Why Question Answering Modality Matters in Visualization Studies and Tests

Anne-Flore Cabouat[†‡], Lily W. Ge[⋆], Tobias Isenberg[†], Samuel Huron[‡], Matthew Kay[⋆], Petra Isenberg[†]
[†]Inria, Université Paris-Saclay, LISN, CNRS
[‡] Télécom Paris, Institut Polytechnique de Paris, i3, CNRS
[⋆] Northwestern University

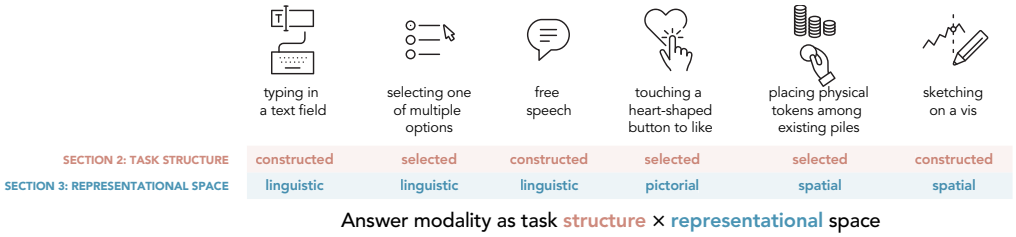


Figure 1: Examples of answer modalities decomposed into two cognitive dimensions discussed in this paper: task structure (*constructed* vs *selected* answers [11]), which affects the strategies available to respondents (Section 2), and representational space, which relates to the mental transformations required to express a response (Section 3).

Abstract

In this position paper, we argue that answer modality choices can affect the validity of visualization evaluations. Visualization evaluation studies frequently rely on specific answer modalities such as multiple-choice questions, open-ended numerical estimates, verbal explanations, or spatial interactions. Yet, answer modality is often treated as a neutral reporting mechanism rather than as a factor that may influence reasoning strategies, representational transformations, and ultimately measured performance and the conclusions researchers can draw from their observations. We discuss two broad classes of modality effects on participant performance: changes in participants' task-solving strategies and changes induced by representational transformations between the visual stimulus and the expected response format. We further argue that answer modality determines what evidence about participants' reasoning is preserved for analysis. We ground our reflections in observations from prior empirical work and propose practical guidance for reasoning about answer modality as part of visualization evaluation design.

Index Terms: Visualization evaluation, Answer modality.

1 Introduction

Visualization research frequently relies on task performance measures to evaluate visualization effectiveness or assess visualization reading ability. In practice, we as researchers implement these evaluations using a large array of answer modalities: participants may, for example, type a numerical estimate or a variable's name in an *open-ended* field, select one or more answer options in a *multiple-choice* question, explain a decision with *spoken words* or by *typing text* in open-ended fields, *point* at a location on a screen (e.g., using a mouse cursor or dragging a visual line), *rank* alternatives, *sketch* a structure, or *physically manipulate* elements. Producing an answer, however, is not merely a passive reporting step: different response modalities may alter the strategies readers use, the information they attend to in the visual artifact, and the mental operations they need to perform to convert a visual percept into a response matching the expected answer modality.

Visualization researchers routinely question assumptions about visual encoding of data, interaction techniques, data input modalities [5], or display environments [3]. Comparatively, few works [30] examine how answer modalities themselves may influence observed performance. For example, existing visualization reading tests [12, 15, 21, 26, 28] overwhelmingly rely on multiple-choice questions. Ge et al. [14] note that, while such tests risk construct misalignment, they are convenient to administer and score, scale well in online settings, and provide seemingly objective measures of correctness, thus facilitating psychometric modeling using Classical Test Theory or Item Response Theory [16]. Our goal is not to argue that one answer modality is inherently better than another, nor that different modalities necessarily capture different phenomena. Rather, we encourage visualization researchers to treat answer modality as a potentially meaningful source of variation whose consequences should be empirically examined.

Based on incidental observations from three past studies (see Table 1), we describe two types of modality effects. First, different answer modalities may allow different cognitive **task-solving strategies**. Selecting among predefined options in multiple-choice data visualization questions does not involve the same processes as generating a value estimate to type in an open-ended field or verbally constructing and reporting an interpretation. We share preliminary evidence of such differences from exploratory analyses of data collected from a cognitive interview study on visualization readability (Study A) and an online study using Mini-VLAT (Study B) in Section 2. Second, answer modalities may affect response content through the **mental transformations** they require, such as converting a visual pattern into a numerical value, a linguistic description, or a spatial response such as pointing at a target. We exemplify this phenomenon in the re-analysis of an online perceptual study (Study C) in Section 3. We further discuss in Section 4 how the analysis of errors in an open-ended modality can surface cases where participants perform a different experimental task than the one intended by the researcher. Finally, we discuss some implications for visualization research in Section 5.

2 Different modalities allow different task-solving strategies

Even though many different question modalities exist, we frequently use selected responses (e.g., multiple-choice questions) and measure performance through accuracy scores (e.g., [19, 24, 35]) in visualization evaluation studies. Selected answers also constitute the dominant answer modality in general-purpose visualization literacy assessments such as the VLAT [21] and Mini-VLAT [26]. Yet, in

Study	Context	Contribution	Type	Original work
Study A	Cognitive interviews conducted during the development of a measuring scale	Observations of reasoning strategies and interpretation processes	Qualitative ($N = 11$)	[8]
Study B	Readability study on 16 visualizations	Relationship between Mini-VLAT scores and performance in study tasks by answer modality; qualification of errors from open-ended answers	Qualitative & Quantitative ($N = 60$)	[9]
Study C	Line average estimate recall	Quantitative	Quantitative ($N = 13$)	[10]

Table 1: Short description of three studies where we found preliminary cues regarding the effect of answer modality effects.

other domains such as educational measurement, answer modality has long been recognized as a factor that can affect how respondents perform tasks and, consequently, what is being measured. For example, in the context of quantitative items—i.e., mathematical reasoning problems that require a numerical answer—Bridgeman [7] argues that open-ended and multiple-choice items differ not only in their scoring procedures, but also in the cognitive processes they afford during task completion. In particular, selected-response formats allow respondents to work backward from the provided answer options, rather than generating an answer independently as they would in many real-world situations. On the other hand, in the medical education domain, constructing an answer does not necessarily make the question more valid, as Hift [18] cautioned. Instead, requiring respondents to generate an answer typically reduces reliability and increases the cost for test administration (e.g., completion time required of the test taker and effort required of the grader).

These arguments point to a broader principle: answer modality may affect not only measured performance but also the ecological and construct validity of an assessment. Accordingly, Hancock [17] argues that assessment items must first be true to the learning objectives educators seek to teach and evaluate rather than "administrational convenience." We argue that the same is true for visualization evaluation tasks and illustrate how similar considerations arise through two empirical examples in the remainder of this section.

2.1 Study A: an illustrative example

In Study A we observed qualitative clues about how multiple-choice questions can change the way participants solve visualization tasks. These clues emerged while we were developing PREVis, a questionnaire to assess how readable people find data visualizations [8]. To evaluate the questionnaire items, we conducted cognitive interviews [2]. At the beginning of each interview, participants completed two reading tasks from the original VLAT [21]. To illustrate two effects of selected-response modality on task-solving strategies, we share here one participant's utterances while answering the question shown in Figure 2. We tagged the different cognitive steps they followed 1 2 3 4 5 6 in Figure 2, and we underline the relevant utterances with the respective color in the following quote.

> "What's the height for the tallest person among the 85 males. I like to do, some height, the tallest is probably like 92 centimeters. Oh, over 92. Okay. Oh, right, right." *[Participant selects the correct answer (197.1 cm)]* *(E: What happened?)* "So yeah, I was looking at the weight line instead of the height line as yet. Anyway... it's because it's a morning is what happened. 195 and 97… I guess it would be closest to 197." *(E: Can you show me with your mouse what you were looking at?)* "I, I know I'm looking at the height and looking at the tallest or I think oh, pretty sure, and then, and then I went across to the 90 bar, I went across to the weight bar instead of height." *(E: At first.)* "Yeah, at first." *(E: And then you went back…)* "And then I noticed that that was none of my choices. And then I went, oh, surely I must be wrong. And then..." *[Participant points at the axis around the 195 value.]* *(E: Okay, great. Thank you.)*

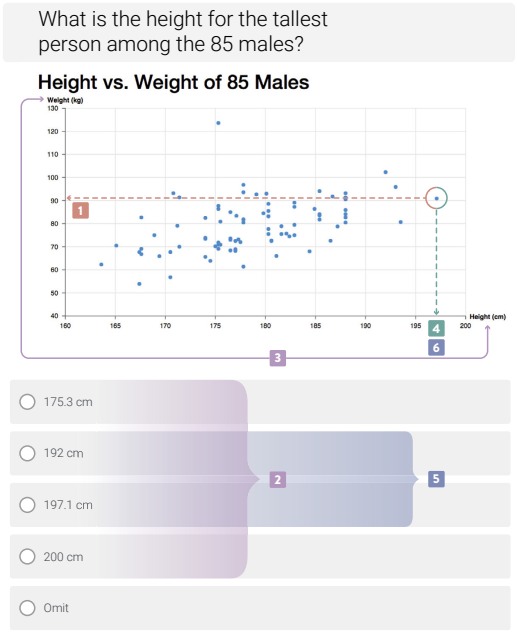


Figure 2: In Study A, we can retrace how a participant answered the question above (*"What is the height for the tallest person among the 85 males?"*, i.e., item #28 from the VLAT [21]) from their think-aloud utterances and answers to follow-up probes.
1 They correctly identify the x-axis extremum as the target point, but mistakenly retrieve a value of 92 from the y-axis.
2 They compare this value against the available answer choices and notice that they are all substantially higher.
3 This discrepancy draws their attention to the axis labels, leading them to realize that they had been reading from the wrong axis.
4 They revisit the visualization and this time appropriately retrieve an approximate value of 195 from the x-axis.
5 This new estimate remains too imprecise to directly determine the answer, but allows them to identify two plausible answer options.
6 Rather than retrieving a more precise value, they perform a final visual comparison to determine which option is closest, ultimately selecting 197.1.

This example shows how multiple-choice items in a visualization literacy test might partly support answer strategies that rely on feedback from the provided options. Since such options are often absent from visualizations in the wild, the multiple-choice format may threaten the validity of the measurement. In their telephone framework, Sarma *et al.* [27] refer to an optimal answer strategy S that the researcher wants to evaluate and the participant's actual strategy S' that the study captures (see Figure 3). Similarly, here, the test's designers intended to evaluate the results of a strategy S, namely, the reading strategy participants would naturally adopt to find the height of the tallest person represented in this chart, but the multiple-choice test item may elicit a strategy S' in which readers use the answer options to guide their interpretation of the chart.

This design is not necessarily problematic. In the example above, following Nobre *et al.*'s [25] taxonomy of barriers to literacy, the

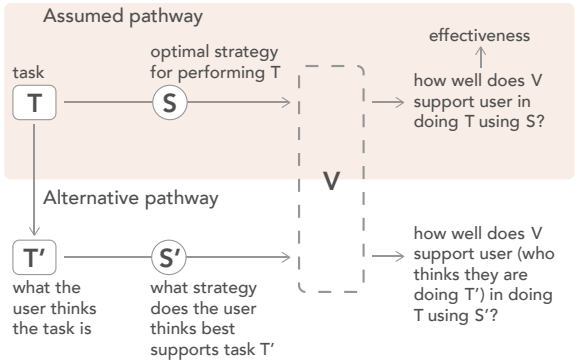


Figure 3: The telephone framework from Sarma *et al.* [27].

Operational barriers — *from Nobre* et al. *[25]*

Occur when a user misreads the question or elements in the visualization, such as data values or color legends.

Definition 1

Conceptual barriers — *from Nobre* et al. *[25]*

Result from a gap in the viewer's understanding of the task or the visual encodings used.

Definition 2

reader's error appears *operational* (Def. 1) rather than *conceptual* (Def. 2). If the purpose of the assessment is to evaluate respondents' *conceptual* understanding of a type of visualization, the multiple-choice format may simply help them recover from an *operational* reading error and does not alter the result's interpretation. The example does, however, raise questions about the exact construct being targeted and measured. In the specific case of a visualization literacy assessment test, one could argue that encoding the height on the horizontal axis rather than the—arguably more congruent [34]—vertical axis imposes additional cognitive effort on the reader to maintain a correct internal representation of the display. Under this interpretation, individuals with high literacy may be better equipped to maintain unintuitive encodings as part of their internal representations—and, therefore, be less likely to face the *operational* barrier this participant encountered in step 1 due to an early interview session time ("it's because it's a morning is what happened"). Then, the recovery 4 facilitated by the answer options in steps 2 3 may actually conceal meaningful literacy differences between readers.

Whether such an operational barrier should be treated as noise or as part of the visualization literacy construct remains an open question. Regardless of how this question is resolved, the example illustrates how answer modality design choices affect which specific phenomena we can observe and measure. In the original VLAT paper [21], this item (#28) is presented as a *Find extremum* task. Yet answering the question also requires retrieving the corresponding value from the chart. The participant indeed performs this sequence of operations in steps 1 and 4. Using Sarma *et al.*'s [27] telephone framework lens in Figure 3, the intended task T is to identify the relevant data point and retrieve its value from the visualization. In steps 2 5 6, however, the participant performs a different task T' instead: evaluating the plausibility of candidate answers.

This example illustrates how, during 1 → 6, part of the task-solving process shifts from reading the visualization to reasoning over the answer options themselves, which echoes concerns raised by Stokes *et al.* [30] on the comparison effects introduced by insight salience ranking tasks compared to open-ended elicitation of insights. Our observations come from a small cognitive interview study and should, therefore, be interpreted as hypothesis-generating. While we

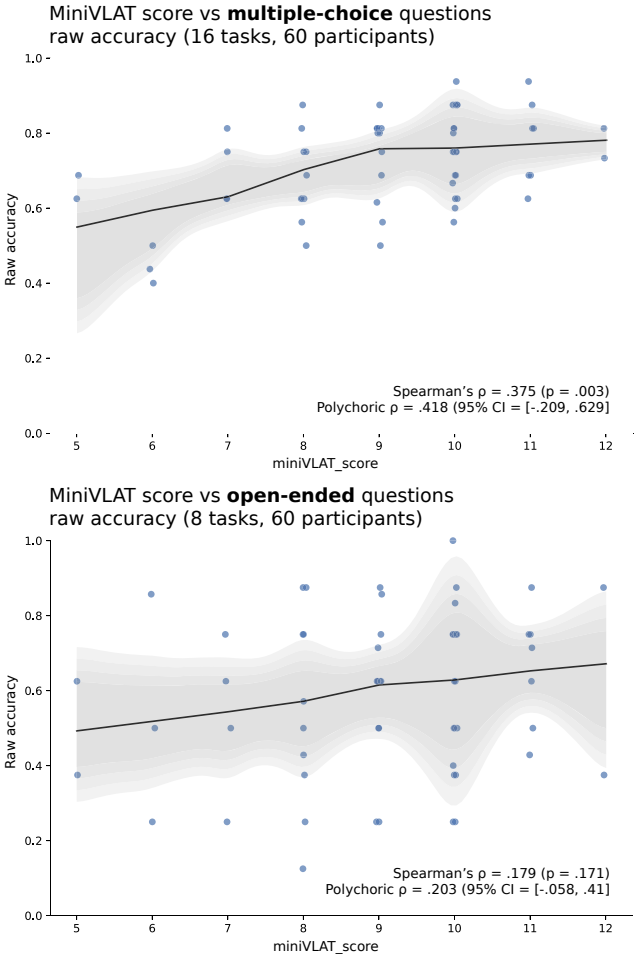


Figure 4: Comparison of participants' raw accuracy on multiple-choice and open-ended questions in our study (y-axes), plotted against their Mini-VLAT scores.

observed the complete error-recovery sequence shown in Figure 2 only once, we encountered other cases in which participants used the answer options to confirm or refine approximate estimates (e.g., selecting 153 after retrieving a value of 150 from a histogram). We next examine whether answer modality may also relate to how visualization literacy measures connect to observed performance.

2.2 Study B: Relating multiple-choice ability test scores to performance on visualization open-ended tasks

During Study B [9], 60 participants recruited through Prolific completed two reading tasks on 16 static visualizations over the course of 4 online sessions. During the first session, participants also answered the Mini-VLAT [21] questionnaire. As part of our pre-registered (`osf.io/7at2d`) exploratory analysis, we investigated the relationship between participants' performance on reading tasks and their Mini-VLAT scores (ranging from 5 to 12 out of 12 in our sample). We observed that this relationship differed depending on task answer modality. Figure 4 shows participants' raw accuracy[1] from our study against their Mini-VLAT scores, faceted by task answer modality. Mini-VLAT scores showed a stronger association with performance

[1] Raw accuracy defined as $\frac{N_{\text{correct}}}{N_{\text{attempted}}}$, where "I don't know" responses were not included as attempted answers. Note that 4 of the 16 visualizations were node-link representations; as these are totally absent from the Mini-VLAT, we exclude them from the data we present here.

Table 2: Overview of the reading tasks used in Study B, and their closest Mini-VLAT equivalent. Each participant completed all tasks.

Answer type *All answers included an "I don't know" option*	***Visual encoding to read***	***Visualization type***	***Example question***	***Target level***	***Task type***	***Closest Mini-VLAT item***
Constructed: open-ended numerical field	Area	Geographical map Treemap	Subregion XX is larger / sent more soldiers than subregion YY. How much larger / more?	Entity	Retrieve value + Compare	Item 9
	Position	Lollipop plot Scatterplot	What is the highest fat content (in g / 100 g) / math score among all items / students?			Item 3
Selected: multiple-choice with all ordinal color bins from legend	Color magnitude	Density map	What is the highest air pollution level in the XX region?	Entity	Retrieve value	Item 3 / 4
		Calendar heatmap	What was the humidity observed on Month+Day?			
Selected: multiple-choice categorical (*True / False / They are the same*)	Area	Geographical map Treemap	Together, regions XX and YY are larger / sent more soldiers than all the other regions combined. True or false?	Dataset	Derive value + Compare	Item 5
	Position	Lollipop plot Scatterplot	Overall, students / dishes from group XX have higher fat content / grades than group YY. True or false?			Item 6
	Color magnitude	Density map Calendar heatmap	Overall, the air pollution / humidity was lower in region XX / month AA than in region YY / month BB. True or false?			Item 4

on the multiple-choice tasks (Spearman's $\rho = 0.375$, $p = 0.003$) than on the open-ended tasks (Spearman's $\rho = 0.179$, $p = 0.171$). This analysis is purely exploratory and should be interpreted with caution: a test for the difference between these dependent rank correlation coefficients [23] was not significant for our sample of 60 participants ($Z = 1.317$, $p = .188$). As such, a cross-sectional study design is required to draw stronger conclusions, with answer modality as an independent variable. Nevertheless, the descriptive correlations suggest a possible difference across answer modalities. Interestingly, the larger association was observed for the multiple-choice questions, even though most of the open-ended tasks are arguably closer to Mini-VLAT items in terms of the low-level tasks they require (see Table 2 for details).

A possible interpretation is that answer modalities do not merely introduce measurement noise, but may alter which abilities are mobilized during task solving. The example from Section 2.1 showed how answer options can become part of the reading process by providing reference points for error detection and self-correction, transforming a data value-retrieval task into an option-ranking task. If such strategies are themselves associated with visualization literacy, then multiple-choice questions may reveal rather than obscure literacy-related differences and improve a test's validity. Conversely, if the strategies are viewed as artifacts of the response format, then the stronger association observed for multiple-choice questions would raise concerns about construct validity.

2.3 Discussion: implications of constructed versus selected answers for experimental validity

In psychological measurement [11], answer modalities are often distinguished according to whether respondents must **construct** an answer themselves or **select** an answer from a predefined set of alternatives. This distinction also cuts across existing visualization evaluation frameworks: Elliott *et al.* [13], for instance, describe *classification* tasks, where viewers assign stimuli to predetermined categories (i.e., selecting an answer), and *identification* tasks, where they identify stimuli through open-ended responses (i.e., constructing an answer). Constructed-response formats include, for example, numerical estimates, free-text explanations, verbal responses, or sketches. Selected-response formats include multiple-choice questions, checkboxes, rankings, and other formats in which respondents choose among predefined alternatives. We emphasize this terminology here because it abstracts beyond specific response formats and highlights a distinction central to the examples discussed in this section: whether information must be generated by the respondent or recognized among candidate answers. Throughout this paper, we use the term *task structure* to refer to this **construct**-versus-**select** distinction, as first illustrated in Figure 1. We acknowledge that task structures are likely distributed along a continuum rather than divided into two discrete and distinct categories. In particular, all answer modalities constrain what participants can express and how they can express it. A numerical estimate entered in a open-ended field that accepts only numerical characters, for instance, affords greater freedom than a multiple-choice question, while still constraining participants to express their answer as a number. A pen-and-paper response affords even greater freedom, allowing participants to write beyond predefined response spaces and supplement their answer with sketches or annotations, but provides less guidance about the expected response format.

The relationship between answer modality and validity is certainly more complex than a simple opposition between ecologically valid constructed-answers and artificial selected-answers. Educational measurement research reports substantial overlap between open-ended and multiple-choice assessments: a recent meta-analysis by Breuer *et al.* [6] found a strong positive correlation between scores obtained from the two formats (mean $r = 0.67$, 95% CI $[0.57, 0.76]$), suggesting that they often capture related underlying abilities. At the same time, students generally obtained lower scores on open-ended assessments than on multiple-choice assessments, and the magnitude of these response-format effects varied across domains. Together, these findings suggest that substantial construct overlap can coexist with meaningful differences in measured performance, and that neither constructed- nor selected-response formats can be assumed to provide a universally more valid measure of a construct. Rather, the implications of answer modality appear to depend on the interaction between the task structure, the abilities being assessed, the design of the assessment items, and the context in which the assessment is administered.

Because our discussion is grounded primarily in modality-focused research from the educational assessment literature, it remains unclear to what extent these findings apply to the visualization tasks commonly used in our field. We present the observations from Study A and Study B not as evidence against multiple-choice questions, but as illustrative examples of how answer modality might shape respondents' behavior and potentially influence the relationship between a construct of interest and measured performance in visualization research. Our observations suggest that the role of

answer modality should be examined explicitly in the context of visualization evaluation. With this idea in mind, we encourage visualization researchers to treat answer modality as a design choice whose implications for construct and ecological validity warrant explicit consideration and discussion.

3 Different modalities involve different mental translations

The examples in the previous section focus on how answer modalities may alter the strategies participants use to solve visualization tasks. Modalities may also influence performance, however, even when the task-solving cognitive strategy remains unchanged. Producing an answer involves more than perceiving a visual representation and solving the expected task. Participants must also translate the outcome of their reasoning (i.e., their *solution*) into the format required by a specific answer modality (i.e., their *response*). This distinction echoes models from the Cognitive Aspects of Survey Methodology literature [32, 33], which distinguish between the mental processes used to arrive at an answer and the process of reporting that answer (Figure 5).

Figure 5: The canonical four stages of survey answering according to Tourangeau [32, 33]: **Comprehension:** understand the task; **Retrieval:** access relevant information; **Judgment:** integrate or evaluate the retrieved information to form a conclusion; **Response:** map that judgment onto the required response format and express it.

Different answer modalities may thus introduce variation to measurement due to the mental transformations between the visual information perceived from the stimulus, the mental solution found by the reader, and the final response they provide for data collection. Expressing information retrieved from an external visual stimulus as a number, a verbalized statement, a written sentence, an interactive selection, a move in a physical space, and so on can all introduce additional sources of imprecision that are independent of the visualization reading process itself.

We observed such translation effects during our re-analysis of data from Study C, a perceptual study by Ceja and Xiong [10] on the estimation of line averages ($N = 13$). Each participant saw 48 line chart stimuli, evenly distributed among 4 line shapes: —— flat, ⌣ bowl, ⌢ hat, and ⩙ noisy. For each line, respondents had to provide an estimate of the line's y-axis average. They answered from memory after seeing a brief moving visual noise mask (see Figure 6). For half of the trials, participants responded by *visually dragging a line* to the estimated position of the average (i.e., *visual* condition); for the other half they responded by *typing a number* in an open text field (i.e., *verbal* condition). The task was identical across both conditions; only the response modality differed.

A mixed-model analysis showed that absolute error was larger in the verbal condition than in the visual condition ($Est = 1.36$, $p = 7.5 \times 10^{-6}$, $SE = 0.3$). This difference varied depending on the line shape condition ⌣ ⌢ ⩙ ——, as shown in Figure 7. The effect of answering modality was even larger for signed error ($Est = 1.93$, $p = 2.7 \times 10^{-7}$, $SE = 0.37$), which is unusual. It might indicate that, in this average estimation task, modality may have elicited a specific error direction bias in which verbal answers were consistently overestimated, and visual answers were underestimated, as Ceja and Xiong mentioned in their analysis [10].

Like our example from Section 2.2, these results are exploratory and should be interpreted with caution. The small sample, furthermore, limits the precision of the estimated effects and their generalizability. They suggest, nonetheless, that response modality may affect performance measurements even when participants perform the same perceptual task. One possible explanation is that

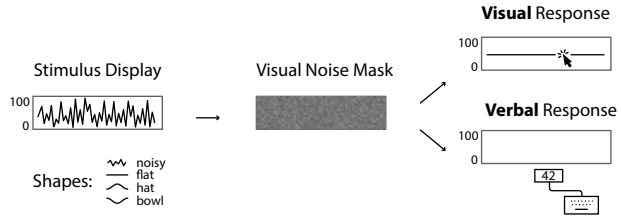


Figure 6: Overview of procedure and experimental conditions in Study C.

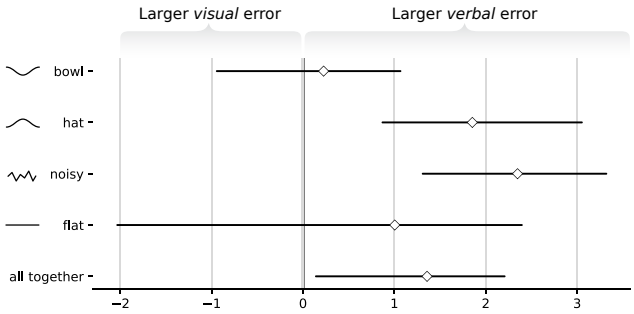


Figure 7: Within-participant absolute error difference between *verbal* and *visual* answering conditions (pairwise t-test and bootstrapped 95% CI).

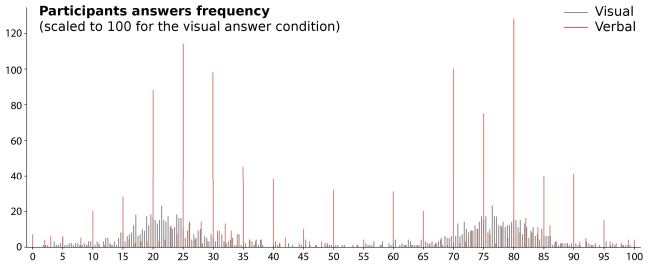


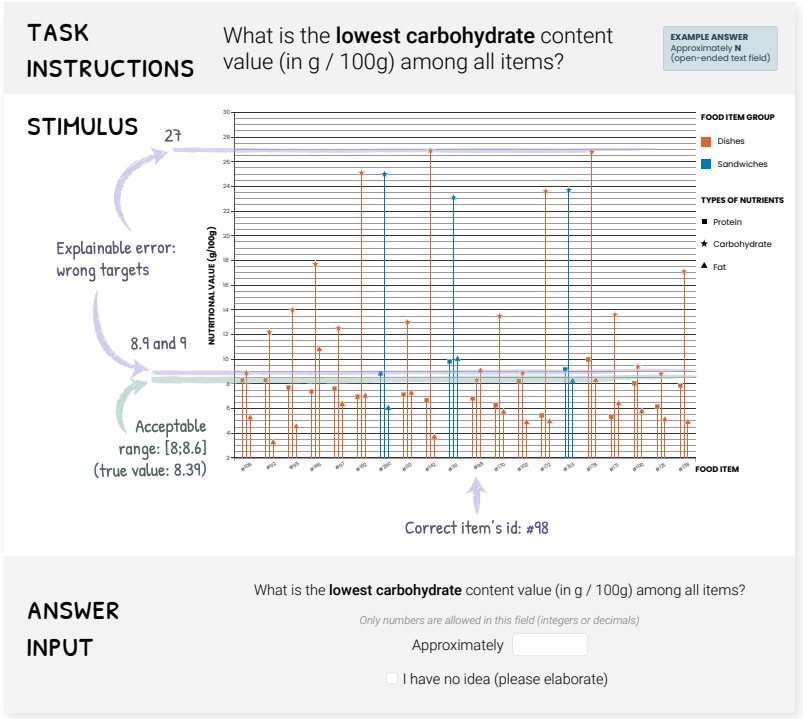
Figure 8: The data collected in Study C show how numerical (typed number) answers in the *verbal* condition are quantized around multiples of 10 and 5 in comparison with *visual* (dragged line) estimates.

the two answer modalities require different representational transformations. In the visual condition, participants could express their estimate using a spatial representation similar to the original visual stimulus. In the verbal condition, participants first had to *translate* their perceptual estimate into a numerical representation before typing a response. This additional transformation may introduce error through mismatches between spatial and numerical representations such as quantization and rounding, as illustrated in Figure 8. The larger absolute error observed in the verbal condition is consistent with the possibility that response modalities may alter measured performance independently of the perceptual judgment itself, because they require greater representational translation. We argue, therefore, that answer modality deserves more explicit consideration in the design, interpretation, and reporting of visualization evaluations.

4 Different modalities affect the conclusions we can draw from errors

In the previous sections, we highlighted how answer modalities might influence performance measures for visualization reading tasks. Similar observations have recently been made in visualization evaluation more broadly. Stokes *et al.* [30] compared free-writing, ranking, and rating methods for studying which insights different charts afford. They found that the observed relationship between

(a) Example task from our study.



(b) Distribution of open-ended answers, including the "I don't know" free text field.

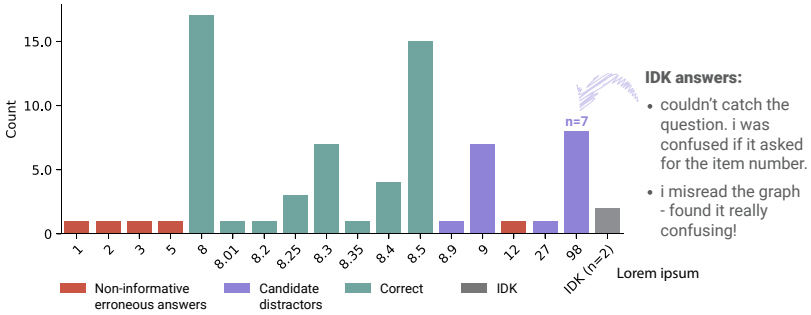


Figure 9: Figure 9a shows an example of a task in Study B. The task instructions, stimulus, and answer input were each presented on a separate screen. Participants were free to navigate back and forth between the instruction and stimulus screens; once they moved to the answer screen, they couldn't go back. Figure 9b describes the answers we collected. We determined the range of acceptable values according to the visual area of the target mark with respect to the grid and other data point distractors. Outside this range, the histogram reveals two qualitatively different error patterns that could inform the design of multiple-choice distractors: erroneous targets and misunderstanding of the task (participants who typed 98).

chart type and insight category depended on the task modality: open-ended responses captured takeaways participants generated spontaneously, whereas selected answers such as ranking and rating foregrounded predefined insight categories and introduced comparison effects. Although all three modalities were applied to the same visualizations and time series data, they led to different conclusions regarding which types of insights were most strongly associated with each type of chart. This finding illustrates a broader point: answer modality shapes not only participants' answering strategies, but also the evidence available for analysis.

We observed a related issue in Study B, where examining the full distribution of constructed-responses revealed patterns that would have been invisible in a binary correct/incorrect scoring scheme. We provide an example in Figure 9. In this task, we asked participants to retrieve the lowest carbohydrate value shown in a lollipop plot and to respond using an open-ended numerical field, as shown in Figure 9a. The distribution of responses shown in Figure 9b suggests several qualitatively different forms of error. Some answers, which we marked as correct, clustered near the correct value (8.39), consistent with estimation imprecision (e.g., 8.3) or quantization (e.g., 8). Other responses matched plausible alternative visual targets (e.g., 8.9 or 9), suggesting that participants may have read a value from the wrong mark despite understanding the task. One participant reported the right value for the wrong task (i.e., 27, which matches the highest carbohydrate value mark in the chart). Another group of responses (e.g., 1, 2, 3, or 5) does not correspond to obvious candidate marks in the visualization and is therefore more difficult to explain; it is possible that they stem from misunderstandings, or even from participants who did not engage faithfully with the task. Particularly informative are the repeated responses of 98. The free-text explanations from a participant who selected "I don't know" indicate that they interpreted the question as asking for the identifier of the item with the lowest carbohydrate value (which was indeed item #98) rather than for the value itself. While this specific participant realized their mistake on the answer screen, seven others apparently did not.

This example demonstrates how constructed answers can provide clues regarding whether errors originate from misunderstanding the task (as described in the telephone framework from Sarma *et al.* [27], Figure 3), an operational (Def. 1 from Nobre *et al.*'s [25] taxonomy) barrier resulting in the selection of the wrong visual target (as exemplified in Section 2.1), a conceptual encoding barrier (Def. 2) resulting in the incorrect decoding of a visual encoding, or simple estimation imprecision in the response formatting stage (from Tourangeau's survey answering model [32], Figure 5, as exemplified in Section 3). Constructed responses may very well provide diagnostic information about task-solving difficulties that remain largely inaccessible with selected-response question scores.

In the context of visualization evaluation, information from open-ended answers may help researchers reason about which errors are most relevant to the construct they seek to assess. If the objective is to evaluate a visualization's support for a specific analytical task, some errors may be reasonably interpreted as peripheral to that task. We do not, however, argue that such errors should necessarily be discarded from a dataset before analysis, nor that constructed-response formats are inherently superior to selected-response formats. Constructed-responses typically require manual review of answers, which makes them harder to reliably scale for large studies. In addition, when quantitative comparison is needed, open-ended answers require the development and validation of scoring procedures, such as coding schemes or scoring rubrics, increasing both the effort and duration of a study. Meanwhile, well-designed multiple-choice questions can also provide diagnostic information, particularly when distractors are grounded in observed or expected error patterns from previous studies. Whether these errors should be considered noise or part of the construct itself remains a validity question that depends on the goals of the evaluation.

Taken together, our examples suggest that answer modality can influence not only how participants solve visualization tasks, but also what researchers are able to observe about those processes. Open-ended questions do not resolve these validity questions, but they can make them visible—at a cost. We next discuss some implications of answer modality for the design of visualization evaluations.

5 Takeaways for Visualization Research

We presented observations that suggest that answer modalities may influence participant performance through at least two mechanisms. First, different task structures (constructed- versus selected-response formats) may encourage different task-solving strategies, as illustrated by the examples discussed in Section 2. Second, different representational spaces (e.g., visuospatial, verbal, or numerical) may require different transformations between a participant's mental solution and their formatted response, as illustrated in Section 3. These two dimensions, presented in Figure 1, offer a useful lens for thinking about answer modality. As further discussed in Section 4, different modalities also vary in the diagnostic information they provide about participants' reasoning processes and errors. These effects are not incidental: they shape what is ultimately measured in a visualization evaluation. Below, we translate our observations into practical considerations that can help researchers align answer modality with their study goals and make the implications explicit.

Match modality to the construct.
The choice of an answer modality should be guided by the construct that researchers intend to evaluate. Different modalities may be appropriate depending on whether the goal is to assess conceptual understanding, perceptual accuracy, decision making, recall, communication, or performance in a real-world task. We suggest three complementary questions researchers may examine to guide their choice of answer modality.

*(1) Which task-solving strategy do you want to ensure participants use (i.e., **constructing** an answer or **selecting** among alternatives)?*
If the objective is, for instance, to assess which solution a viewer would prefer out of a limited set of possible decisions, a selected-response format is appropriate. Conversely, if the goal is to evaluate whether readers can retrieve specific information from a visualization, a constructed-response format may be better, as we discuss in Section 2. Within constructed-answer tasks, researchers can set the degree of freedom afforded to participants; Suh *et al.* [31], for example, discuss how varying prompt specificity in inferential tasks can narrow the range of observations participants need to consider.

(2) Which aspects of participants' reasoning and mistakes should remain observable in the collected responses?
Researchers may wish to distinguish between different types of errors (e.g., misunderstanding the task versus selecting the wrong target) in which case constructed answers (e.g., open-ended text fields) would be more appropriate as we discuss in Section 4. If the researchers' goal, however, is to evaluate how quickly a visualization supports correct decision making, they may primarily care about the final decision reached by participants and when they reached it. In this case, combining training tasks (as suggested by Sarma *et al.* [27]) and a selected-answer modality might be more appropriate.

*(3) Do the **mental transformations** required to express an answer align with what you want to evaluate?*
When the objective is to study perceptual judgments, requiring participants to convert a visual estimate into a numerical value may introduce additional variability such as quantization of numerical estimates (as illustrated in Figure 8), which is unrelated to the studied phenomenon. Visualization researchers already identified whole-number bias as a potential limitation in evaluation frameworks [13], and accounted for such effects in previous studies (e.g., [20, 36]). Translating a visual pattern into a numerical estimate may, however, be appropriate when evaluating a person's ability to retrieve numerical information from a visualization. Conversely, if the construct of interest concerns broader mental constructs such as qualitative interpretations (e.g., the *gist* [1]) or attitudes and beliefs, the response modality should ideally reflect how these constructs are naturally externalized. Recent work by Li *et al.* [22], for example, showed that participants spontaneously relied on different visual and textual representations to express attitudes and beliefs. Our argument echoes their call to design elicitation techniques around the representational strategies people naturally use, and extends this consideration beyond attitudes and beliefs to other constructs evaluated in visualization research.

Use constructed responses to inform selected-response design.
Selected-response formats are often preferred because they are faster to administer, easier to score consistently, and readily support large-scale studies. These practical advantages, however, require empirically grounded distractors that allow researchers to distinguish, for example, misunderstandings of the visualization from task misinterpretation or imprecision introduced by the response format. One possible compromise is to collect constructed responses during pilot studies or early exploratory experiments, and to use these observations to design subsequent selected-response options. Similar recommendations have been made in survey methodology and educational measurement [4, 29]. Rather than merely increasing item difficulty, such distractors can better discriminate between qualitatively different forms of reasoning while preserving many of the practical advantages of selected-response formats.

Control the reasoning strategies that the answer modality allows.
Answer modalities do not merely record participants' conclusions; they also shape the strategies available to them when solving a task. Whether these strategies constitute undesirable shortcuts (e.g., reverse engineering the correct answer option in multiple-choice questions) or legitimate parts of the target construct depends on the evaluation's objectives. Researchers should, therefore, make deliberate choices about which reasoning strategies their evaluation

allows. This reasoning may involve separating answer options from the visualization to enforce a particular reading order, considering whether participants can revisit previous screens, providing explicit "I don't know" options with an open-ended field, or intentionally embracing selected-response formats when these better reflect real-world usage.

Report and explain answer modality choices.
The influences of answer modality we discussed throughout this paper are directly relevant to construct validity, yet answer modality choices are rarely discussed explicitly in visualization research publications. We thus encourage researchers to report and justify answer modality as a deliberate methodological decision rather than treating it as an implementation detail. As with choices regarding participants, tasks, datasets, or visualization designs, authors should explain why a particular response format is appropriate for the construct they seek to measure, and acknowledge the limitations it introduces. Making these choices explicit would facilitate comparison across studies, improve the interpretation of empirical results, and encourage future work investigating the methodological consequences of different answer modalities.

Figure credits

Figure 3 is a reproduction of Figure 3 from Sarma *et al.*'s work [27] (which we use under its CC BY 4.0 license) and we created Figure 6 as an adaptation of Figure 1 from Ceja and Xiong's work [10]. All other figures in this article are our own. All figures except Figure 3 remain under our own personal copyright, with the permission to be used here. We also make them available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license at `osf.io/muvrh`.

Acknowledgments

The authors sincerely thank Cindy Xiong Bearfield for generously sharing the data collected with Cristina R. Ceja, and for her encouragement in pursuing this work. We also thank the reviewers for their helpful feedback and suggestions. This works received the support of the French National Research Agency (ANR) and was supported by the Investments for the Future Program (PIA), CONTINUUM (ANR-21-ESRE-0030).

References

[1] K. Batziakoudi, C. Bortolaso, and J.-D. Fekete. We need to talk (more) about numeracy! Toward a numeracy-informed visualization design process. In *Proc. VisComm*, pp. 10–15. IEEE Computer Society, Los Alamitos, 2025. doi: 10/hb9gb3

[2] P. C. Beatty and G. B. Willis. Research synthesis: The practice of cognitive interviewing. *Public Opin Q*, 71(2):287–311,2007. doi: 10/b7pdj4

[3] A. Bezerianos and P. Isenberg. Perception of visual variables on tiled wall-sized displays for information visualization applications. *IEEE Trans Vis Comput Graph*, 18(12):2516–2525, 2012. doi: 10/hbrzmw

[4] M. Birenbaum and K. K. Tatsuoka. Open-ended versus multiple-choice response formats—It does make a difference for diagnostic purposes. *Appl Psychol Meas*, 11(4):385–395, 1987. doi: 10/fp9w77

[5] N. Bressa, J. Louis, W. Willett, and S. Huron. Input visualization: Collecting and modifying data with visual representations. In *Proc. CHI*, art. no. 499, 18 pp. ACM, New York, 2024. doi: 10/g88ndg

[6] S. Breuer, T. Scherndl, and T. M. Ortner. Effects of response format on achievement and aptitude assessment results: Multi-level random effects meta-analyses. *R Soc Open Sci*, 10(5), art. no. 220456, 25 pp., 2023. doi: 10/hb9gb2

[7] B. Bridgeman. A comparison of quantitative questions in open-ended and multiple-choice formats. *J Educ Meas*, 29(3):253–271, 1992. doi: 10/bqmsvg

[8] A.-F. Cabouat, T. He, P. Isenberg, and T. Isenberg. PREVis: Perceived readability evaluation for visualizations. *IEEE Trans Vis Comput Graph*, 31(1):1083–1093, 2025. doi: 10/njnz

[9] A.-F. Cabouat, P. Isenberg, T. Isenberg, S. Huron, K. Kurzhals, and L. Matzen. Characterizing perceived readability in data visualization: Design and reader factors. Working paper, 2026. url: `hal.science/hal-05654024`.

[10] C. R. Ceja and C. Xiong. Show or tell? Visual and verbal representations bias position recall. In *Posters at IEEE VIS*, 2021. doi: 10/rdqp

[11] R. J. Cohen, W. J. Schneider, and R. Tobin. *Psychological Testing and Assessment*. McGraw Hill LLC, New York, 10th ed., 2022. url: `directtextbook.com/isbn/9781260837025`.

[12] Y. Cui, L. W. Ge, Y. Ding, F. Yang, L. Harrison, and M. Kay. Adaptive assessment of visualization literacy. *IEEE Trans Vis Comput Graph*, 30(1):628–637, 2024. doi: 10/gtjwqh

[13] M. A. Elliott, C. Nothelfer, C. Xiong, and D. A. Szafir. A design space of vision science methods for visualization research. *IEEE Trans Vis Comput Graph*, 27(2):1117–1127, 2021. doi: 10/ghgt5q

[14] L. W. Ge, A.-F. Cabouat, K. Bonilla, Y. Cui, Y. Ding, N. Rakotondravony et al. An autoethnography on visualization literacy: A wicked measurement problem. *IEEE Trans Vis Comput Graph*, 32(1):1394–1404, 2026. doi: 10/rdhj

[15] L. W. Ge, Y. Cui, and M. Kay. CALVI: Critical thinking assessment for literacy in visualizations. In *Proc. CHI*, art. no. 815, 18 pp. ACM, New York, 2023. doi: 10/gtjwqc

[16] R. K. Hambleton and R. W. Jones. An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educ Meas Issues Pract*, 12(3):38–47, 1993. doi: 10/bqcgrb

[17] G. R. Hancock. Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *J Exp Educ*, 62(2):143–157, 1994. doi: 10/bzs3fk

[18] R. J. Hift. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Med Educ*, 14(1), art. no. 249, 18 pp., 2014. doi: 10/gb3365

[19] N. W. Kim, B. Bach, H. Im, S. Schriber, M. Gross, and H. Pfister. Visualizing nonlinear narratives with story curves. *IEEE Trans Vis Comput Graph*, 24(1):595–604, 2018. doi: 10/gcqbpv

[20] N. Kong, J. Heer, and M. Agrawala. Perceptual guidelines for creating rectangular treemaps. *IEEE Trans Vis Comput Graph*, 16(6):990–998, 2010. doi: 10/fsqt5r

[21] S. Lee, S.-H. Kim, and B. C. Kwon. VLAT: Development of a visualization literacy assessment test. *IEEE Trans Vis Comput Graph*, 23(1):551–560, 2017. doi: 10/f92d38

[22] S. Li, R. Deva, A. Narechania, A. Karduni, C. X. Bearfield, and E. Wall. Does a picture paint a thousand words? Using visual and textual channels to understand attitudes and beliefs. In *Proc. CHI*, art. no. 1231, 17 pp. ACM, New York, 2026. doi: 10/hb9gb5

[23] X.-L. Meng, R. Rosenthal, and D. B. Rubin. Comparing correlated correlation coefficients. *Psychol Bull*, 111(1):172–175, 1992. doi: 10/d3mxjg

[24] R. Netzel, M. Burch, and D. Weiskopf. Comparative eye tracking study on node-link visualizations of trajectories. *IEEE Trans Vis Comput Graph*, 20(12):2221–2230, 2014. doi: 10/f6qj65

[25] C. Nobre, K. Zhu, E. Mörth, H. Pfister, and J. Beyer. Reading between the pixels: Investigating the barriers to visualization literacy. In *Proc. CHI*, art. no. 197, 17 pp. ACM, New York, 2024. doi: 10/n98g

[26] S. Pandey and A. Ottley. Mini-VLAT: A short and effective measure of visualization literacy. *Comput Graph Forum*, 42(3):1–11, 2023. doi: 10/gtjwqb

[27] A. Sarma, S. Long, M. Correll, and M. Kay. Tasks and telephones: Threats to experimental validity due to misunderstandings of visualisation tasks and strategies (position paper). OSF preprint, 2024. doi: 10/hb9gbx

[28] A. Saske, L. Koesten, T. Möller, J. Staudner, and S. Kritzinger. A multidimensional assessment method for visualization understanding (MdamV). *IEEE Trans Vis Comput Graph*, 32(3):2695–2708, 2026. doi: 10/qq2n

[29] H. Schuman and S. Presser. The open and closed question. *Am Sociol Rev*, 44(5):692–712, 1979. doi: 10/b72rd8

[30] C. Stokes, K. Lin, and C. X. Bearfield. Write, rank, or rate: Comparing methods for studying visualization affordances. *IEEE Trans Vis Comput Graph*, 32(1):309–319, 2026. doi: 10/hb9gb4

[31] A. Suh, A. Mosca, S. Robinson, Q. Pham, D. Cashman, A. Ottley et al. Inferential tasks as an evaluation technique for visualization. In *EuroVis Short Papers*, pp. 13–17. Eurographics Association, Goslar, 2022. doi: 10/rjg8

[32] R. Tourangeau. Cognitive sciences and survey methods: A cognitive perspective. In *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, pp. 73–100. National Academy Press, Washington DC, 1984. doi: 10/rdqr

[33] R. Tourangeau. The survey response process from a cognitive viewpoint. *Qual Assur Educ*, 26(2):169–181, 2018. doi: 10/hbfqnq

[34] B. Tversky, J. Bauer Morrison, and M. Betrancourt. Animation: Can it facilitate? *Int J Hum-Comput Stud*, 57(4):247–262, 2002. doi: 10/d5frg6

[35] M. Wallinger, B. Jacobsen, S. Kobourov, and M. Nöllenburg. On the readability of abstract set visualizations. *IEEE Trans Vis Comput Graph*, 27(6):2821–2832, 2021. doi: 10/gtgz8g

[36] L. Yao, A. Bezerianos, R. Vuillemot, and P. Isenberg. Visualization in motion: A research agenda and two evaluations. *IEEE Tran Vis Comput Graph*, 28(10):3546–3562, 2022. doi: 10/gr633b