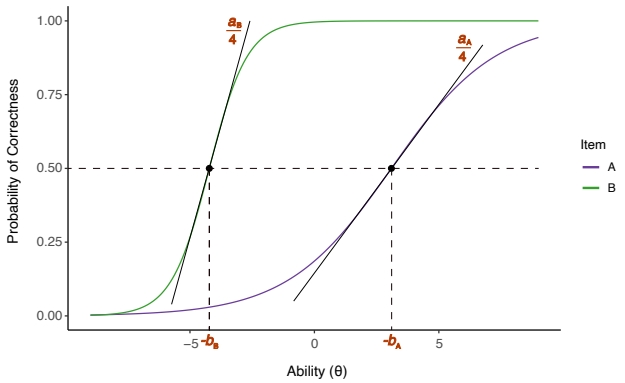


Figure 8: ICC curves for two example items, A and B. The θ values corresponding to the two points on the curves are the difficulty values (negation of easiness) for those two items, and the slopes of the curves at those points are proportional to the discrimination values of the items. In this example, item A is more difficult than item B, while item B is more discriminating than item A.

Item Easiness. To understand item easiness, consider the case where $\theta = -b_i$. Then by Equation 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 8 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 1 is $\frac{a_i}{4}$ at $\theta = -b_i$. For instance, Figure 8 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.

5.2 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 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 [11].

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 [11], which was preregistered on OSF (see https://osf.io/pv67z/).

5.2.1 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 [11]. 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 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 Section 3.5). 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.

5.3 Results

Figure 10 contains the analysis results of each item, including the item easiness and discrimination estimates, correctness, $rate_{wm}$ (see Section 7.1), $chance_{wm}$ (see Section 7.1), and the content validity index (CVI) (see Section 7.2).

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 10 shows the coefficient plots displaying the median easiness and discrimination of each item with 95% and 66% credible intervals (CI).

In Figure 9, we show the average item easiness and discrimination estimates for each misleader. *Overplotting*, *Missing Data*, and *Missing Normalization* are the most 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.

6 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 5.3 and 1 item based on the responses to the open-ended questions. Below, we explain our reasons for these revisions.

6.1 Revision from IRT Analysis

Results from Section 5.3 were used to revise the items in the item bank. The revision process is necesssarily holistic with many factors to balance [14], 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

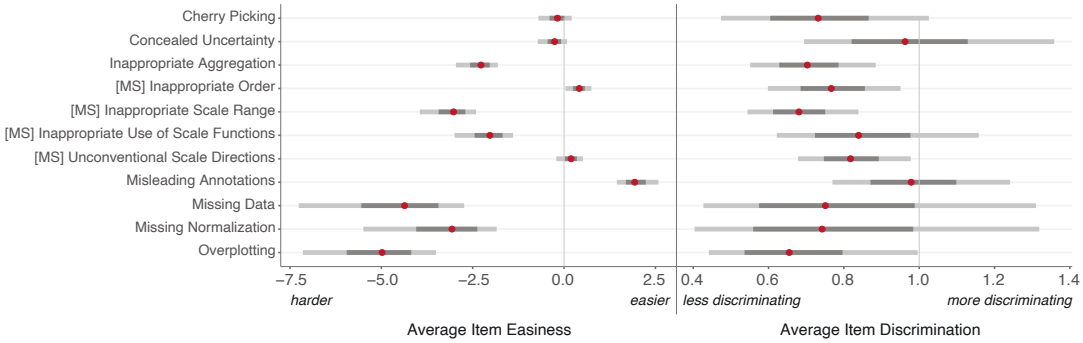


Figure 9: Average item easiness and discrimination estimates for each misleader. The dots represent the median average easiness/discrimination estimates, the lighter bar represents the 95% CI, and the darker bar represents the 66% CI.

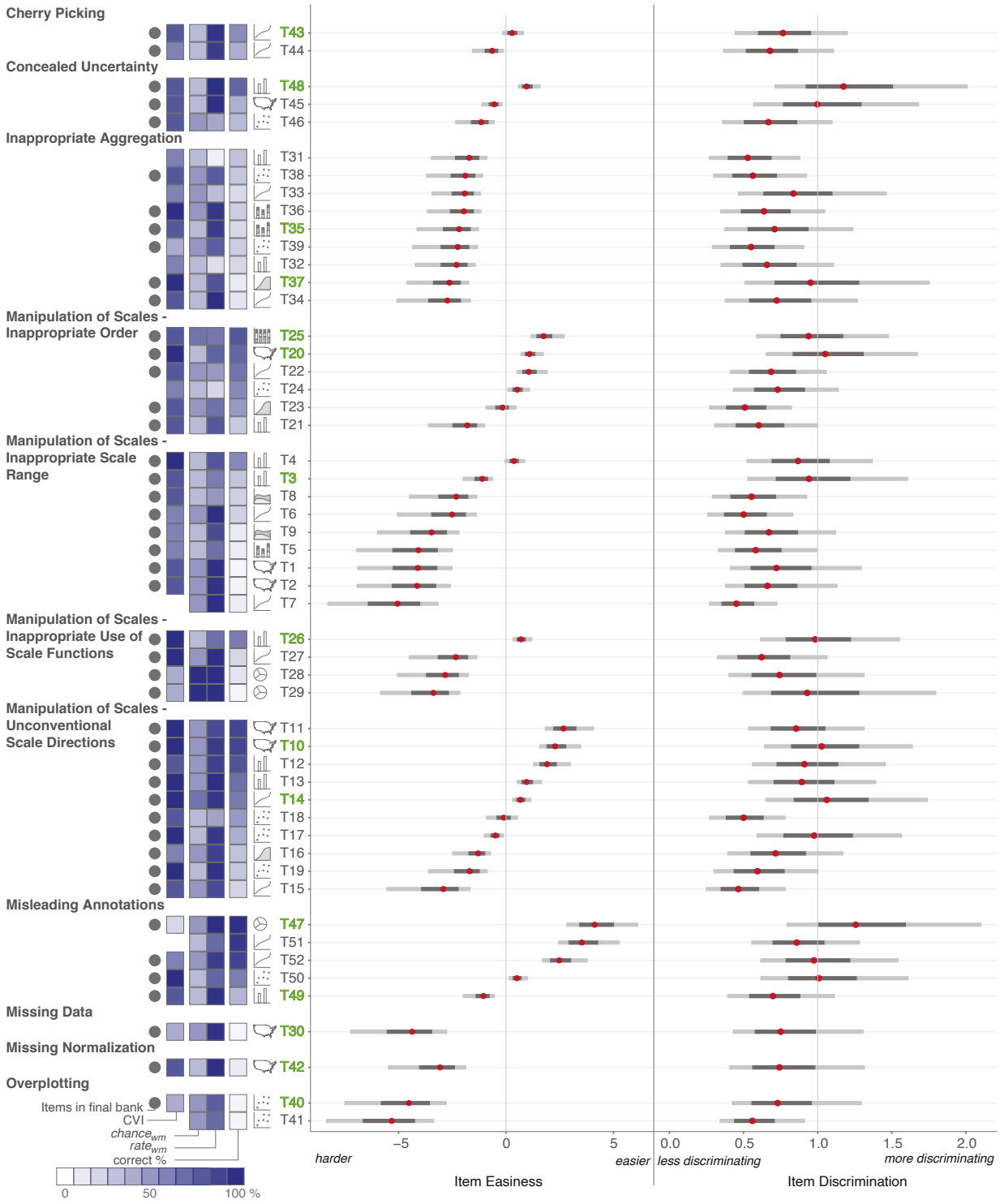


Figure 10: Item easiness and discrimination for all of the items in our item bank during test tryout. For each item, the dot represents the median estimate of its easiness/discrimination parameter, the lighter bar represents the 95% CI, and the darker bar represents the 66% CI. Additionally, shown on the left of the item IDs are chart type, correctness, $rate_{wm}$ (see Section 7.1), $chance_{wm}$ (see Section 7.1), CVI (see Section 7.2), 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 9.1).

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.

6.2 Revision from Qualitative Analysis of Open-ended Questions

As described in Section 3.5, 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 7.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 7.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 7.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 [27]. 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).

7 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 [6], 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) [35] for each item.

7.1 Wrong-due-to-misleader Score

We constructed (Section 3.2) and wrote the items (Section 3.3) 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 6). 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 3.5): 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.

7.2 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 [35]. We invited five domain experts to rate the 49 items in our bank on a scale of 1 (not relevant) to 4 (highly relevant) [35]. 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) [35].

7.3 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 6.2. 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 6.2, 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 10).

8 FINALIZED BANK AND RELIABILITY

8.1 Finalized Bank of Items

The finalized bank of items spans a wide range of difficulty and discrimination (as shown in Figure 11 (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 (see Section 9.1.2).[11]

As we mentioned in Section 2.3, 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 [14]. 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 research emerge in the field [14].

8.2 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 [18]. However, Dunn et al. pointed out some difficulties with using α and outlined advantages for using ω [20]. 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 [37]. 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 ($\omega_t = 0.81$).

9 DISCUSSION

9.1 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 10); 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 11 (middle), and compared to the ICC curves of all 45 items in Figure 11 (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 ($\omega_t = 0.82$).

9.1.1 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.

9.1.2 Customizing Future Tests. Although the target test taker population is the general public for our study, it could potentially be

[11] The item bank can be found in supplemental materials.

[12] A scatterplot of raw score vs. ability and the correlation coefficient can be found in supplemental materials.

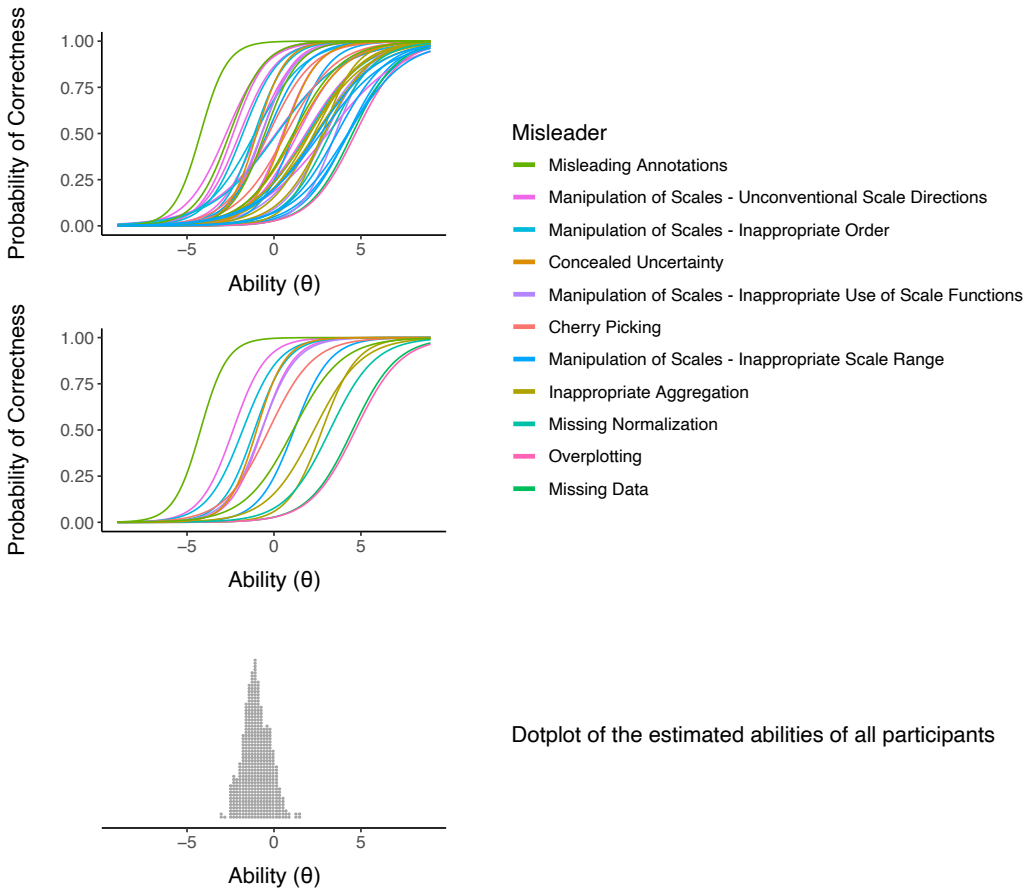


Figure 11: ICC curves for all 45 items (top) in the bank show that the bank covers a wide range of abilities. The 15 items (middle) we selected also cover a similarly wide range of abilities, which can be used to test a general audience. The ICC curves are colored by the 11 misleaders. The dotplot (bottom)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

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, 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.

9.2 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 9.1).

9.3 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 6.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 2.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.

9.4 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?

9.5 Are All Misleaders Actually Misleading?

As we explained in Section 3.2, 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 [15]; 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.

10 CONCLUSION

In this work, we proposed a definition of visualization literacy that incorporates the critical interpretation of visualizations, which is a crucial skill to have to effectively navigate around potentially misleading visualizations in everyday media consumption. Drawing upon prior research in visualization misinformation, we developed a design space for misleading visualizations containing 11 misleaders and 9 chart types, then systematically developed a test to assess people's critical thinking ability in the face of potentially misleading visualizations. We finalized a bank of 45 items with their item easiness and item discrimination parameters, which can be used to tailor future tests. CALVI is a first step in a series of many toward a more comprehensive understanding of visualization misinformation, and more broadly, the different aspects of visualization literacy that extend beyond the core ability to read and extract information from visualizations.

ACKNOWLEDGMENTS

We thank Steven Franconeri for early feedback, and many thanks to Elizabeth Tipton and William Revelle for their practical advice during different phases of the project. Special thanks to Hyeok Kim, Fumeng Yang, Abhraneel Sarma, Xiaoying Pu, and members of the MU Collective at Northwestern for their valuable feedback. We also greatly appreciate the participants and domain experts for their time, as well as the anonymous reviewers for their helpful comments. This work was supported by a grant from the National Science Foundation (#1815790).

REFERENCES

[1] R. Amar, J. Eagan, and J. Stasko. 2005. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE, Minneapolis, MN, USA, 111–117. https://doi.org/10.1109/INFVIS.2005.1532136

[2] Kazuhiro Aoyama. 2007. Investigating a hierarchy of students' interpretations of graphs. *International Electronic Journal of Mathematics Education* 2, 3 (2007), 298–318.

[3] Kazuhiro Aoyama and Max Stephens. 2003. Graph interpretation aspects of statistical literacy: A Japanese perspective. *Mathematics Education Research Journal* 15, 3 (2003), 207–225.

[4] Ryan Best, Elena Mejía, Jasmine Mithani, Anna Wiederkehr, Julia Wolfe, and Yutong Yuan. 2020. The 40 Weirdest (And Best) Charts We Made In This Long, Strange Year. Retrieved August 26, 2022 from https://fivethirtyeight.com/features/the-40-weirdest-and-best-charts-we-made-in-2020/

[5] R Darrell Bock and Robert D Gibbons. 2021. *Item response theory*. John Wiley & Sons, Hoboken, NJ.

[6] Denny Borsboom, Gideon J Mellenbergh, and Jaap Van Heerden. 2004. The concept of validity. *Psychological review* 111, 4 (2004), 1061.

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

[8] Matthew Brehmer and Tamara Munzner. 2013. A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2376–2385. https://doi.org/10.1109/TVCG.2013.124

[9] Katy Börner, Andreas Bueckle, and Michael Ginda. 2019. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1857–1864. https://doi.org/10.1073/pnas.1807180116 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1807180116

[10] Katy Börner, Adam Maltese, Russell Nelson Balliet, and Joe Heimlich. 2016. Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization* 15, 3 (2016), 198–213. https://doi.org/10.1177/1473871615594652 arXiv:https://doi.org/10.1177/1473871615594652

[11] Paul-Christian Bürkner. 2021. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software* 100, 5 (2021), 1–54. https://doi.org/10.18637/jss.v100.i05

[12] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[13] Jorge D. Camba, Pedro Company, and Vetria L. Byrd. 2022. Identifying Deception as a Critical Component of Visualization Literacy. *IEEE Computer Graphics and Applications* 42, 1 (2022), 116–122. https://doi.org/10.1109/MCG.2021.3132004

[14] Ronald Jay Cohen, W. Joel Schneider, and Renée Tobin. 2022. *Psychological Testing and Assessment* (10th ed.). McGraw Hill LLC, New York, NY.

[15] Michael Correll, Enrico Bertini, and Steven Franconeri. 2020. Truncating the Y-Axis: Threat or Menace?. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376222

[16] Michael Correll and Jeffrey Heer. 2017. Black hat visualization. In *Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVe), IEEE VIS*, Vol. 1. IEEE, Phoenix, Arizona, USA, 10.

[17] Michael Correll and Jeffrey Heer. 2017. Surprise! Bayesian Weighting for De-Biasing Thematic Maps. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 651–660. https://doi.org/10.1109/TVCG.2016.2598618

[18] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[19] Christine DeMars. 2010. *Item Response Theory*. Oxford University Press, Oxford, United Kingdom. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

[20] Thomas J. Dunn, Thom Baguley, and Vivienne Brunsden. 2014. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology* 105, 3 (2014), 399–412. https://doi.org/10.1111/bjop.12046 arXiv:https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/bjop.12046

[21] Carla Evans. 2020. Measuring Student Success Skills: A Review of the Literature on Critical Thinking. 21st Century Success Skills. https://www.nciea.org/library/measuring-student-success-skills-a-review-of-the-literature-on-critical-thinking/

[22] Arlen Fan, Yuxin Ma, Michelle Mancenido, and Ross Maciejewski. 2022. Annotating Line Charts for Addressing Deception. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 80, 12 pages. https://doi.org/10.1145/3491102.3502138

[23] Iddo Gal. 2002. Adults' Statistical Literacy: Meanings, Components, Responsibilities. *International Statistical Review* 70, 1 (2002), 1–25. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-5823.2002.tb00336.x

[24] Thomas M. Haladyna. 1999. *Developing and validating multiple-choice test items* (2nd ed. ed.). L. Erlbaum Associates, Mahwah, N.J.

[25] Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, and Doheon Lee. 2003. A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery* 7, 1 (Jan. 2003), 81–99. https://doi.org/10.1023/A:1021564703268

[26] Gordon Kindlmann and Carlos Scheidegger. 2014. An Algebraic Process for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2181–2190. https://doi.org/10.1109/TVCG.2014.2346325

[27] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2019. 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300576

[28] Marc Lallanilla. 2014. Misleading Gun-Death Chart Draws Fire. Retrieved August 26, 2022 from https://www.livescience.com/45083-misleading-gun-death-chart.html

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

[30] Leo Yu-Ho Lo, Ayush Gupta, Kento Shigyo, Aoyu Wu, Enrico Bertini, and Huamin Qu. 2022. Misinformed by Visualization: What Do We Learn From Misinformative Visualizations? *Computer Graphics Forum* 41, 3 (2022), 515–525. https://doi.org/10.1111/cgf.14559 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14559

[31] Morgan McFall-Johnsen. 2020. A 'cuckoo' graph with no sense of time or place shows how Georgia bungled coronavirus data as it reopens. Retrieved August 26, 2022 from https://www.businessinsider.com/graph-shows-georgia-bungling-coronavirus-data-2020-5

[32] Andrew McNutt, Gordon Kindlmann, and Michael Correll. 2020. Surfacing Visualization Mirages. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376420

[33] Carlos Monteiro and Janet Ainley. 2007. Investigating the interpretation of media graphs among student teachers. *International Electronic Journal of Mathematics Education* 2, 3 (2007), 187–207.

[34] Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, and Enrico Bertini. 2015. 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* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1469–1478. https://doi.org/10.1145/2702123.2702608

[35] Denise F. Polit, Cheryl Tatano Beck, and Steven V. Owen. 2007. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health* 30, 4 (2007), 459–467. https://doi.org/10.1002/nur.20199 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/nur.20199

[36] Washington Post. 2022. Washington Post | FlowingData. https://flowingdata.com/tag/washington-post/

[37] William Revelle and David M. Condon. 2019. Reliability from α to ω: A tutorial. *Psychological Assessment* 31, 12 (2019), 1395–1411. https://doi.org/10.1037/pas0000754 Place: US Publisher: American Psychological Association.

[38] The New York Times. 2022. Graphics - The New York Times. https://www.nytimes.com/spotlight/graphics

[39] Eliane R. A. Valiati, Marcelo S. Pimenta, and Carla M. D. S. Freitas. 2006. 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* (Venice, Italy) *(BELIV '06)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/1168149.1168169

[40] VisLies. 2017. Vis Lies 2017 Gallery. Retrieved September 01, 2022 from https://www.vislies.org/2017/gallery/

[41] VisLies. 2020. VisLies 2020. Retrieved August 26, 2022 from http://www.vislies.org/2020/

[42] VisLies. 2021. VisLies 2021. Retrieved August 26, 2022 from https://www.vislies.org/2021/

[43] Jane Watson and Rosemary Callingham. 2002. Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal* 2 (11 2002). https://doi.org/10.52041/serj.v2i2.553