https://courses.cs.washington.edu/courses/cse442/22wi/ (Perception)

Supplemental materials: https://osf.io/3e2qt/

In Data Visualization, some chart types/encodings are more "precise", based on asking people to judge a ratio between two values. In this work, the authors asked people to reproduce up to 8 values. Cleveland & McGill's ranking did not hold.

Rethinking the Ranks of Visual Channels

https://twitter.com/fumeng_yang/status/1451635789016805377

Caitlyn M. McColeman*, Fumeng Yang*, Timothy F. Brady, and Steven Franconeri

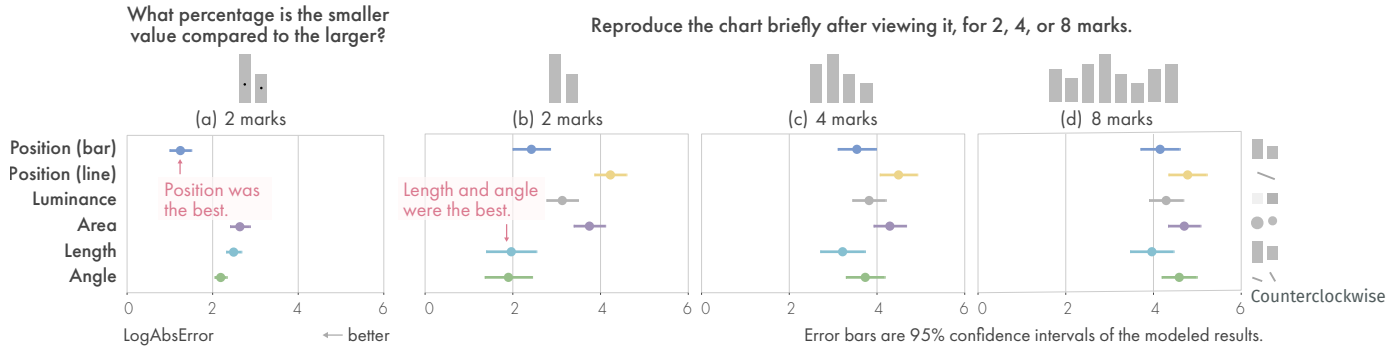


Fig. 1. One core guideline for data visualization design is that some visual channels offer better perceptual precision than others, drawing those precision estimates from two-value ratio judgment tasks [17]. (a) This figure depicts typical data (from [33], 50 participants) showing these judgments are more precise for position (e.g., bar graphs) than for area (e.g., bubble charts). We tested whether that ranking generalizes to the new task of reproducing 2 to 8 previously seen values, and analyzed reproduction bias, precision, and error using a Bayesian modeling approach. (b) This figure shows our modeled results (49 participants). The ranking did not hold, and other factors besides channel choice—like the number of values in the series—had an order of magnitude more influence on performance.

Abstract—Data can be visually represented using visual channels like position, length or luminance. An existing ranking of these visual channels is based on how accurately participants could report the ratio between two depicted values. There is an assumption that this ranking should hold for different tasks and for different numbers of marks. However, there is surprisingly little existing work that tests this assumption, especially given that visually computing ratios is relatively unimportant in real-world visualizations, compared to seeing, remembering, and comparing trends and motifs, across displays that almost universally depict more than two values.

To simulate the information extracted from a glance at a visualization, we instead asked participants to immediately reproduce a set of values from memory after they were shown the visualization. These values could be shown in a bar graph (position (bar)), line graph (position (line)), heat map (luminance), bubble chart (area), misaligned bar graph (length), or 'wind map' (angle). With a Bayesian multilevel modeling approach, we show how the rank positions of visual channels shift across different numbers of marks (2, 4 or 8) and for bias, precision, and error measures. The ranking did not hold, even for reproductions of only 2 marks, and the new probabilistic ranking was highly inconsistent for reproductions of different numbers of marks. Other factors besides channel choice had an order of magnitude more influence on performance, such as the number of values in the series (e.g., more marks led to larger errors), or the value of each mark (e.g., small values were systematically overestimated). Every visual channel was worse for displays with 8 marks than 4, consistent with established limits on visual memory. These results point to the need for a body of empirical studies that move beyond two-value ratio judgments as a baseline for reliably ranking the quality of a visual channel, including testing new tasks (detection of trends or motifs), timescales (immediate computation, or later comparison), and the number of values (from a handful, to thousands).

Paper summary

Index Terms—DataType Agnostic; Human-Subjects Quantitative Studies; Perception & Cognition; Charts, Diagrams, and Plots.

Putative: commonly believed or deemed to be the case; accepted by supposition rather than as a result of proof

https://en.wiktionary.org/wiki/putative

Motif: a recurring or dominant element; a decorative figure that is repeated in a design or pattern

https://en.wiktionary.org/wiki/motif

1 Introduction

Metric values can be efficiently transmitted to the human visual system across a set of channels, including position, length, or intensity [6] (see [56] for review). When creating a visualization, designers face a choice of which channel to depict metric values, with a major constraint being a ranking of putative *perceptual precision* of that channel. This ranking is organized by either expert judgment [50] or operationalized by a particular task. The most referenced operationalization is the precision of making ratio judgments between two values [17,30,33,77].

For example, in Fig. 2a, the viewer might use the position channel to estimate that the value for A is 85% of the value for B, close to the correct answer of 80%. The typical (log) error for this judgment is shown in Fig. 1a. It is typically the lowest error of any channel. Making the same judgment in Fig. 2b between A and B (now separated vertically) is a bit tougher, as reflected by the larger error value for ratio judgments of length in Fig. 1a. Finally, Fig. 2c shows the same data plotted as luminances. While we do not know of empirical measurements of ratio judgment error for this channel, expert judgment [17] (as well as ours) suggests that the error would be quite high [50].

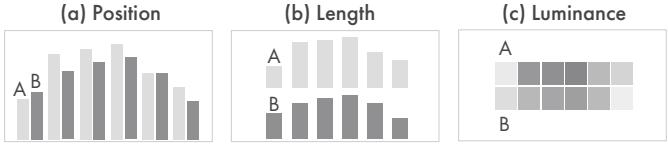


Fig. 2. **Examples of visualization designs that use three different visual channels.** (a) This bar chart relies on the position channel for comparison, (b) this bar chart relies on the length channel for vertical comparisons between A and B, and (c) this heatmap relies on the luminance channel. A two-value ratio judgment is precise in (a), and progressively less precise from (b) to (c).

- *Caitlyn McColeman* is with Northwestern University. E-mail: caitlyn.mccoleman@gmail.com.*
- *Fumeng Yang* is with Brown University. E-mail:fy@brown.edu.*
  ** Both authors contributed equally to this work.*
- *Timothy Brady is with University of San Diego. E-mail:timbrady@ucsd.edu.*
- *Steven Franconeri is with Northwestern University. E-mail:franconeri@northwestern.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/

Show this example starting with the chart on the right ("in this chart, what percentage is A compared to B?")

Unease: anxiety or discontent

1.1 Beyond a ranking based on two-value ratio precision

The channel ranking derived from error measurements of two-value ratio judgments likely deserves its role as a key factor that determines the choice of a channel for depicting metric values. But there is an implicit assumption that it should extrapolate broadly across the types of lower-level visual tasks that viewers execute in real-world visualizations and visual analytics. This is a bold assumption, because visualizations require that we extract, remember, and compare sets of statistics, trends and motifs, across visualizations that almost universally depict more than two values, leading to increasing unease about the dominance of this method of ranking channels [7]. A taxonomy of such operations presents ten low-level perceptual tasks used in analyzing a set of datasets [2]. Interestingly, computing ratios does not appear as a task. 'Retrieving a value' is present, and it is plausible that this task is the foundation of a two-value ratio judgment for charts [17]. A more recent survey includes 'computing derived value' but also reveals concerns about task-dependent effectiveness [66]. In Fig. 2a, if the viewer knew that the maximum value of the *y*-axis were 10, computing a ratio between any bar and that number would allow the viewer to extract the value of a single bar from an unlabeled (or sparsely-labeled) axis.

In Fig. 2, for each of the three visualizations, where are the three highest, or lowest, values for A or B? Where are the largest (or average) differences for each value pairing across the series? There are dozens of such critical comparisons that all involve more than two points (see [7, 27, 57] for review), and there is insufficient empirical work that evaluates whether the ranking of channels extracted from two-value ratio tasks also applies to them (see Sec. 2).

1.2 The present study: reproduction as a proxy for various comparison tasks

How might one compare performance for each channel across such a long list of potential comparison tasks? We start with the assumption that many of these comparison tasks require that one set of values be held in visual memory, and that memory is compared to a subsequently perceived set. For example, in any panel of Fig. 2, computing a two-value ratio might not feel like it requires a heavy memory component. But comparing the global shape of series A versus series B feels far more capacity-limited [78] and memory-intensive [21]. Indeed, evidence from the visual memory and attention literatures suggest that for such more complex comparisons, one must first inspect A, hold the set in memory, and then compare that memory to set B [38, 84, 85]. At the very least, comparisons that are not 'within the eyespan' [79], requiring an eye movement or turn of a page, certainly require, and will be limited by, visual memory. See section 2.3 for examples

Visual memory is highly capacity-limited [73]. As we attempt to remember more information, precision plummets, bias quickly increases, and storage capacity hits ceiling limits (see Sec. 2.3). Therefore, we would expect the number of data values involved in comparison tasks to predict whether the viewer is successful. Because memory serves as a critical gateway for performance in comparison tasks, the present study measures how a viewer's memory precision, bias, and overall error is affected by the channel used to encode a dataset, and how those measures are affected by the number of data values that the viewer is asked to process and remember.

The present study measures memory using a reproduction task, under the assumption that this measurement will generalize to a variety of comparison tasks. If we had instead used a more specific comparison task, which would we pick? Comparisons of data distributions? A search for the longest set of relatively low values? Ask for the differences in the global shape across the two series? If so, what type of difference, and how would it be reported? And how different should the two data series be, and in what ways? The present reproduction task allows a first look at how channel and number of marks affects reproduction performance, without the need to consider these more specific operationalizations of the various types of visual comparisons. We hope that after this initial exploration, the field can begin to ask more targeted empirical questions for particular comparison tasks.

We asked participants to immediately reproduce a set of values seen moments earlier across six channels and three numbers of marks $\{2, 4, 8\}$. Our results from a Bayesian multilevel model show that the previous ranking [18] does not hold, even for reproducing only 2 marks. The new probabilistic ranking also varies with the number of marks. Other factors besides channel choice have an order of magnitude more influence on performance, such as the number of marks in the series, or the value of each mark. Across every visual channel, performance drops precipitously when more than just a few marks have to be stored, consistent with the known limits on visual memory.

1.3 Contributions

This work challenges the assumption that the ranking derived from the precision of judging a ratio between two visual marks will extrapolate to new tasks, especially those that involve more than two marks.

Our primary contributions are as follows.

- **Experimental study results** on the effects of six typical encoding channels, and the number of marks $\{2, 4, 8\}$, on a task of reproducing a set of visualized data, leading to a reassessment of the value of rankings based on two-value ratio tasks.
- **A contextual, probabilistic ranking** of the six visual channels on three statistical measures: bias, precision, and error.
- **A publicly-accessible dataset** of 28,602 responses measuring that reproduction performance, as well as a Bayesian multilevel model to describe the dataset. The dataset, analysis script, and model files are available at `https://doi.org/10.17605/OSF.IO/3E2QT`.

2 Related Work

Here we surveyed work in visual perception, information visualization, and visual working memory to gather considerations for factors that may impact visual reproduction performance.

2.1 Context and bias effects on visual judgements

While Cleveland and McGill [18] tested the precision of ratio judgements with only two relevant values for the judgment itself, they also showed decreased precision for displays where those values were crowded by adding other values in the display [17, 33]. More recent work [77, 88] identified similar impairments. In other reproduction tasks, like the one used in the present study, surrounding values in a display created memory biases, such that recollections of a single relevant value were repulsed from the 0, .5, and 1.0 proportion of a second larger reference bar [52]. Memory bias has been shown even for values presented alone, such that tall bars with a high height:width ratio were underestimated, and wide bars with a low height:width ratio were overestimated [14].

Recollection: the action of remembering something

2.2 Evaluations beyond two-value ratio precision

After one study showed that correlation judgments follow a systematically measurable profile of perceptual precision for scatterplots using the position channel [69], a later study ranked the relative precision of correlation judgments across other visualizations, finding that position-based scatterplots offered the highest precision, but position-based line charts offered the lowest precision [31]. Angle, a channel with low precision on a two-value ratio task, showed the second-highest precision [31]. Though in this case, the correlation judgment may not have been perceptually extracted by angle *per se*, but emergent shapes created by the angles for high negative correlations.

With judgments of aggregate properties of mean, average, or spread, the typical ranking can reverse, such that typically low-ranked values like luminance (in this case, a ramp combining luminance and color saturation) can actually lead to the best performance in those tasks [1] (see [76] for review). Judgments of minimum, maximum, or range were still best for visualizations that used position channels. Another study asked participants to complete four tasks—read value, compare values, find maximum, and compare averages—across visualizations that relied on position, size, or color (similar to the luminance ramp used here). They found similar results, where extracting one value, or comparing two single values, was fast and accurate for position, but for aggregate properties like comparing averages, the color condition showed equal performance [44]. Another study, similar in spirit, tested

Eyespan: the distance between the eyes

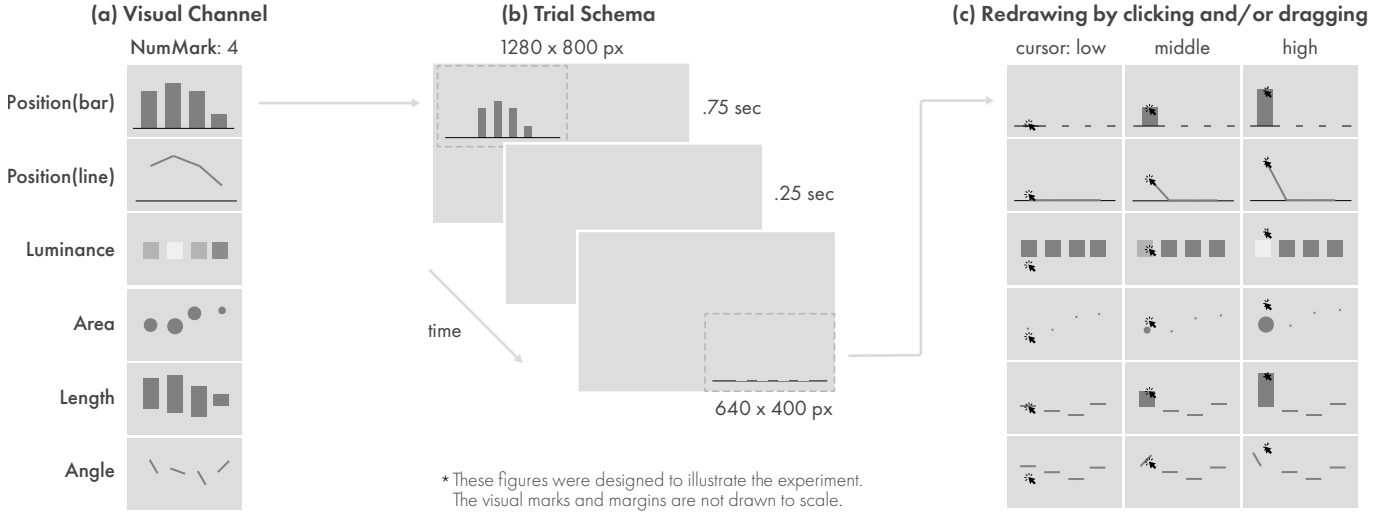


Fig. 3. **Visual channels and the reproduction task**. (a) Examples of visual channels for 4 visual marks. (b) Each trial followed a "show-remove-reproduce" procedure to indicate their responses. (c) In reproduction, participants clicked on the screen or dragged the mouse to redraw the previously-seen marks. In all conditions, the visual channel changed as a linear function of the vertical position of the mouse cursor, such that even angle and area were changed by dragging the mouse up-and-down. For area channel, participants adjusted based on area not radius. More details are available in Appendix A.

the speed, accuracy, and preference for ten data visualization tasks across scatterplots, bar charts, pie charts, and line charts [72]. They found, for example, advantages for bar charts in finding value clusters, or that scatterplots show advantages for anomaly detection, but not for cluster detection.

Others evaluated the visual channels for comparison (measured by staircasing threshold differences that could be detected in a limited time [86]) across two tasks, finding the maximum difference among two paired values in a display similar to the left bar chart in Fig. 2, or the stronger correlation between two such pairings of values. The study included bar, line, and donut charts, was focused on comparing value arrangements within each chart type (e.g., juxtaposed vs. interleaved values). Those charts—and their underlying channels—could in theory be compared in their effectiveness for supporting those comparison tasks, but differences in the methods between chart types make that comparison difficult [59].

Similar to the cited studies, the present study relies on a single task, but we regard reproduction as a starting point for more generalizable results, compared to two-value ratio precision or a single visual comparison task.

2.3 Visual memory

Working memory is the ability to hold information actively in mind, and to manipulate that information to perform a wide variety of cognitive tasks [3]. For visual memory in particular, when asked to remember visual information across eye movements (e.g., for comparisons) or across interruptions [35], studies typically claim a capacity limit of only '3-4 items' (e.g., [19]). Even for fewer than 3-4 items, when participants recall the sizes, colors, or angles, of previously seen objects, they are notably less accurate in recalling 2 items than 1 item (e.g., [5, 87]).

Remembering more complex conjunctions of visual channels (e.g., both the color and orientation of a mark) is extremely difficult when more than 1-2 objects must be remembered [29, 58]. The performance cost of increasing memory load from just 1 item held in mind at once to 2 items is larger than the cost of increasing the load from 4 to 8 items (e.g., [73]). Thus, the profile of memory performance for tasks that involve only 1 or 2 items at a time may not predict the profile for more complex visual displays [11]. There are also strong contextual dependency effects where values are stored in compressed ways, as relative to other values [10]. In a visualization, increasing the number of memorized values will lead to performance changes that are hard to predict. Since nearly all data visualizations include more than 1 or 2 marks, it is critical to study these cases directly rather than assume the lessons drawn from studies of 1 or 2 marks will generalize to these larger value sets.

In the present study, participants were asked to reproduce data displays that fall *within* (2 marks), *at* (4 marks), or *beyond* working memory capacity (8 marks) to gather data from qualitatively different memory loads. Participants in this task rely on reproduction of values, as opposed to semantic recall of the main message of a visualization [47] or whether they have encountered an entire image before [8]. This task is an analogy to typical visual working memory tasks, acting as a proxy for how one retains values of marks across eye movements and delays (as when reading the text associated with the visualization).

3 Methods

This section presents and justifies our design decisions, along with the description of the stimuli generation process, the experimental design and procedure, and the data collected.

3.1 Visual channels

As introduced above, we chose six visual channels (denoted by *VisualChannel*) to cover a wide range of the original ranks by Cleveland and McGill [18]: position (bar) (bar chart), position (line) (line chart), luminance (heatmap), area (bubble chart), length (misaligned bar chart), and angle (wind map). We show an example of each of the six visual channels in Fig. 3a.

3.2 The number of marks

We tested three different numbers of marks (denoted by *NumMark*): 2, 4, and 8 (Fig. 4a). The 2-mark condition requires that the viewer extract the value of two data visualization marks, replicating the earlier studies based on two-value ratio judgments (e.g., [17, 30, 33, 77]). The

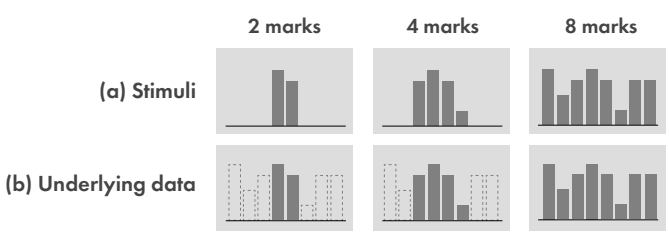


Fig. 4. **Different numbers of visual marks**. We used the same pre-generated datasets across different *NumMark* and *VisualChannel* and removed the side values when showing 2 or 4 marks.

difference in our task is the nature of that extraction, in that participants must redraw it rather than reporting a ratio value. The 4-mark condition aligns with the boundary of typical working memory capacity, and the 8-mark condition exceeds even the most optimistic estimates for human visual working memory. These three conditions have categorically different loads for working memory, allowing us to infer how the working memory limits affect reproduction.

Why 4 and 8 marks/ items

3.3 Experimental design

We split the visual channels into two experiments based on whether the channel uses a common baseline. This split was decided to align participants' mental models and to keep the experiment duration approximately 30 minutes to avoid severe fatigue effects. The first experiment tested position (bar), position (line), and luminance. The second experiment tested length, area, and angle. Each experiment tested all three numbers of marks $\{2, 4, 8\}$. Each participant did the task with 3 visual channels and all 3 numbers of marks, but with different channels for different experiments.

Each pair of *VisualChannel* $\times$ *NumMark* was a block with a series of trials. The first experiment used 13 trials per block. The second experiment used 15 trials per block; this is because, in the pilot study, we found that the second experiment was more difficult: the mapping between the vertical mouse click and the visual change was challenging, and responses were noisier. Thus, we included the additional two trials to offset this additional noise. Within each of the two experiments, the order of visual channels was counterbalanced.

3.4 Generating stimuli

All the values were in the numeric range of [0.01, 1.0] and encoded to the visual channels as follows (see Appendix A for more details). The dimensions of marks were decided to maximize the varying range but to avoid overlapping. The background was set to `rgb(.75,.75,.75)` (light grey) to control visual contrast effects. As a result, position (bar) has the height of each bar ranging from 3.9 pixels to 390 pixels. Position (line) has the height of each line end ranging from 3.9 pixels to 390 pixels. Luminance has the color of each square ranging from `rgb(.5,.5,.5)` (grey) to `rgb(1.0,1.0,1.0)`(white) such that its middle point was the same as the background color. Area has the area of each circle ranging from $\pi(5+1.19 \text{ pixels})^2$ to $\pi(5+37.5 \text{ pixels})^2$; the 5 pixels offset was to ensure that all the circles were visible all the time. Length has the height of each bar ranging from 3.75 pixels to 375 pixels. Lastly, angle has each segment rotated counter-clockwise in the range of 1.8° to 180°. For area, length, and angle, the vertical position of the marks were randomly generated in the range of the y-axis, spanning .0 of its height (i.e., the bottom of the axis range) to .9 of its height.

All datasets were pre-generated, and the same datasets were repeated within the same experiment for different *VisualChannel* $\times$ *NumMark* blocks. Each dataset consisted of 8 numeric values, and each value was randomly and uniformly sampled from the standardized values of $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. When the *NumMark* was 8, participants saw all 8 values in a trial. When the number of marks was 2 (or 4), participants saw the middle two (or four) values; the remaining values were not displayed (Fig. 4b). Each participant viewed different datasets within a *VisualChannel* $\times$ *NumMark* block, and repeated the same datasets across different blocks. The order of the datasets and the values within a dataset were otherwise randomized.

3.5 Procedure

The experimenter first collected informed consent from the participants and then shared an instruction presentation displaying the format, structure and response modality for all trial conditions. The experimenter was present for training and answered clarifying questions the participant had about how to make their response.

Trial As discussed above, in each trial, participants performed a reproduction task. They first saw the stimuli visualization for .75 seconds. The stimuli were then replaced with a blank screen for .25 seconds. Immediately after this, participants were asked to reproduce each visual element (e.g., a bar) as they clicked and/or dragged the mouse to change the pre-marked visual elements on the screen (Fig. 3c). The stimuli visualization was randomly placed in one of the four quadrants (Fig. 3b) and redrawn in the diagonal quadrant. For example, if participants saw the stimuli in the upper left, they redraw the stimuli in the bottom right.

The short duration exposure, along with unlabeled axes, prevent participants from recoding stimuli into other forms [51] and suppress top-down effects like prior knowledge. The duration is adequate for testing visual working memory [51] and provides ample time for the vision system to encode information (e.g., comparing correlation in scatterplots [67], estimating two-value ratio in bar charts [52], etc.). The inclusion of a blank screen as a mask and a different redrawing location together eliminated visual aftereffects.

Participants Thirty and twenty-nine participants were recruited for the two experiments, respectively. They were undergraduate students from the same institution, enrolled in introductory psychology classes, for which they earned partial credit in exchange for their time. Participants were between 18 and 23 years old ($\mu = 19.02$ years, $\sigma = 0.96$; 22 female, 34 male, 3 unspecified), all with normal, or corrected-to-normal vision. The same author and experimenter proctored all the experiment sessions and finished them before the COVID-19 pandemic.

"supervisionou"

Apparatus The experimental system was implemented using Psychophysics Toolbox [12, 46] and MATLAB 2018a, running on a Mac Mini (OS 10.10.5). Stimuli were displayed on a 23" monitor with a resolution set to 1280 $\times$ 800 pixels and a 60 Hz refresh rate. Participants were sat approximately 18.5" from the display.

≈ 47 cm

3.6 Response data

All the raw data from all the participants were considered for analysis with two exceptions. First, 3 and 7 participants from the two experiments, respectively, contributed to the pilot study or were unable to finish the experiment; they were excluded for the purpose of balancing learning and fatigue effects. Second, in the angle condition, when showing a maximum value 1.0 (180°) as the reference, 45.79% of the responses were the same default value of 0.001 (∼0°), resulting in a very large error (100% error). Because both 0°and 180°were a flat segment (see Fig. 3), we think, if not all, the majority of the participants misinterpreted 180° as 0°. To ensure the comparability of our results, we transferred the reference value (1.0) to 0.0 (180° to 0°) for angle.

We recorded the reproduced value of each mark, the order of visual marks, the reference values shown on the screen, the reaction time, *VisualChannel*, *NumMark*, and the trial index. We collected 6,129 trials = 3 *VisualChannels* $\times$ 3 *NumMarks* $\times$ (13 trials $\times$ 27 participants + 15 trials $\times$ 22 participants). Together we analyzed 28,602 responses = 3 *VisualChannels* $\times$ (2 + 4 + 8) marks $\times$ (13 trials $\times$ 27 participants + 15 trials $\times$ 22 participants).

4 Analyses

To analyze the response data, we first decided the measures to quantify the effects, followed by a description of the modeling approach and the model to support the inference.

4.1 Measures

We follow the literature on visual memory and used three statistical measures to compare participants' responses: bias, precision, and individual response level error [10] (see Fig. 5).

Among these, bias is how the mean of the responses deviates from the actual value presented as the reference. Think of bias as systematic error or the tendency to make mistakes in a certain direction, such as exhibiting a bias to overestimate wide bars [14]. Precision is the consistency of participants' responses; they may consistently report the same value, regardless of the reference value. Bias and precision are different facets for the same set of responses. Participants could be precise but consistently underestimate (or overestimate) the value [17, 52]. They could be imprecise but generally right on average. Alternatively, error measures how each response deviates from the reference value. These three measures are different facets for the same distribution of the responses, capturing variations in visual error and reproduction performance through different lenses.

Metrics explained (see Figure 5 caption for a concise description)

Here we used a `student_t(`μ`,`σ`,`ν`)` distribution for a more robust understanding. Bias, the mean of responses, is described by the location

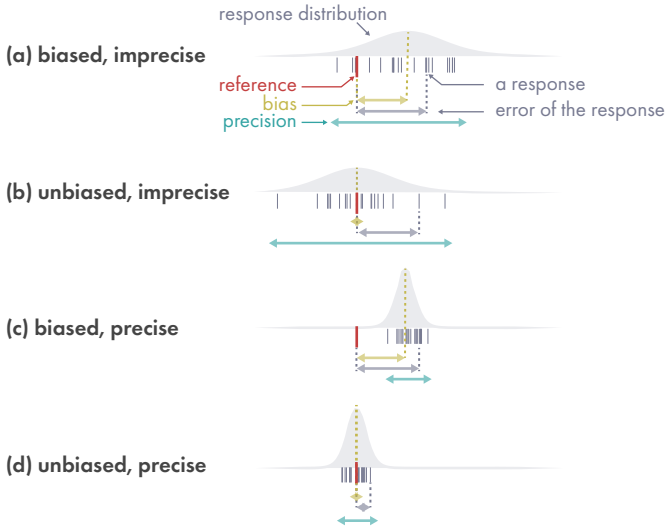


Fig. 5. **Bias, precision, and error.** Bias and precision describe the average properties of a set of responses, while error is a measure for a single response. In this work, error is defined as the deviation from the reference, mean of errors is defined as bias, and standard deviation of errors is defined as precision.

parameter μ; and precision, the consistency of responses, is described by the dispersion parameter σ[1]. The errors of individual responses combine both bias (μ) and precision (σ) of the responses into one measure. If we fit the distribution with the response data collected, then knowing μ and σ, we are able to draw samples from the distribution and calculate error of each draw.

It is important to note that bias and precision describe the average properties of a set of responses (e.g., responses from one or more experimental conditions, one or more participants). However, error is a measure for a single response, combining variance from bias and precision; hence it is with more uncertainty than bias and precision.

Because each of the three measures is associated with a reference, in the remainder of this paper, we subtract the reference value from each response and transform all the raw responses to errors (i.e., relative responses = raw responses − reference values).

4.2 Bayesian multilevel (hierarchical) modeling

We adopted a Bayesian modeling approach to estimate the error distribution. The mean and standard deviation parameters of this distribution, as described above, are considered bias and precision of the responses.

We followed a process of model expansion with *regularization* [53, 65]. It allowed us to understand how each predictor affects the model, to capture more variance in the data while reducing overfitting, and to explore the effects of secondary variables. We started with a *minimal model*, which contained only experimental variables, and a list of potential predictors, ordered by their importance in our subjective beliefs. We then progressively added the predictors and evaluated each intermediate model by inspecting their posterior predictions and posterior distributions of the coefficients. We compared each intermediate model to the last model using WAIC (widely applicable information criterion) and LOO (Leave-One-Out Cross-Validation) for out-of-sample prediction accuracy, and examined their Akaike weights (the probabilities of the differences in these predictions) [49, 53]. We also started with weakly informative priors and gradually regularized the priors as the model expanded [53]. We chose the final model which was the best at addressing our research questions, describing the current data, and predicting future observations.

We implemented the modeling processes using `R` packages `brms` [13], `CmdStanR` [24], `bayesplot` [22, 23], `ggdist` [41], and `tidybayes` [42]. We provide the analysis script and the resulting model files as supplementary materials (the `analysis.Rmd|html` and `*.rds` files).

[1] Strictly, the σ parameter (standard deviation) describes *imprecision*.

4.3 Model specification

Formula Using a syntax similar to `brms`'s [13] extended Wilkinson-Rogers-Pinheiro-Bates notation [64, 80], our final model is

```
1  Response | cens(Censoring) ~ mixture(Student_t(μ1,σ1,ν1), normal(μ2,σ2), θ1)
2  θ1 = MarkChanged + NumMark * VisualChannel * ReferenceValue
3  μ1 = NumMark * VisualChannel * ReferenceValue +
4       ExperimentalTrial +
5       VisualChannel * DataMean +
6       (1 + NumMark * VisualChannel | ParticipantID),
7  σ1 = NumMark * VisualChannel * ReferenceValue +
8       ExperimentalTrial +
9       VisualChannel * DataMean +
10      (1 + NumMark * VisualChannel | ParticipantID)
11 μ2 = DefaultError
```

Explanation

line 1 We treat all the responses as arising from a mixture of two distributions: a `student_t` distribution for all the genuine reproduction responses, and a `normal` distribution for those made without an intention to reproduce a value, termed the 'default' distribution. This is because sometimes participants did not move the mouse to make a response, resulting in a cluster of likely irrelevant responses at a small (known) value. The mixture model separates these two sorts of responses; a mixture model like this is ubiquitous in the visual memory literature [10, 87] for modeling responses.
In the model, the mixture parameter θ_1, the mean (μ_1; bias), and standard deviation (σ_1; precision) of the `student_t` distribution vary with the experimental variables. The mean (μ_2) of the `normal` distribution captures the default responses (see line 11 below). We assumed that the ν_1 parameter of the `student_t` distribution and the standard deviation (σ_2) of the `normal` distribution do not vary. We also left censored the responses to reduce the impact of erroneous responses.

line 2 This line describes the probability of a response coming from the genuine reproduction (cf. default) distribution. This probability could be affected by if the mark was changed (1 or 0), the number of marks, the visual channel used, and the reference value.

The mean (μ_1; bias) of the reproduction distribution is a joint function of a set of linear predictors with varying intercepts and slopes:

line 3 The experimental variables *NumMark* and *VisualChannel* are of the most importance. *ReferenceValue* acknowledges that perceptual errors are likely to be affected by the magnitude of stimuli (e.g., Weber-Fechner's [26, 32], Stevens's power [74], and Guilford's laws [28]) without making a strong assumption about this relationship is the same for different numbers of marks and visual channels; this aligns with the observations that Weber's law appears not to hold for extreme values [25] nor perception of area and angle [74] (see Appendix B for more discussion). The interaction between these variables further generalize this relationship.

line 4 *ExperimentalTrial* captures learning and fatigue effects over the course of the experiment such that we can later divest these effects by conditioning on the median trial.

line 5 *DataMean* is the average of the shown data in a trial. It approximates the context of a response. If the reference value is small but the data mean is large, it may indicate that this response was made in the presence of other large values, and *vice versa*. The interaction with *VisualChannel* is motivated by the speculation that participants may use perceptual proxies for mean [39, 60], and the proxies may be different for different visual channels [31].

line 6 The group-level effects ("random intercepts and slopes") capture the correlation within a participant and also allow each participant to vary for different experiments and experimental conditions.

lines 7-10 The same predictors were used for bias (μ_1) and precision (σ_1) to ensure compatibility.

line 11 The responses from the default distribution, when participants may not be trying to reproduce the value, are always near a small, known value (denoted by *DefaultError*), specified via the informative priors for the mean (μ_2)and standard deviation (σ_2) .

Divest: deprive (someone) of power, rights, or possessions

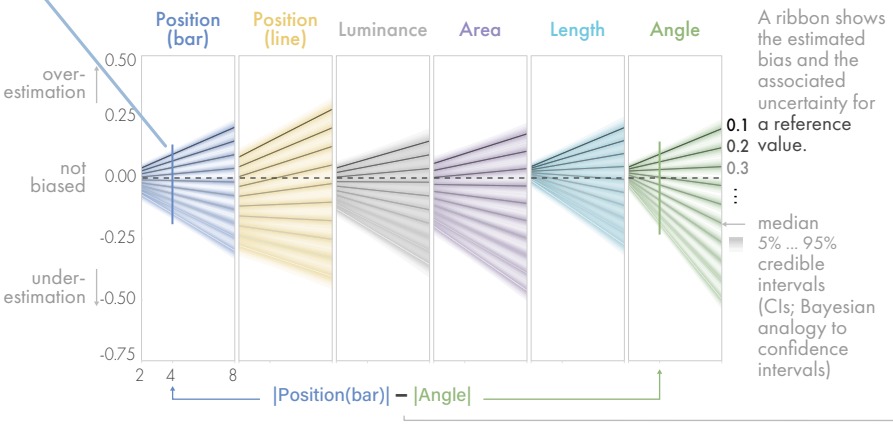
(a) Primary effects

These figures show how bias and precision vary with the number of marks and the reference value. Remember that bias (mean; μ) and precision (standard deviation; σ) are the aggregated properties of a set of responses; they are distributional parameters of the modeled responses (error).

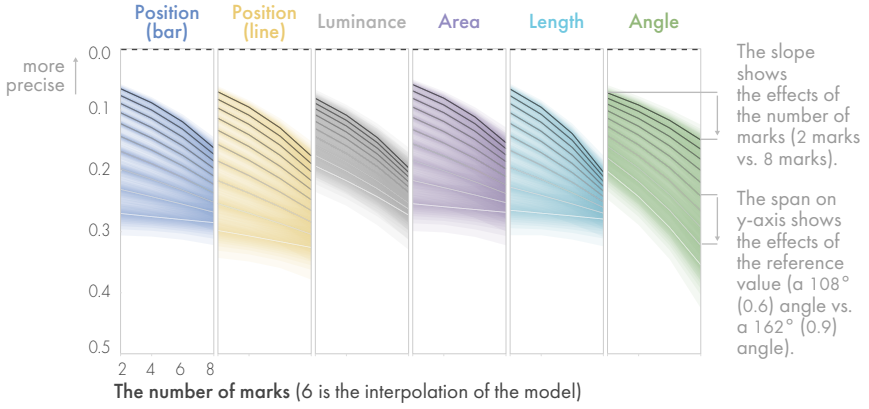
The **slope** shows the effects from the number of marks, and the **distance** between two "ribbons" shows the effects from different reference values. Each ribbon averages across all participants and conditions on the median trial and an average case when data mean is equal to the reference value.

These vertical lines relate to the value 4 on the X-axis

i. Bias (μ) is how the reponses systematically deviate from the truth.



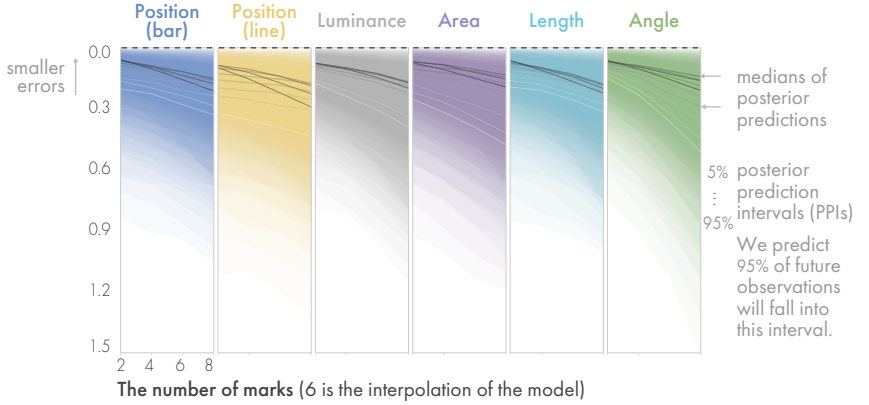
ii. Precision (σ) is how consistent the reponses are.



Mean and standard deviation jointly define the response distributions, and samples from the distribution express the predictions of a future response as the posterior predicted error.

Unlike aggregated properties (bias and precision), errors describe individual responses, in which randomness dominates the differences among reference values and visual channels. Thus, the ribbons overlap with each other.

iii. Error of individual response expresses the prediction of a future response.



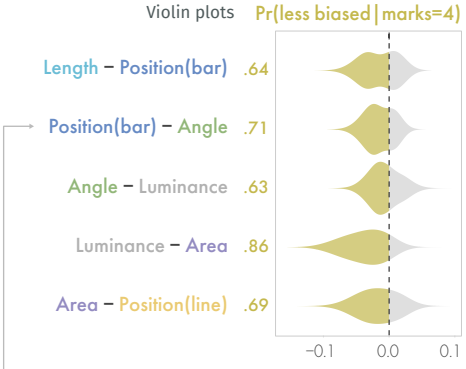
(b) Deriving probabilistic ranks

A ranking list of visual channels with uncertainty is considered **a chain** on which a previous node (a visual channel) is more likely (>50%) to be better than any of its later nodes.

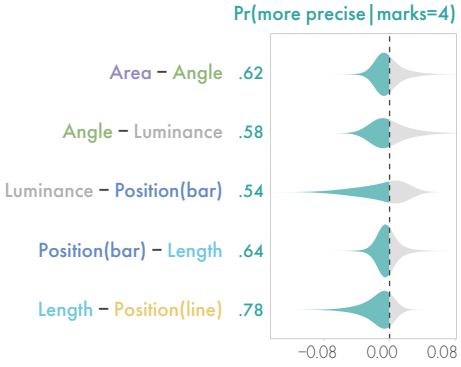
Explanation for the charts below

Such probabilities are found by **subtracting the posterior samples of two channels**. Negative values mean better (e.g., less bias). The proportion of negative values over the entire distribution is the probability.

This figure shows the subtractions and the chain for 4 marks, which becomes the basis of Fig. 8i (4 marks).



The subtractions and the chain for precision. This is the basis of Fig. 8ii (4 marks).



The subtractions and the chain for error. This is the basis of Fig. 8iii (4 marks).

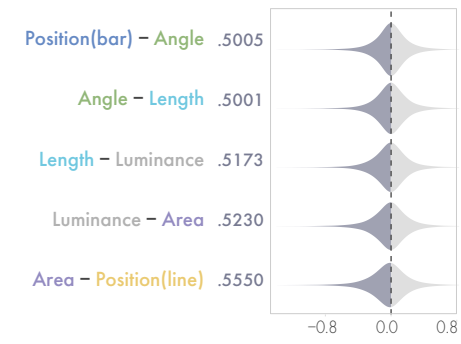


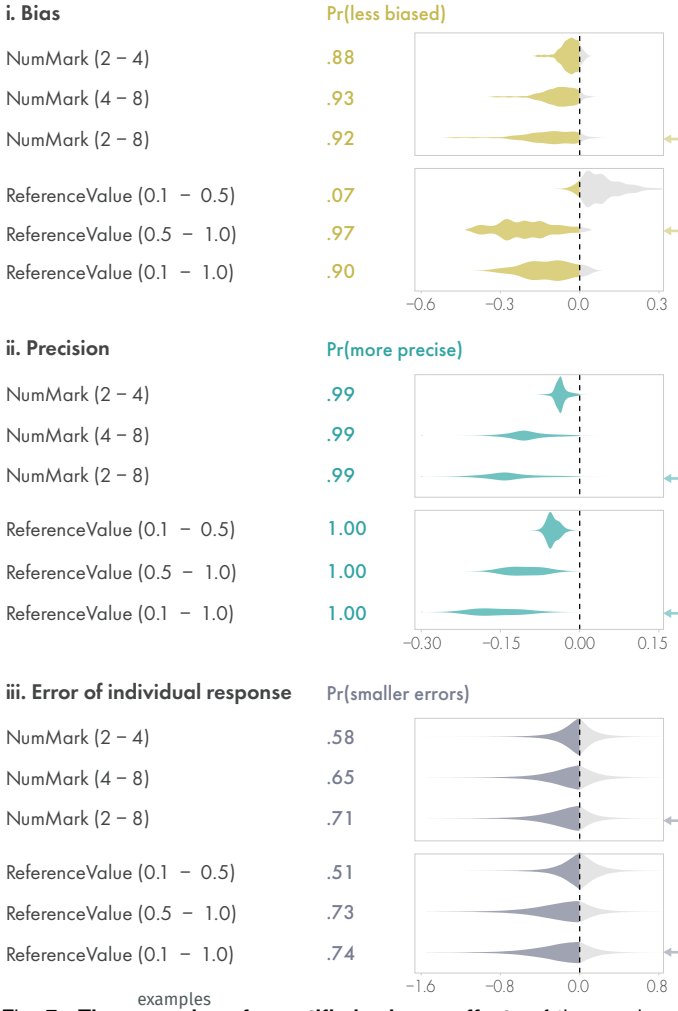
Fig. 6. (a) **The primary effects** modeled from the experimental observations; and (b) how we compare two visual channels, calculate the probabilities of being better, and finally **derive the probabilistic ranks.**

5 Results

To understand the differences in visual channels for the reproduction task, we report various effects on each of the precision, bias, and error measures. We then derive ranks for the visual channels.

Distributions explained

We base our inference on the first distribution of the mixture model and the posterior distributions (marginal, conditional, and predictive distributions). Marginal posterior distributions summarize all the known information for one parameter; conditional posterior distributions tell us the expected value of one parameter in a specific situation; and posterior predictive distributions provide unobserved data conditioning on the observed data and the fitted model.



examples

Fig. 7. **The exmaples of quantified primary effects** of the number of marks and reference values. We take subtraction and calculate the marginal probabilities of being better (see Fig. 6b), averaging across visual channels and reference values (or different numbers of marks).

5.1 Primary effects

The model suggests that the two experimental variables—the number of marks (*NumMark*) and the reference value (*ReferenceValue*)—both have very strong effects on the reproduction responses across the three measures. To show these effects, we take an average participant (to eliminate individual differences), conditional on the median trial (to rid learning/fatigue effects) and on the case where data mean is equal to the reference value (to remove the effects of the other marks in the same trial).

Fig. 6a shows all of the modeled effects, including the tendencies and the interactions between variables. Fig. 7 provides examples of quantified primary effects by showing how likely an average participant's responses are better (less biased/more precise/smaller errors).

i. Bias (Figs. 6i and 7i)

Number of marks. The average participant is very likely to be *less* biased in the reproduction, when the number of marks is *small*. For an average visual channel and an average reference value, the estimated probability that the average participant is less biased in a chart with 2 marks than with 8 marks is .92. That is, for the same reference value, we expect 92% of responses with 2 marks to exhibit less bias than the responses with 8 marks.

Reference value. The average participant is very likely to *overestimate* a small reference value and seriously *underestimate* a large reference value, and are least biased with a reference value around .4 or .5 (median). For an average visual channel and an average number of marks, the estimated probability that the participant is less biased in a chart with the median reference value (.5) than the minimum value (.1) is .93 (this is 1-.07). Similarly, the estimated probability that an average participant is less biased in a chart with the median reference value (.5) than the maximum value (1.0) is .97.

Interaction effects. The effects of *NumMark* and *ReferenceValue* interact, and each interacts with *VisualChannel*. For most of the visual channels but position (line), response bias increases when the number of marks is large and a reference value deviates from the median further. Overall, angle is the visual channel where response bias is most sensitive to either a change in the number of marks or the reference value; position (line) is where bias is sensitive to the reference value, but robust to the number of marks for large reference values.

ii. Precision (Figs. 6ii and 7ii)

Number of marks. The average participant is very likely to be *more* precise (more consistent) when the number of marks is *small*. For an average visual channel and an average reference value, the estimated probability that the participant is more precise with a chart of 2 marks than a chart of 8 marks is .99.

Reference value. The average participant is *more* precise with reproducing a *small* reference value and much less precise with reproducing a large reference value. For an average visual channel and an average number of marks, the estimated probability that the participant is more precise with the minimum reference value (.1) than the median or maximum reference value (.5 or 1.0) is 1.00 (nearly deterministic).

Interaction effects. The effects of these two variables on precision interact with each other and further with visual channels. Response precision is *more* affected by the number of marks when the reference value is *smaller*, except angle, where precision is more affected by the number of marks when the reference value is *large*. Similarly, precision is *more* affected by the reference value with *fewer* marks, except angle, where precision is *more* affected by the reference value with *more* marks. Overall, luminance is the visual channel where precision is least sensitive to the reference value, and position (line) is where precision is most sensitive to the reference value.

iii. Error of individual response (Figs. 6iii and 7iii)

The samples drawn from the posterior distributions provide an estimation of errors in individual responses; for the convenience of comparison, we took the absolute values.

Number of marks. The average participant is likely to make *smaller* errors with *fewer* marks. For an average visual channel and an average reference value, the probability that a single future response exhibits a smaller error with 2 marks than with 8 marks is .71.

Reference value. The average participant is likely to make smaller errors with a *smaller* reference value. The estimated probability that a single future response will have a *smaller* error for the minimum reference value (.1) than the maximum (1.0) is .74.

Interaction effects. Reproduction error is likely affected by the number of marks slightly more in larger reference values for area and angle, less for position (bar) and position (line), and similarly across different reference values for luminance and length. These interactions effects are milder than those observed for bias and precision, owing in part to increased uncertainty in this measure relative to the aggregated properties described by bias and precision.

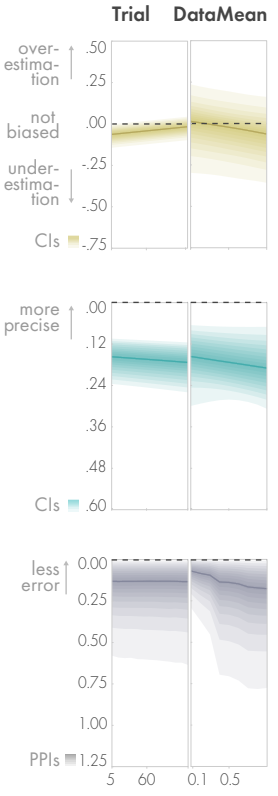
5.2 Secondary effects

The model also suggests several moderate effects. To show the learning/fatigue effect, we condition on the average case where both reference value and the associated data mean are at their median (.5, .55, respectively). To show the effect of data properties (e.g., the mean of all the data values in a trial), we condition on the average case where reference value is at its median (.5) in the median trial, and sampled all the possible values of data mean. We also marginalize out the number of marks and visual channels and use an average participant.

i. Bias The participant appears to underestimate reference values at the beginning of the experiment. In later trials, the participant generally increases the reproduced values and becomes less biased. In reproducing an average value, the participant seems not affected by other small reference values, but is likely to underestimate the median value when other reference values are large.

ii. Precision The participant appears to become slightly less precise as the experiment goes on, possibly due to the fatigue effect. In reproducing an average value, the average participant seems less precise when other larger values are present in the same trial; these larger values possibly distract the participant's judgment and reproduction.

iii. Error of individual response It appears that learning or fatigue effects do not strongly affect response error. In reproducing an average value, the participant is likely to make smaller errors when other reference values are small, and to make larger errors when other values are large. The error of a response seems to largely increase when data mean is above .25, half of the median.



Belie: fail to give a true notion or impression of (something)

5.3 Deriving probabilistic ranks

One primary goal of this work is to derive ranks for the visual channels based on the reproduction task and compare them to those from two-value ratio judgment tasks [18, 30, 33]. A ranking list may provide a summary of effects for others to digest the results (e.g., [31, 70]). However, a rank list may cause *dichotomous thinking* (e.g., "A is always better than B."), which belies the nuances in the ranking. In the spirit of rethinking the previous ranks and in the context of Bayesian statistics, we will derive probabilistic ranks that acknowledge the *uncertainty* in the observations and the modeling processes.

A rank in the probabilistic domain may mean that one is *more likely* to be better than another. Hence we start by calculating these probabilities for visual channels. We marginalize all reference values and condition on the median trial with the median data mean. We subtract the absolute values of each measure; if A is better (less biased/more precise/smaller errors) than B, we expect negative values after subtraction (see Fig. 6b). The estimated probability of A being better than B is the proportion of the negative values over the entire subtracted distribution. When this probability is larger than 50% (larger than chance or other thresholds), we say A is *more likely* to be better than B, and A "wins."

We derive the probabilistic rank lists by pairwise subtraction and then build a chain of visual channels where any previous node on this chain *always* "wins" any comparison to the later nodes for a given measure. This is essentially a *constraint satisfaction problem* [48, 55], and there are many methods to find a solution [48]. For the scale of our problem, we can build the chain by hand or apply heuristics (e.g., sorting by how many times a visual channel "wins"). We construct a chain for each measure and visualize them in Fig. 8, augmented with the associated probabilities to convey uncertainty.

CSP (e.g., Sudoku)

The ranking produced by the precision of two-value ratio judgments does not hold for each of the three measures nor any of the modeled numbers of marks. The ranks changes with different numbers of marks across different measures, which suggests that the previous channel rank is likely not generalize to other visual comparison tasks.

6 Discussion

The varying rankings and the effects of the number of marks and the reference values bring our discussion on the *context* of a visualization below, which further invites a discussion on the implications of this work, along with an acknowledgment of the known limitations.

6.1 The context of a visualization

We find that showing more marks adds substantial noise to memory representations, and has an order of magnitude more influence on performance than the choice of channel. Squeezing more data into one visualization may cause viewers to increasingly remember (and compare) data as statistics or rough global shapes [9, 16], rather than precise representations of individual values. As memory for each value likely also depends on its relation to the distribution of other values, as in work on neighborhood effects [4, 10, 88] and distractor effects [77].

We also indicate that the value of a mark (the reference value) has a more powerful effect on reproduction than the channel chosen: *tall* bars are more biased than *small* areas, even though position (bar) is one of the least biased channels overall. The context of the reference value also indicates strong bias (e.g., angle), similar to past work where participants tend to be more biased and less precise with a value further from the ends of the range [34, 52] (e.g., "edge effects"); it also aligns with psychophysical observations [26], where low and high ends of the data range can serve as perceptual anchors (e.g., "the angle is 10° from 90° is perceptually congruent). Alternatively, for a channel like position (bar), participants perform better around the median value. This is probable that they resort to near-mean estimations when their

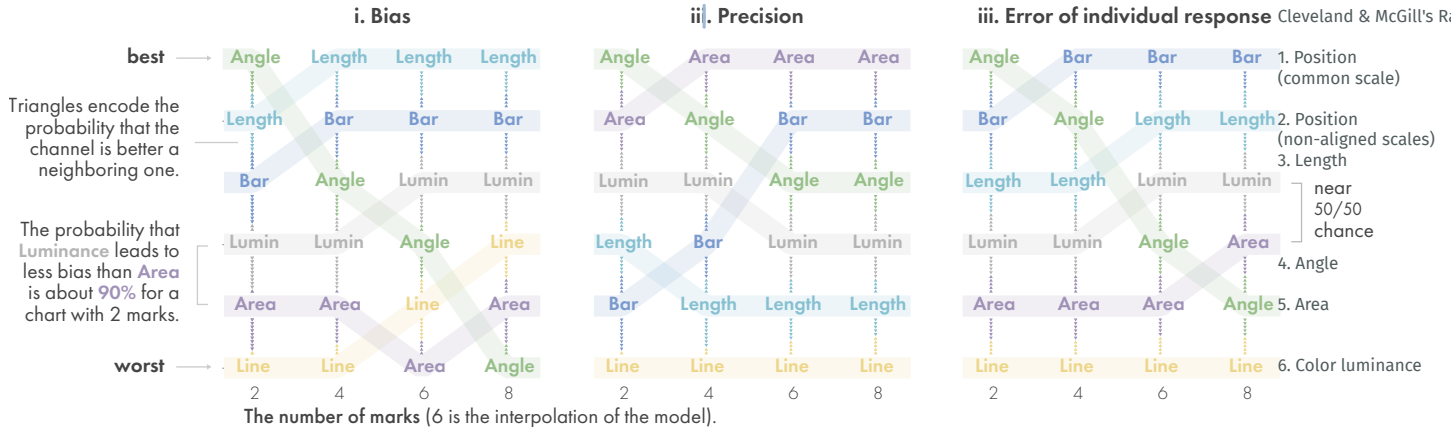
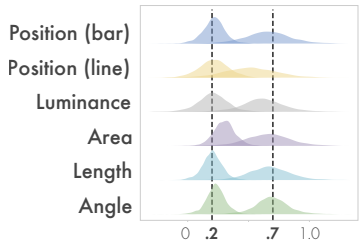


Fig. 8. **The probabilistic rankings of visual channels** on the three measures, augmented with the probability of the associated channel being better (less biased/more precise/smaller errors) than another.



The model's prediction of what Participant 221 would respond when recalling two values {.2, .7}. Length seems to be least biasing for this participant and this set of values, while our overall ranks suggest Angle. If the goal is to eliminate bias, Length should be recommended.

Fig. 9. An example of how the model could support design decision making via predicting a viewer's responses.

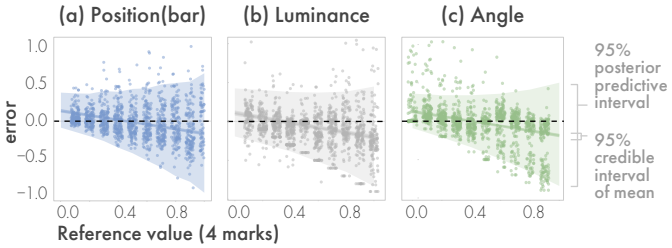


Fig. 10. The moderate non-linearity and symmetry in responses for (b) luminance and (c) angle, compared to (a) position (bar).

Falter: start to lose strength or momentum

memory falters, also consistent with better performance around the mean of the possible range of values [37].

The mark-based reproduction task itself may also influence the context in the eye of the viewer. Line charts use the position channel but were always redrawn lower down than the reference values. They may be perceived as a single complex shape, or set of contrasting slopes, for the purposes of redrawing. Line charts indicate relative changes but may also create more bias in average value judgments for comparisons of lines with different baselines (consistent with [81]). This suggests that the reproduction task may undervalue visual channels that provide good relative, but poor absolute information about the values.

6.2 Implications and future directions

The previous channel ranking based on two-value ratio judgments appears to be attractive as a rule. That task *feels* like a visual comparison distilled down to an atomic unit, which may lead to an assumption that a ranking based on that task should extrapolate to new ones. However, the present study denies this assumption (Fig. 8), with a different ranking for the two-value condition in a reproduction task. As such, designers ought to be increasingly skeptical of the channel ranking produced by two-value ratio tasks [18], which may not serve as a generalizable guideline for their usage.

The present study also shows how the number of marks, the reference value, and other secondary factors such as data mean may strongly, progressively, and interactively affect reproduction. Like previous work that had identified data category [50] and distribution [44] as design factors, other factors beyond visual channels could be critical inputs for design recommendations.

Our rankings reveal the tendencies of the visual channels on different measures and the associated uncertainty, given the reproduction task. If a designer is pursuing the exact or a similar task, these rankings could be used as a reference. For example, length and position (bar) generally lead to less bias and smaller errors, likely desired in reducing bias. Area is surprisingly more precise but could lead to more bias and larger error, likely preferred in improving precision. Luminance is relevantly stable across different numbers of marks and measures, likely suitable when data size is varying. Angle seems sensitive to different numbers of marks and may be most useful for two marks. Position (line) seems ineffective in this reproduction task but may reduce bias for a larger dataset. Knowing one's risk appetite for the misperception of *bias* or *precision* will inform the choice of visual channels. Moving forward, parameterizing the influence of data properties (number of marks, values) and the designers' desire to optimize for lower bias, higher precision or lower error may help to inform visualization designers' decisions.

Thinking of multiple factors may be difficult for designers, not to mention the possible conflicts and other specialized design considerations. Our analysis methods may shed light on resolving this complexity. We were inspired by the recent modeling work [20, 31, 40, 43, 75] and appealed to psychophysical laws [31, 52, 69, 83], entropy [15, 68, 70, 71], perceptual proxies [39, 58], serial-position and ordering effects [36], visual memory (e.g., [3, 10, 54]), neighborhood effects [88], and distractor effects [77]. The final model used, incorporating empirical knowledge, is capable of providing preliminary recommendations given the inputs (see Fig. 9). Thus, a modeling approach like ours may harmonize different factors and provide design candidates.

It would be premature to derive other firm guidelines based on the present study; additional studies will be needed or establish whether there exist broadly applicable guidelines. Our task relies on a purposely abstract reproduction task as a first step toward inspiring future work using more concrete comparison tasks. Those studies will need to expand how the visualization research community operationalizes different visualizations tasks (e.g., detecting trends and motifs, immediate or later comparison, and viewing a visualization with thousands of data points), and what 'good performance' means in a task (e.g., precision, bias, error, speed [44, 72], etc.). In addition, other factors such as top-down effects like prior knowledge [82] and expectations [45, 63] may impact reproduction task performance, and individual differences [62] and spatial ability [61] may affect strategies that subsequently impact task performance, which could be promising directions.

Serial-position effect: tendency of a person to recall the first and last items in a series best, and the middle items worst

https://en.wikipedia.org/wiki/Serial-position_effect

6.3 Limitations

For the sake of comparability, we treated all visual channels equally in designing the experiment and analyzing the data. This makes the channels easier to compare, potentially at the cost of the usability for some. The redrawing method may add noise and bias to data, as it might not be equally intuitive for all the visual channels. For example, for angle, we mapped the y-coordinate of the cursor to the degrees of the angle, which might be more difficult to draw than others (e.g., position (bar) that maps the cursor to the height of the bar). Similarly, always dragging up from the zero value might result in a bias towards smaller values for the two position channels, possibly explaining the underestimation in position (line) noted above. These response methods likely have an impact on the result, and a comparison of different response methods will be critical to generalization of these results. The model also always assumes linearity between errors of the responses and all the variables. While most of the data meet this assumption, visual channels like angle display moderate non-linearity across different reference values (Fig. 10), likely affecting the estimation of the model.

By users/ participants

7 Conclusion

We revisited the ranking of visual channels [17] using a visual reproduction task as a proxy of various visual comparison tasks. We tested participants' reproduction performance with six visual channels: position (bar), position (line), luminance, area, length, and angle across different numbers of marks and data values. With a Bayesian multilevel model, we show that both the number of marks and the reference value strongly affect the bias and precision in a set of responses, as well as errors of individual responses; the number of marks gradually dominates the differences in visual channels and reference values, reflecting a strong limit on working memory, that likely serves to limit most comparison tasks in data visualization. We further derive probabilistic rankings from the model for each measure and show that the previous ranking [18] does not hold. We demonstrate the limitations of the previous ranking [17], offer the preliminary new rankings based on a reproduction task, and present a Bayesian modeling approach to rank visual channels, all for future work to continue exploring this area.

Acknowledgments

Thank you to Satoru Suzuki and members of the Visual Thinking Laboratory at Northwestern University for their suggestions during the experimental design. The authors also thank the anonymous reviewers for their feedback. This work was supported in part by grants BCS-1653457 and IIS-1901485 from the National Science Foundation.

REFERENCES

[1] D. Albers, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 551–560, 2014. doi:10.1145/2556288.2557200.

[2] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*, pages 111–117, 2005. arXiv:15334406, doi:10.1109/INFVIS.2005.1532136.

[3] A. Baddeley. The concept of episodic memory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1413):1345–1350, 2001. doi:10.1098/rstb.2001.0957.

[4] G.-Y. Bae and S. J. Luck. Interactions between visual working memory representations. *Attention, Perception, & Psychophysics*, 79(8):2376–2395, 2017. doi:10.3758/s13414-017-1404-8.

[5] P. M. Bays, R. F. Catalao, and M. Husain. The precision of visual working memory is set by allocation of a shared resource. *Journal of vision*, 9(10), 2009. doi:10.1167/9.10.7.

[6] J. Bertin, W. J. Berg, and H. Wainer. *Semiology of graphics: diagrams, networks, maps*, volume 1. University of Wisconsin press Madison, 1983.

[7] E. Bertini, M. Correll, and S. Franconeri. Why shouldn't all charts be scatter plots? Beyond precision-driven visualizations. *CoRR*, 2020. URL: https://arxiv.org/abs/2008.11310.

[8] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics*, 22(1):519–528, 2015. doi:10.1109/TVCG.2015.2467732.

[9] T. F. Brady and G. A. Alvarez. Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological science*, 22(3):384–392, 2011. doi:10.1177/0956797610397956.

[10] T. F. Brady and G. A. Alvarez. Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of vision*, 15(15):6–6, 2015. doi:10.1167/15.15.6.

[11] T. F. Brady and G. A. Alvarez. No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(3):921, 2015. doi:10.1037/xlm0000075.

[12] D. H. Brainard and S. Vision. The psychophysics toolbox. *Spatial Vision*, 10:433–436, 1997.

[13] P.-C. Bürkner et al. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi:10.18637/jss.v080.i01.

[14] C. Ceja, C. McColeman, C. Xiong, and S. Franconeri. Truth or square: Aspect ratio biases recall of position encodings. *IEEE Transactions on Visualization and Computer Graphics*, 27:1054–1062, 2020. doi:10.1109/TVCG.2020.3030422.

[15] M. Chen and H. Jäenicke. An information-theoretic framework for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1206–1215, 2010. doi:10.1109/TVCG.2010.132.

[16] C. Chunharas, R. L. Rademaker, T. Brady, and J. Serences. Adaptive memory distortion in visual working memory. 2019. doi:10.31234/osf.io/e3m5a.

[17] W. S. Cleveland and R. McGill. Graphical methods graphical perception: Theory , experimentation , and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[18] W. S. Cleveland and R. McGill. An experiment in graphical perception. *International Journal of Man-Machine Studies*, 25(5):491–500, 1986. doi:10.1016/S0020-7373(86)80019-0.

[19] N. Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114, 2001. doi:10.1017/S0140525X01003922.

[20] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018. doi:10.1145/3173574.3173718.

[21] S. L. Franconeri. The nature and status of visual resources. pages 147–162. Oxford University Press, 2013. doi:10.1093/oxfordhb/9780195376746.013.0010.

[22] J. Gabry and T. Mahr. bayesplot: Plotting for bayesian models, 2021. URL: https://mc-stan.org/bayesplot/.

[23] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society*, 182:389–402, 2019. doi:10.1111/rssa.12378.

[24] J. Gabry and R. Češnovar. CmdStanR: the R interface to CmdStan, 2020. URL: https://mc-stan.org/users/interfaces/cmdstan.

[25] H. F. Gaydos. Sensitivity in the judgment of size by finger-span. *The American journal of psychology*, 71(3):557–562, 1958. doi:10.2307/1420251.

[26] G. A. Gescheider. *Psychophysics: the fundamentals*. Psychology Press, 2013.

[27] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011. doi:10.1177/1473871611416549.

[28] J. Guilford. A generalized psychophysical law. *Psychological Review*, 39(1):73–85, 1932. doi:10.1037/h0070969.

[29] K. O. Hardman and N. Cowan. Remembering complex objects in visual working memory: Do capacity limits restrict objects or features? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(2):325, 2015. doi:10.1037/xlm0000031.

[30] L. Harrison, D. Skau, S. Franconeri, A. Lu, and R. Chang. Influencing visual judgment through affective priming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 2949–2958. Association for Computing Machinery, 2013. doi:10.1145/2470654.2481410.

[31] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE transactions on visualization and computer graphics*, 20(12):1943–1952, 2014. doi:10.1109/TVCG.2014.2346979.

[32] S. Hecht. The visual discrimination of intensity and the weber-fechner law. *The Journal of general physiology*, 7(2):235–267, 1924.

[33] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. *Proceedings of the 28th Annual CHI Conference on Human Factors in Computing Systems*, pages 203–212, 2010. doi:10.1145/1753326.1753357.

[34] J. Hollands and B. P. Dyre. Bias in proportion judgments: the cyclical power model. *Psychological review*, 107(3):500, 2000. doi:10.1037/0033-295X.107.3.500.

[35] A. Hollingworth, A. M. Richard, and S. J. Luck. Understanding the function of visual short-term memory: transsaccadic memory, object correspondence, and gaze correction. *Journal of Experimental Psychology: General*, 137(1):163, 2008. doi:10.1037/0096-3445.137.1.163.

[36] G. Hua. *Shaping the Scientific Hypothesis Generation Process through the Design of Visual Analysis Tools*. PhD dissertation, Brown University, 2017. doi:10.26300/vyf3-qw80.

[37] L. Huang. Distinguishing target biases and strategic guesses in visual working memory. *Attention, Perception, & Psychophysics*, 82:1258–1270, 2019. doi:10.3758/s13414-019-01913-2.

[38] L. Huang and H. Pashler. A boolean map theory of visual attention. *Psychological review*, 114(3):599, 2007. doi:10.1037/0033-295X.114.3.599.

[39] N. Jardine, B. D. Ondov, N. Elmqvist, and S. Franconeri. The perceptual proxies of visual comparison. *IEEE transactions on visualization and computer graphics*, 26(1):1012–1021, 2019. doi:10.1109/TVCG.2019.2934786.

[40] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, 26:272–282, 2021. doi:10.1109/TVCG.2020.3030335.

[41] M. Kay. *ggdist: Visualizations of Distributions and Uncertainty*, 2020. R package version 2.4.0. doi:10.5281/zenodo.3879620.

[42] M. Kay. *tidybayes: Tidy Data and Geoms for Bayesian Models*, 2020. doi:10.5281/zenodo.1308151.

[43] M. Kay and J. Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE transactions on visualization and computer graphics*, 22(1):469–478, 2015. doi:10.1109/TVCG.2015.2467671.

[44] Y. Kim and J. Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. In *Computer Graphics Forum*, volume 37, pages 157–167. Wiley Online Library, 2018. doi:10.1111/

cgf.13409.

[45] Y.-S. Kim, K. Reinecke, and J. Hullman. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 1375–1386, 2017. doi:10.1145/3025453.3025592.

[46] M. Kleiner, D. Brainard, D. Pelli, A. Ingling, R. Murray, C. Broussard, et al. What's new in psychtoolbox-3. *Perception*, 36(14):1, 2007.

[47] H.-K. Kong, Z. Liu, and K. Karahalios. Trust and recall of information across varying degrees of title-visualization misalignment. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019. doi:10.1145/3290605.3300576.

[48] V. Kumar. Algorithms for constraint-satisfaction problems: A survey. *AI magazine*, 13(1):32–32, 1992. doi:10.1609/aimag.v13i1.976.

[49] B. Lambert. *A student's guide to Bayesian statistics*. Sage, 2018.

[50] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Tranactions on Graphics*, 5(2):110–141, 1986. doi:10.1145/22949.22950.

[51] R. A. McCarthy and E. K. Warrington. Short-term memory. In *Cognitive Neuropsychology*, pages 275–295. Academic Press, Boston, 1990. doi:10.1016/B978-0-12-481845-3.50016-3.

[52] C. McColeman, M. Feng, L. Harrison, and S. Franconeri. No mark is an island: Precision and category repulsion biases in data reproductions. *IEEE Transactions on Visualization and Computer Graphics*, 27:1063–1072, 2021. doi:10.1109/TVCG.2020.3030345.

[53] R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2016. doi:10.1201/9781315372495.

[54] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956. doi:10.1525/9780520318267-011.

[55] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics*, 25(1):438–448, 2018. doi:10.1109/TVCG.2018.2865240.

[56] T. Munzner. *Visualization analysis and design*. CRC press, 2014.

[57] C. Nothelfer and S. Franconeri. Measures of the benefit of direct encoding of data deltas for data pair relation perception. *IEEE transactions on visualization and computer graphics*, 26(1):311–320, 2019. doi:10.1109/TVCG.2019.2934801.

[58] K. Oberauer and S. Eichenberger. Visual working memory declines when more features must be remembered for each object. *Memory & cognition*, 41(8):1212–1227, 2013. doi:10.3758/s13421-013-0333-6.

[59] B. D. Ondov, N. Jardine, N. Elmqvist, and S. Franconeri. Face to face: Evaluating visual comparison. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):861–871, 2018. doi:10.1109/TVCG.2018.2864884.

[60] B. D. Ondov, F. Yang, M. Kay, N. Elmqvist, and S. Franconeri. Revealing perceptual proxies with adversarial examples. *IEEE transactions on visualization and computer graphics*, 25(1), 2020. doi:10.1109/TVCG.2020.3030429.

[61] A. Ottley, E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, H. A. Taylor, P. K. J. Han, and R. Chang. Improving bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):529–538, 2016. doi:10.1109/TVCG.2015.2467758.

[62] A. Ottley, H. Yang, and R. Chang. Personality as a predictor of user strategy: How locus of control affects search strategies on tree visualizations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, page 3251–3254, 2015. doi:10.1145/2702123.2702590.

[63] E. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 473–482. Association for Computing Machinery, 2013. doi:10.1145/2470654.2470723.

[64] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, S. Heisterkamp, B. Van Willigen, and R. Maintainer. *Package 'nlme'*, 2017. URL: https://CRAN.R-project.org/package=nlme.

[65] X. Pu and M. Kay. The garden of forking paths in visualization: A design space for reliable exploratory visual analytics: Position paper. In *2018 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)*, pages 37–45. IEEE, 2018. doi:10.1109/BELIV.2018.8634103.

[66] G. J. Quadri and P. Rosen. A survey of perception-based visualization studies by task. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–22, 2021. doi:10.1109/TVCG.2021.3098240.

[67] R. A. Rensink. On the prospects for a science of visualization. In *Handbook of human centric visualization*, pages 147–175. Springer, 2014. doi:10.1007/978-1-4614-7485-2_6.

[68] R. A. Rensink. An entropy theory of correlation perception. volume 16, page 811, 09 2016. doi:10.1167/16.12.811.

[69] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. In *Computer Graphics Forum*, volume 29, pages 1203–1210. Wiley Online Library, 2010. doi:10.1111/j.1467-8659.2009.01694.x.

[70] P. Rosen and G. J. Quadri. Linesmooth: An analytical framework for evaluating the effectiveness of smoothing techniques on line charts. *IEEE Transactions on Visualization and Computer Graphics*, 27:1536–1546, 2020. doi:10.1109/TVCG.2020.3030421.

[71] G. Ryan, A. Mosca, R. Chang, and E. Wu. At a glance: Pixel approximate entropy as a measure of line chart complexity. *IEEE transactions on visualization and computer graphics*, 25(1):872–881, 2018. doi:10.1109/TVCG.2018.2865264.

[72] B. Saket, A. Endert, and Ç. Demiralp. Task-based effectiveness of basic visualizations. *IEEE transactions on visualization and computer graphics*, 25(7):2505–2512, 2018. doi:10.1109/TVCG.2018.2829750.

[73] M. W. Schurgin, J. T. Wixted, and T. F. Brady. Psychophysical scaling reveals a unified theory of visual memory strength. *Nature human behaviour*, 4(11):1156–1172, 2020. doi:10.1038/s41562-020-00938-0.

[74] S. S. Stevens. On the psychophysical law. *Psychological Review*, 64(3):153–181, 1957. doi:10.1037/h0046162.

[75] D. A. Szafir. Modeling color difference for visualization design. *IEEE transactions on visualization and computer graphics*, 24(1):392–401, 2017. doi:10.1109/TVCG.2017.2744359.

[76] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of vision*, 16(5):11–11, 2016. doi:10.1167/16.5.11.

[77] J. Talbot, V. Setlur, and A. Anand. Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 20:2152–2160, 2014. doi:10.1109/TVCG.2014.2346320.

[78] L. M. Trick and J. T. Enns. Clusters precede shapes in perceptual organization. *Psychological Science*, 8(2):124–129, 1997. doi:10.1111/j.1467-9280.1997.tb00694.x.

[79] E. R. Tufte, N. H. Goeler, and R. Benson. *Envisioning information*, volume 2. Graphics press Cheshire, CT, 1990.

[80] G. Wilkinson and C. Rogers. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 22(3):392–399, 1973. doi:10.2307/2346786.

[81] C. Xiong, C. R. Ceja, C. J. Ludwig, and S. Franconeri. Biased average position estimates in line and bar graphs: Underestimation, overestimation, and perceptual pull. *IEEE transactions on visualization and computer graphics*, 26(1):301–310, 2019. doi:10.1109/TVCG.2019.2934400.

[82] C. Xiong, L. Van Weelden, and S. Franconeri. The curse of knowledge in visual data communication. *IEEE Transactions on Visualization and Computer Graphics*, 26(10):3051–3062, 2020. doi:10.1109/TVCG.2019.2917689.

[83] F. Yang, L. Harrison, R. A. Rensink, S. Franconeri, and R. Chang. Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics*, 25:1474–1488, 2018. doi:10.1109/TVCG.2018.2810918.

[84] D. Yu, D. Tam, and S. L. Franconeri. Gestalt similarity groupings are not constructed in parallel. *Cognition*, 182:8–13, 2019. doi:10.1016/j.cognition.2018.08.006.

[85] D. Yu, X. Xiao, D. K. Bemis, and S. L. Franconeri. Similarity grouping as feature-based selection. *Psychological Science*, 30(3):376–385, 2019. doi:10.1177/0956797618822798.

[86] L. Yuan, S. Haroz, and S. Franconeri. Perceptual proxies for extracting averages in data visualizations. *Psychonomic bulletin & review*, 26(2):669–676, 2019. doi:10.3758/s13423-018-1525-7.

[87] W. Zhang and S. J. Luck. Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192):233–235, 2008. doi:10.1038/nature06860.

[88] M. Zhao, H. Qu, and M. Sedlmair. Neighborhood perception in bar charts. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019. doi:10.1145/3290605.3300462.

Rethinking the Ranks of Visual Channels

Appendices

A Visual Channels Details

In all the conditions, for all the *visual channels*, participants redrew the stimulus by dragging their mouse above the data mark. This may be obvious for a visual channel such as a bar graph, where the height of the bar reflects the underlying value. It is a little less obvious for a wind map, where the orientation of the line reflects the underlying value. The motivation for this consistent response is to ensure that differences in the observed errors are not because of different motor demands. Using a consistent response method across different visual channels may introduce variance in the data.

For all visual channel types, the data marks were presented within an axis that was nearly half the height of a 23" screen. Running with a resolution of 1280 $\times$ 800 pixels, the usable Y range for each data display was $\frac{800}{2}$ - 20 pixels. The 20 pixels was dynamically determined ($\frac{1}{40}$ of the height) to keep the highest values from ever hitting the top of the screen. Similarly, the *x*-axis was determined to be half of the width, minus an edge buffer ($\frac{1}{30}$ of the width) to keep the data marks from hitting the edge of the screen. The maximum height for a bar is 390 pixels. The bars were 1/16th of the *x*-axis wide (37.3 pixels). As with all of the data marks used in the current experiments, participants were presented with a random selection of values from 0.1 - 1.0. The minimum bar height (0.1) was then 39 pixels.

Experiment A: The visual channels with a common baseline

The bottom of the each visual mark in the first experiment was randomly selected within the y axis range for each trial, so participants could not rely solely on position to remember the values they were shown in the graph. The background for all conditions was light grey (25% black).

Position-bar Participants respond using a computer mouse and clicking/dragging above the data mark to draw it to the size they remember seeing in the initial data presentation. Note that the initial value was 0 for the bar. For this, and all conditions, the response was initialized to the lowest value in the range of possible responses. Participants can click or adjust the same mark multiple times. Their response is fully self-timed. If participants click away from the indicated response space, the graph briefly flashes off the screen to provide unintrusive feedback about the viable response regions.

Position-line The heights of the points on the line were the same as the heights of the bars in the position condition, such that the maximum height for a line chart vertex was 390 pixels (the height of the Y range). The stimulus was created by joining the randomly selected point heights. The line was four pixels wide.

Participants respond by re-drawing the line in the same manner as the bars. They click or drag the point above the zero-mark to adjust the height of the line, in an attempt to match the line chart they saw in the initial stimulus as closely as possible.

Luminance The heat map marks were 37 $\times$ 37 pixels (same width as the bars). While the aligned bars and the line chart represented changing values by changing in height, the heat maps changed how light/dark the presented marks were. The maximum value was white, such that as the participant dragged their mouse higher, the box that they were adjusting became lighter. 0% was represented the same color as the stimuli in the other conditions: RGB values [127.5, 127.5, 127.5] or 50% black.

Participants respond by click-and-dragging above the data point just as in the bar and line graphs. That is, to make the mark darker, they drag the mouse cursor higher on the screen. Just as with the bar and line charts, the initial response started at 0, and participants had to drag the mouse to change the value of the data mark. The only difference between the heat map and the previous conditions' response method, is that, for the heat maps, when participants dragged the mouse up and down, the position of the mark stayed constant and the color changed; in the line and bar charts, when the participants dragged the mouse up and down the position of the mark changed and the color stayed constant.

Experiment B: The misaligned charts

Length As with the other visual channels in this experiment, the misaligned bars do not share a common baseline. Since the baseline varies within the *y*-axis range, the maximum height of these bars is reduced so that there is enough room to have the 0.1 - 1.0 variation while still having the participants respond in the opposite corner from where the bars were originally presented. Otherwise, the bars were the same as in Experiment 1A: the aligned bars condition.

Angle The wind map line segments were as long as the bars were wide ($\frac{1}{16}$ of the *x*-axis). The maximum value (1.0) was presented by a horizontal line (180°). The remaining presented values (0.1 - 1.0) were proportions of the 0-180°range, such that 0.1 was 18°.

Participants make their responses just as in the other conditions: by adjusting the height of the mouse cursor above the data mark. Because there is 180°of possible response space it could be difficult to tell whether a nearly-flat was close to 0 or 180°. To address this concern for participants, we ensured that they were well trained before they began the task, and that there was an origin for the line segment, such that a line falling to the right was closer to 0°and a line falling to the left was closer to 180°. Just as with the other visual channels, participants' response screens were initialized to 0, so it was clear to participants that the initial screen was 0 degrees, they would exceed the 0.5 response only adjusting the line past 90°.

Area The area charts used circle visual channel, where the area of the circle mark changed in direct linear proportion to the size of the presented value. The maximum value (1.0) was represented by a circle with a diameter the same width as the bars ($\frac{1}{16}$ of the *x*-axis, or 37.3 pixels). The remaining values were represented as a proportion of that circle's area, such that the radius for the circle representing 0.1 was $\frac{\sqrt{0.1 * maximum\,area}}{\pi}$.

Participants re-drew the area values just as they re-drew the bar, line, and heatmap above: they dragged their mouse on top of the data mark to make their response. Note that the adjustment of this mark was scaled to the change the area of the circle (not, the diameter) since people tend to perceive the differences in the area of the circle as the natural data mapping. One unit increase in mouse cursor height, then, corresponds to one unit increase in mark area.

B Alternative Modeling Approaches

An alternative approach to incorporate *ReferenceValue* as a predictor is to divide all the responses by their corresponding presented values. This approach assumes that each visual channel follows Weber's law, and therefore division is able to normalize errors. However, this assumption is

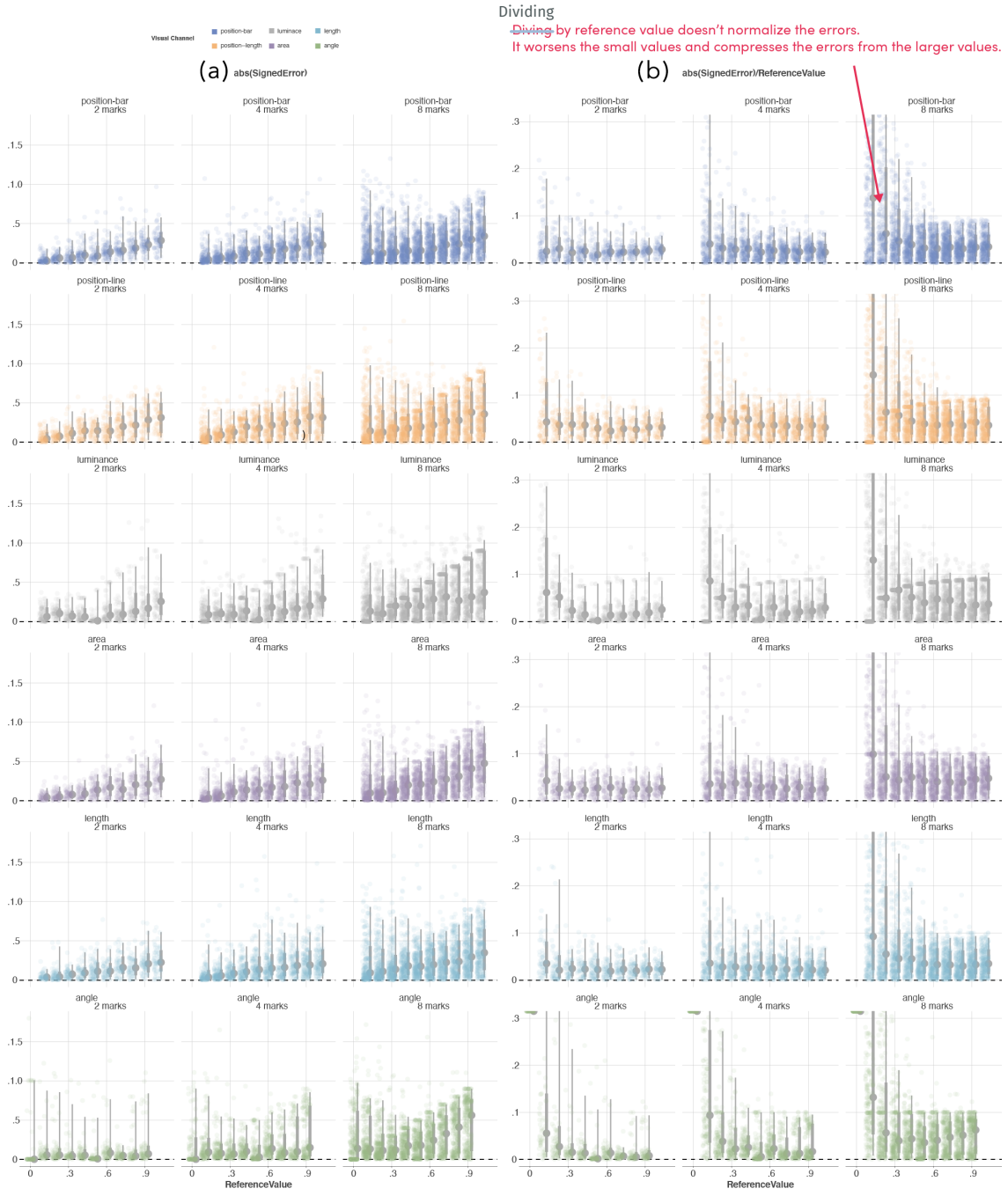


Fig. B.1. Alternative approaches of using our knowledge about perception. We could transfer data by dividing reference value, but this still did not normalize error distributions. Bayesian approaches can directly model skewed distributions. So, using reference value as a predictor is rational.

too strong, and we found that after dividing presented value, errors still vary with presented value, and errors for small presented values were exacerbated. Fig. B.1a shows absolute errors. Fig. B.1b shows absolute errors divided by presented value, where errors are still non-linear.

C Calculating LogAbsError (Ratio) for Our Data

We replicate Cleveland and McGill's analysis to facilitate comparison. Cleveland and McGill's study and their successors based the ranks of visual visual channels on a task of ratio estimation and the log-transformed absolute errors.

$$LogAbsError = log_2(|bias\ of\ percentage\ error| + .125)$$

We utilize the same ratio measure to the modeled bias in our experiment and its reference:

$$LogAbsError = log_2\left(\left|\frac{bias}{Reference}\right| * 100 + .125\right)$$

We apply this measure to our modeled bias and calculate t confidence intervals. Since the model has 49 participants levels, we can consider that we have 49 participants. The results are used to generate Fig. 1.