No Mark is an Island: Precision and Category Repulsion Biases in Data Reproductions

Caitlyn M. McColeman, Lane Harrison *Member, IEEE*, Mi Feng, Steven Franconeri, *Member, IEEE*

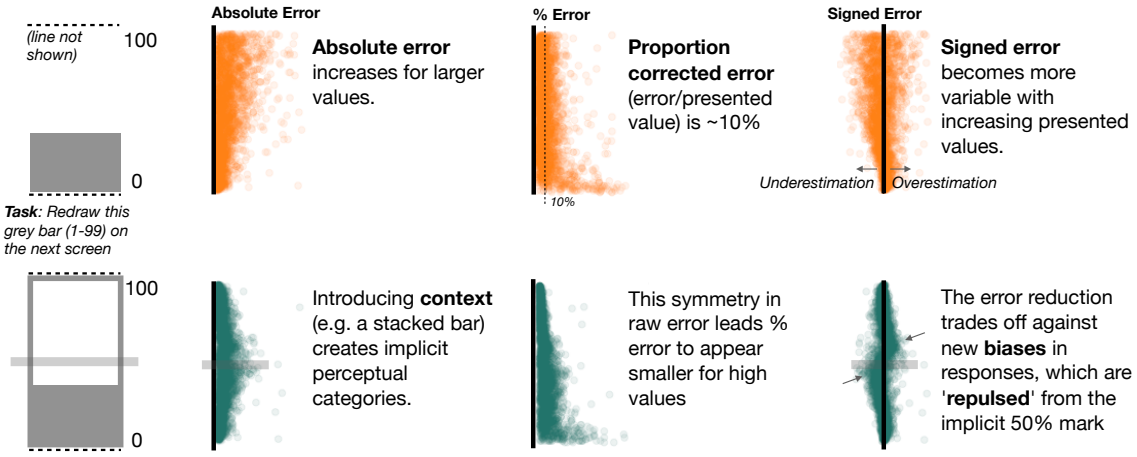

Fig. 1. We measured raw perceptual precision for reproducing marks in simple bar graphs and dot plots. Reproductions of smaller values are overestimated, and larger values underestimated. Reproduction error was approximately 10% of the actual value, regardless of whether the reproduction was done on a common baseline with the original. When a reference point is introduced to the value, there is a repulsion effect such that values are remembered as being farther from that reference.

Abstract—Data visualization is powerful in large part because it facilitates *visual* extraction of values. Yet, existing measures of perceptual precision for data channels (e.g., position, length, orientation, etc.) are based largely on verbal reports of ratio judgments between two values (e.g., [7]). Verbal report conflates multiple sources of error beyond actual visual precision, introducing a ratio computation between these values and a requirement to translate that ratio to a verbal number. Here we observe raw measures of precision by eliminating both ratio computations and verbal reports; we simply ask participants to reproduce marks (a single bar or dot) to match a previously seen one. We manipulated whether the mark was initially presented (and later drawn) alone, paired with a reference (e.g. a second '100%' bar also present at test, or a y-axis for the dot), or integrated with the reference (merging that reference bar into a stacked bar graph, or placing the dot directly on the axis). Reproductions of smaller values were overestimated, and larger values were underestimated, suggesting systematic memory biases. Average reproduction error was around 10% of the actual value, regardless of whether the reproduction was done on a common baseline with the original. In the reference and (especially) the integrated conditions, responses were repulsed from an implicit midpoint of the reference mark, such that values above 50% were overestimated, and values below 50% were underestimated. This reproduction paradigm may serve within a new suite of more fundamental measures of the precision of graphical perception.

Index Terms—Cognition and perception, Graphical perception, Perceptual biases, Ratio perception

1 Introduction

When choosing how to visualize data, one major constraint is choosing a data encoding that most precisely conveys data values to the eye. This decision of which data encoding to use is typically based on past work that ranked visual encodings by how precisely participants could verbally report the ratio between two values (e.g., [14, 6, 39]). This part-to-whole judgment is a useful metric for many tasks, but it also confounds three sources of error: the actual visual precision for extracting the individual values, the ratio computation between these values, and the translation of that 'visual' ratio to a symbolic number (Figure 4, part-to-whole task). Here we explore an alternative measure that captures visual error independently from these other sources.

In three studies, participants briefly viewed a stimulus (bar chart, or dot chart), and redrew the value in a different location. Figure 1 above summarizes the major findings. Participants tended to overestimate small values and underestimate large values. Absolute errors were roughly proportional to the presented value (around 10% on average). Unlike verbal part-to-whole ratio reports (e.g., [6]), the error value did not differ for redrawing values on a shared baseline and on a non-shared baseline. Finally, analyses of signed error show over- and under-estimation biases, where responses were repulsed away from nearby categorical boundaries (e.g., the 50% mark of a reference bar). The novel findings of graphical precision, and response bias may provide new guidance for designers seeking to construct effective visualizations, as well as for formal recommendation systems [25].

- *Caitlyn McColeman is with Northwestern University. E-mail: caitlyn.mccoleman@northwestern.edu.*
- *Mi Feng is with Worcester Polytechnic Institute. E-mail: mfeng2@wpi.edu.*
- *Lane Harrison is with Worcester Polytechnic Institute. E-mail: ltharrison@wpi.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/

2 BACKGROUND

Verbal reports of ratio judgments represent a popular and important way to measure the precision of value extraction from visualized data [6]. In these studies, participants compare values and report "what percentage is the smaller of the larger" for marked values in common graphs such as pie charts, bar charts, and stacked bar charts (Figure 4). These findings reveal a ranking of precision, for example showing that judgments in bar charts lead to higher precision than for pie charts.

However, verbal reports can misrepresent actual percepts. The act of verbalizing a percept can impact what the observers remember seeing, for example, inflating the perceived differences between stimuli belonging to different categories (see [21] for review). Alternative methods, such as the method of constant sum and the method of adjustment used in the current study, have the advantage of not requiring a translation to a verbal representation (see Figure 4).

2.1 Measuring perceptual precision

Perceptual precision can often be captured by Weber's Law [7, 6, 14], where precision is proportional to the value of a presented item. If an observer can reliably see a 1-pixel change in a 10-pixel bar, it should take about 10 pixels of change in a 100-pixel bar to match the same level of performance. This law can capture precision for simple encoding channels like position or length, as well as more abstract judgments of correlation across various types of visualizations, including scatterplots, parallel coordinates, stacked area, bar charts, and line charts [6, 7, 13, 14, 16, 17, 29, 41].

Steven's Power Law allows for non-linearity between the actual value and the impression of that value. The predicted subjective intensity is a function of the stimulus intensity (I), exponentiated by α: I^{α}. For length, α is roughly one, so that a line that is twice as long correctly feels twice as longBrightness has an α of around 0.5, meaning that as an object gets twice as bright, our percept is far less than that. Electric shock has an alpha of 3.5, meaning that doubling the intensity of a shock far more than doubles the sensation [37].

This law fails to predict responses when values are paired with external references, such as grid lines or marks on a measuring cup (e.g. 1/4, 1/2). These additions lead to repulsion effects, such that responses tend to veer away from the values marked by the references. In one study, children indicated the relative proportion between two graphed values using a single-dimensional slider, and responses showed a repulsive effect from zero, 100%, and an implicit 50% division withing the longer line. They consistently overestimated values from 0-25%, underestimated 25-50%, overestimated, 50-75% and underestimated the 75-100% values [36]. The repulsion effect of the reference marks can be fit to a cyclical variant of Steven's Law, where the generalized model treats each indicator as new 'zero' or endpoint. The model was able to account for biases in proportion judgments with a different number of indicators in the response space (for example, marks on a measuring cup; see Figure 2, right) [15].

In the present experiments, the results of the 'alone' conditions (the orange plots in Figure 7), can largely be modeled more simply by Weber's law. But once single values are shown with reference to, or integrated with, other values, cyclical patterns emerge (the green and blue plots in Figure 7).

2.2 Metric judgment is biased by categorical perception

Why do these repulsive effects arise once references are added? One explanation is that we automatically compress metric values by categorizing them [4, 19, 12]. A sound halfway between 'ba' and 'da' is heard as either sound instead of a blend of the two, revealing a repulsion away from a their categorical boundary [22]. Pairs of colors, equidistant in objective space, appear further apart if they cross a verbal category boundary: a yellow-green and a blue-green appear more similar when they are both classified as "green" than a blue-green and a green-blue which are classified as "green" and "blue", respectively [28, 38, 10].

There are similar categorical influences on shape recognition (e.g. faster recognition of a prototypical shape, [1]), size judgment [18], and even perception of emotional facial expressions, such that happy versus sad faces are easier to distinguish than happy versus very happy faces [8](see also Figure 3). For the case of axes next to graphed values, the reference line becomes a category boundary, pushing percepts and memories toward either side. Categorical perception might stem from a repulsion from the boundary (e.g. 0%, 50%, 100%) or attraction toward a category prototype (e.g., 25% is a prototype for 'lower half', similar for 75%). In another study, children make proportion judgments using 1D bar (length), 2D bar, 3D bar, and pie charts [36], by using a 'method of constant sum' response (see Figure 4). Older children responded in a manner that generally mirrored the predictions of Stevens' Law with an exponent of 1, and so performed the proportion judgements well. However, when younger children judged pie charts, they overestimated values 0-24%, underestimated values 25-49%, overestimated 50-74% and underestimated 75-100%. This result suggests repulsion from categorical boundaries such as 0%, 100%, and implicit 50% categorical boundary of 'half the pie'.

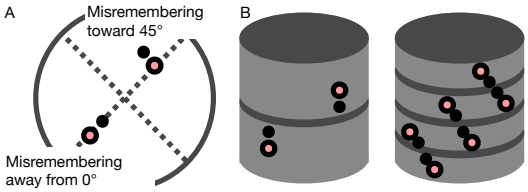

Fig. 2. (A) When people are shown dots in a circle (black dots), they tend to reproduce their positions (pink dots) in locations biased toward the 45 and 135 degree angles, and away from the horizontal/vertical axes [27], even though these references are not shown in the display. (B) When an reference (e.g. a line on a measuring cup) is present in a cylinder to mark 50%, reports of position (either by verbal ratios or moving sliders) are biased away from that indicator, even for multiple references [15]

2.3 Contributions

Some previous work eliminates the verbal report for ratio judgements, instead relying on a linear slider as a way to report ratios as percentages (e.g. [36]). These tasks typically rely on the "method of constant sum" [23] where participants report the value conveyed by two (or more) items as a proportion of a whole. For example, if a pie chart displayed a smaller and a larger slice, participants could say that the smaller slice is 30% of the pie and the larger is 70%, as long as the two values add up to 100%. In one study, participants judged two values displayed in line, bar, 3D bar and pie charts, and reported the proportion of each of the two values using a slider [36] rather than having to verbalize a number.

However, participants still had to translate their impression of the

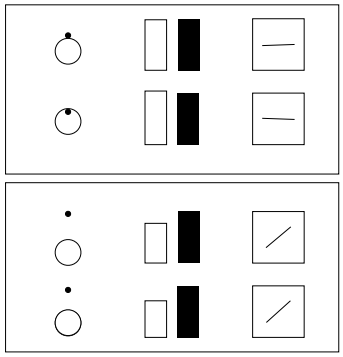

Fig. 3. Categorical perception causes the changes between the pairs in the top panel to be easier to see than changes of the same metric size in the bottom panel, for dot height [20], bar length, and line orientation [18].

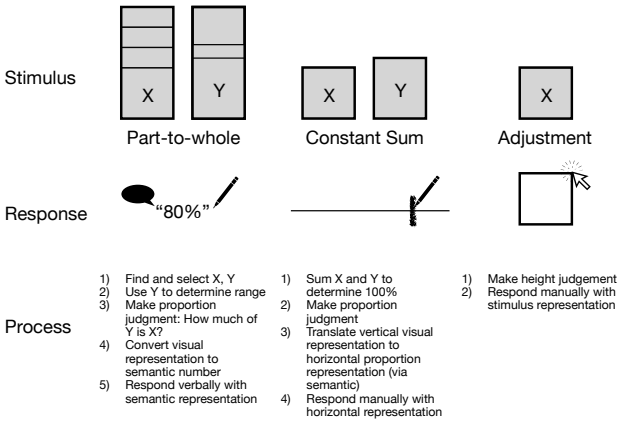

Fig. 4. Various ways to measure participants' perception of data marks, with an example of a participant judging the value of X. The part-to-whole(e.g. [6]) requires more processing steps than the method of constant sum (e.g. [35]) and the method of adjustment, used in the current studies. The method of adjustment allows more specific measurement of the error associated with perception of a single target.

graphical value from its original encoding (the size of the pie slice) to a linear representation for the slider response. The additional translation step likely requires that the viewer's visual system translate that visual magnitude into a domain-general magnitude representation that can also apply to the linear slider, which can introduce variability in the response [9] above and beyond the perception of the pie slice itself. Instead of translating signals to verbal codes or other representations, we test visual representation more directly using the psychophysical "method of adjustment" [11] where participants simply redraw the mark with the mouse.

We also sample more of the ratio space in our presented values than the earlier studies, showing patterns of bias that would be difficult to detect with more granular sampling. In the majority of the reported studies, we presented graphs that represent every 1% within 1-100%, while previous work only showed a portion of the full scale. We found results consistent with this past work, where participants were likely to underestimate data marks representing 25-50% and overestimate marks representing 51-75%, but with a higher resolution sample. We additionally found stronger under- and overestimation patterns in the stacked bar graphs, which had not been previously tested with an adult population.

By adding a dot-only study, we isolate the error pattern for position encodings. When the values represented by bars are higher, the position of their tops become higher, but their length and area also grow. By using a single dot we are able to partial out how much of the observed bias was a function of error in the position encoding compared to the other changing visual features.

We also tested the observed error for the bar alone-condition compared to the reference condition (Figure 6), which is similar in structure to [6], because the value of one bar is perceived in the context of a second neighboring bar, just as the one bar is judged as a proportion of a second bar in their part-to-whole judgement task. In contrast to the method of constant sum, the neighboring bar does not explicitly need to be considered for the participant to redraw their response.

Finally, by integrating the target bar into a stacked bar, the participant can perceive the entire range of values, so that the graphed value is viewed as a proportion of a whole instead of as a independent magnitude. The perceptual experience of viewing a value as a part of a full range may differ from viewing a value as a context-free magnitude. Earlier re-investigations of [6] actually propose the proximity of compared marks as an explanatory factor of why some perceptual tasks (adjacent bars on a common scale) are easier than others (separated bars segments on a misaligned scale) [39]).

Past work has required the translation of "part-to-whole" perception into verbal ratio judgements, where participants must actually state/write a number to represent the presented value. In contrast, the method of adjustment does not require an overt ratio representation.

Work in cognitive and perceptual psychology shows that including context may bias the way that metric values are seen and remembered by adding salient categorical boundaries, e.g., [5, 36, 26, 32, 40, 33]. A better understanding of the visual elements that bias the perception of visualized data may lead to design guidelines that allow displays to convey data in more accurate and less biased ways.

3 Experiment Summary

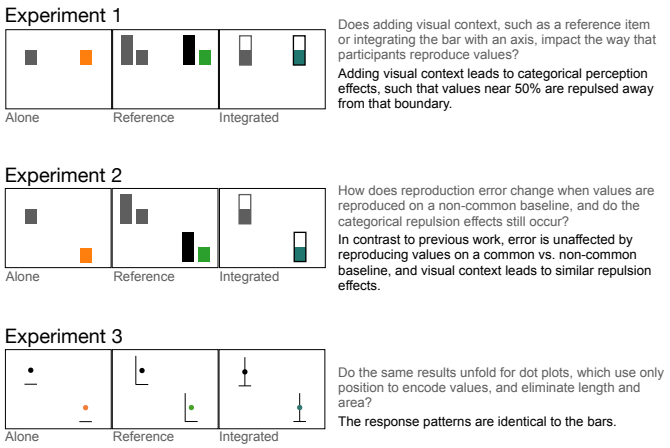

Fig. 5. An overview of the primary questions addressed in this paper, and how experiments answer them.

Through three experiments, we test the impact of added context on error biases by having participants visually recreate a stimulus representing values between 1-99% of a whole. We analyze both absolute and signed errors: unsigned error reflects participants' ability to accurately recreate a presented stimulus, while signed error reflects biases to over- or underestimate the magnitude of a presented stimulus.

3.1 Experiment 1a: Rigid sampling of ratio space

This experiment required the reproduction of values for a single bar (*alone* condition), a value with a reference bar (*reference* condition), or as a stacked bar (*integrated* condition), with values chosen from an even sampling of the ratio space.

3.1.1 Experiment Procedure

Before this and all subsequent experiments, the participant provided informed consent and participated in a brief instruction sequence that included text, images, and a video that briefly described their task. Participants viewed a stimulus (0.5 seconds) that presented a value between 1% and 100%, before recreating the value from memory (0.5 seconds later) in a different location on the screen (Figure 6). Participants redrew the initial bar (or dot, in Experiment 3) by adjusting the height of a response bar (or dot, in Experiment 3) with their mouse. The bar height was adjusted either by clicking above an initial dash shape (which indicated the response location (Figure 6, top) or by dragging up from that dash. The height of the bar after their last adjustment (before they pressed the spacebar to advance) was accepted as their drawn value. For all three experiments, only the height of the mark changes to reflect ratio value. The bars are always vertically oriented. After all trials in the experiment are completed, participants provided additional comments and were debriefed. De-identified data from our studies are available at https://osf.io/sdt2b/).

Experiment 1a Methods Experiment 1a tested participants' ability to recreate bars in different context conditions (alone, reference, integrated) using a within-subjects design. Participants performed three blocks, such that they saw 33 trials from each of the three context conditions (see Figure 6). The block order was counterbalanced, to mitigate possible order effects.

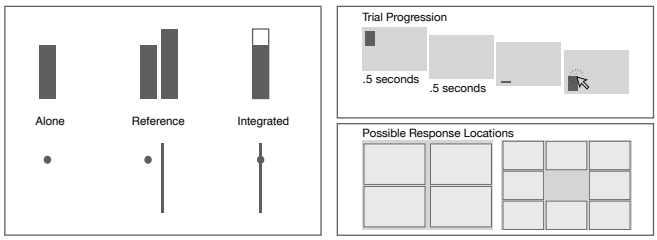

Fig. 6. The different experimental conditions (left) and how they were tested during the experiments. During a trial, participant saw a stimulus, it disappeared, and then they redrew it in a different part of the screen. Experiment 1, 2b and 3b used 4a possible locations for the stimuli and response. Experiment 3a and 4a used more of the screen, and had 8 possible redraw locations.

In Experiments 1a and 1b, The participants always redrew the stimulus in a position shifted horizontally from the originally presented bar. The value presented on each trial was from 1..4..7..100%, in a randomized order, but blocked by condition (alone, reference, integrated, Figure 6 with the blocked condition order counterbalanced between participants)

3.1.2 Online Experiment Methods

For this experiment, and all other online experiments (Experiment 1b, 2b, 3b), we recruited participants on Mechanical Turk. Workers received $2.25 USD (estimation completion time <20 minutes with a target remuneration of $10 US per hour).

The stimulus display canvas was adaptive to the size of the participant's browser window. Canvas dimensions were required to exceed a minimum (900 x 600 px) size to proceed. If the canvas exceeded a maximum size (1680 x 1,050 px) size, the experiment was displayed in the top-center of the browser, with the additional space left blank. The background was neutral grey $RGB(189,189,189)$

The relative size of elements (e.g., the bars) was also adaptive to the canvas size. Specifically, we defined an adaptive *canvas unit* as proportional to the canvas height, $CanvasHeight$ / $Max_{CanvasHeight}$ x C, where $Max_{CanvasHeight}$ is 1,050 px, and C = 2.1 is a constant multiplier. Height was chosen over width as browser height is generally more stable than browser width (i.e., consumer screens typically vary in width more than height). We computed the dimensions of the bar, dot, and line stimuli based on this canvas unit.

Experiment 1a Results One participant was excluded from analysis (but still compensated) for having a median error exceeding 10% across all trials, leaving 35 participants. Trials with unsigned error exceeding 3 standard deviations were excluded (1.3% of total).

We computed unsigned error as a measure of accuracy, proportion-corrected unsigned error to attempt to use Weber's Law to reduce performance across values to a single percentage, and signed error to measure over- and under-estimation biases. We analyzed all three measures with ANOVAs to check for differences between bar types. The interpretations are Bonferroni corrected, such that the p-value must be less than .05/3 to yield a 'significant' result. The following results are visually summarized in Figure 8.

There was a significant difference between the conditions' unsigned error, as indicated by a within subjects ANOVA, $F(2,68) = 43.38$, $p < .001$, $\eta^2 = .202$. No difference was detected in the proportion corrected data (the unsigned error divided by the presented value), $F(2.68) = 1.85$, $p = .17$. An omnibus ANOVA did not suggest a difference between the three groups' overall signed error, $F(2,68) = 1.85$, $p = 0.165$ (Figures 7 and 8).

We expected only the reference and integrated to display a categorical repulsion with values approaching 50% being underestimated and values exceeding 50% being overestimated, because only those conditions allow the viewer to notice the 100% highest value, and the implicit 50% mark for that value. The presented values from 25-49% were contrasted with the presented values from 51-75% to test whether signed error reliably changed between these ranges. We calculated the median error for each participant for each quadrant and performed a t-test on the means of the medians. We used the difference of the second and third quartiles to test for repulsion for this and all of the remaining experiments in the current study.

For this, and all subsequent experiments, the results of the paired t-test are shown in Figure 7: an arrow shows the direction of the difference for significant tests. In the Alone condition, Q2 (mean = .271) was higher than Q3 (mean = -2.36), $t = 2.38, D = .57, p = .021$, after error correction. There was no difference in the Reference condition between Q2 (mean = -1.58) and Q3 (mean = -2.25), $t = -.06, D = .01, p = .952$. There was a significant difference between Q2 (mean = -1.07) and Q3 (mean = .655) error in the Integrated condition, $t = -4.25, D = 1.02, p < .001$, suggesting that the stacked bar graph elicits a repulsion from 50%. Participants typically underestimate Q2 values and overestimate Q3 values.

Finally, to test for the bias patterns at the extreme ends of the ratio space, we compare the mean bias values for the first (Q1: $< 25\%$) and fourth (Q4: $> 75\%$) quartiles. This analysis is exploratory, to test whether values at the ends of the ratio space are differently impacted by including context. The test does show a difference for the Alone condition, $t(38) = 2.79, p = .008$) between Q1 (mean = 1.3) and Q4 (mean = -2.19) and Reference condition, Q1 (mean = 1.13), Q4 (mean = -1.89), $t(67.6) = 5.38, p =< .001$. The bias differs between Q1 (mean = 1.06) and Q4 (mean = -1.11) in the Integrated condition, $t(66) = -4.67, p < .001$. This is taken as preliminary evidence that lower values (Q1) are overestimated compared to higher values (Q4) which are underestimated.

3.2 Experiment 1b: Ratio judgments and categorical perception

In Experiment 1a we used an even sampling of the ratio space to develop a base understanding of how response error changes as a function of the graphed value presented to participants either as a single bar (alone), with a reference bar (reference) or as a stacked bar (integrated).

The randomly presented values in Experiment 1b ensured that the patterns of results we observed in Experiment 1a were not somehow tied to its consistently-sampled ratio values, and allows us to uncover potentially peculiarities of specific untested ratio values (e.g. 45%). Perhaps most importantly, random sampling allowed us to observe additional values around 50%, where the apparent repulsion from the halfway mark was observed in Experiment 1a, compared to the rest of the ratio response range.

Experiment 1b Design Experiment 1b employed a within-subjects design. Participants performed three blocks, such that they saw thirty trials from each of the three conditions: alone, reference and integrated (see Figure 6). The block order was counter-balanced, to avoid for order effects in participants' responses. The value presented on each trial was randomly selected for each participant from 1-99%. The task was identical to the task in Experiment 1a, with the exception of the random sampling procedure.

Experiment 1b Results 27 participants were recruited from Mechanical Turk, and 24 of them met our inclusion criterion of <10% median unsigned error for all trials in the experiment. Trials were excluded from analysis if the unsigned error exceeded 3 standard deviations from the mean. Only 0.04% of trials met this criterion, and the remaining trials were analyzed as described below.

A within subject ANOVA compared the unsigned error between conditions. The Alone condition showed a higher unsigned error than the others, $F(2,50) = 15.359, p =< .001, \eta^2 = .225$). The proportion correct error was not statistically significant, F(2,50) = 2.94, p = .06, although there was some interesting variation observed in Figure 7. A final within subjects' ANOVA did not reveal a difference between the conditions for the overall signed error value, $F(2,50) = .211$, $p = 0.810$ (see Figure 8.)

To test whether there was a bias in the signed response around the 50% mark, we test each subjects' median score between 25-49% and

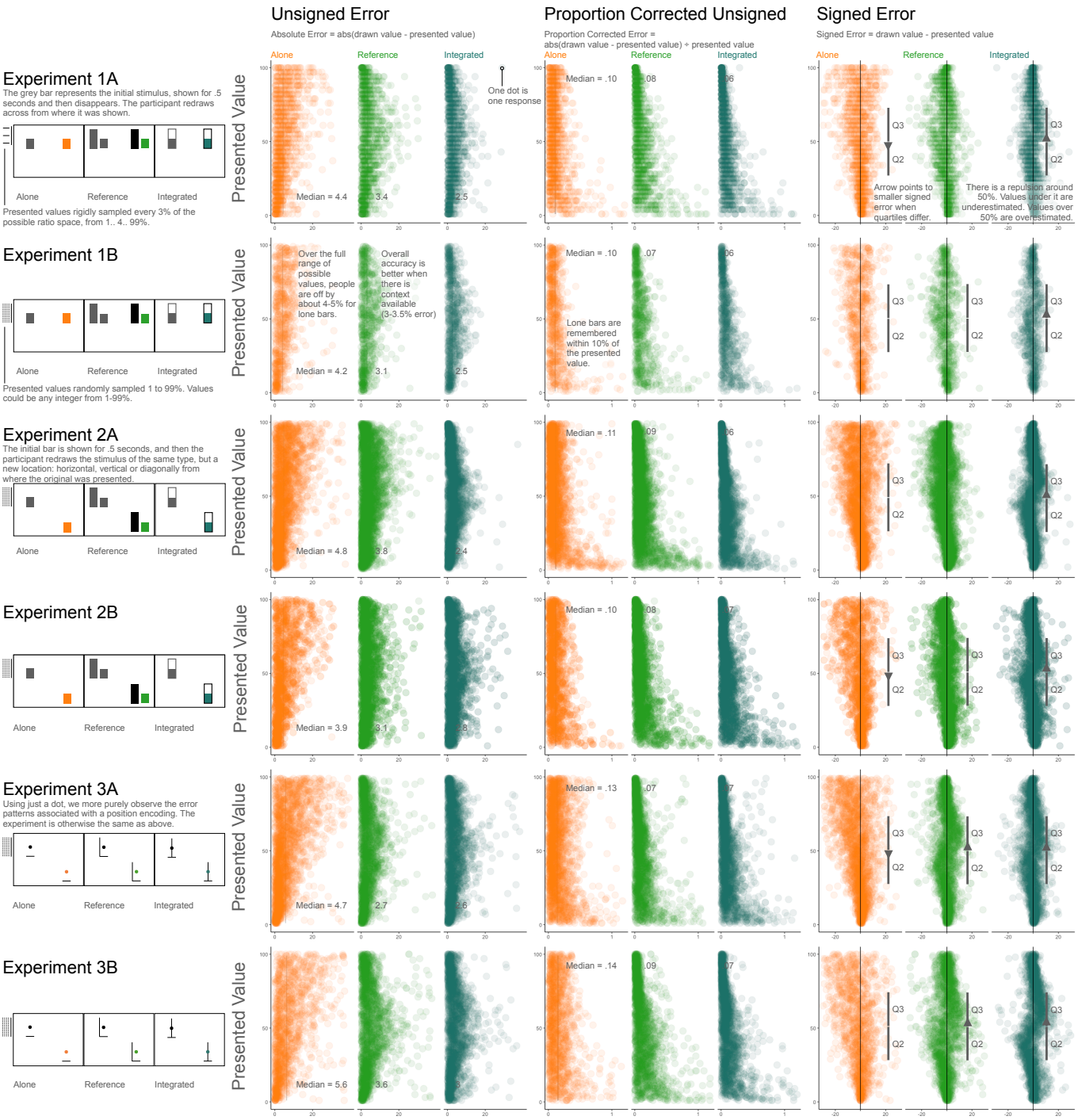

Fig. 7. Absolute, proportion corrected and signed error for all experiments. The experiment panel (left) shows schematics of the manipulations in each experiment. The grey mark reflects the initial stimulus presented to participants. The colorful mark represents what (and where) a participant would redraw the initial stimulus, which participants drew in the same grey as the initial stimulus but is colorized here for illustration.
Each colored dot represents a participant's response for one trial in the study. Unsigned error is the distance between the participant response and the presented value. Proportion Corrected Error is the the difference between participant response and the correct answer in a single trial, divided by the presented value. Signed error reflects under- and overestimation in participant responses.
Long vertical grey lines represent the condition medians for unsigned and proportion corrected error. Short grey lines encompass the second and third quartiles of presented values. Q2 represents the second quartile of presented values (25-49%) and Q3 represents the third quartile of presented values (51-75%). Arrow heads reflect the direction of change when Q2 and Q3 are significantly different. In the Signed Error data (right) we observe an apparent repulsion from the 50% mark, wherein participants underestimate values approaching 50% in Q2 and overestimate values exceeding 50% in Q3.

between 51-75%. There were no significant differences in the signed error for the Alone condition (Q2 mean = -.636; Q3 mean = -1.23, $t(46) = -.56, D = .156, p = .579$) or the Reference condition (Q2 mean = -1.42; Q3 mean = -2.86, $t(41.3) = 1.16, D = .321, p = .254$). But in the Integrated condition, bias values when redrawing presented values in Quartile 2 (mean = -1.92) were lower than presented values in Quartile 3 (mean = -.481), $t(43.2) = -2.54, D = .705, p = .015$. The difference between Q2 and Q3 in the Integrated condition replicate the pattern observed in Experiment 1a. Values less than 50% were be underestimated more than values over 50% which tend to be overestimated.

To determine if there were differences between the biases from Q1 and Q4, the error observed for each of the quadrants was compared with t-tests. In the Alone condition, Q1 (mean = 1.66) and was greater than Q4 (mean = -2.66), $t(13) = 2.84, p = .014$. Reference condition (between Q1 (mean = 1.71) and Q4 (mean = -4.04), $t(15) = 6.44, p < .001$) and the Integrated condition (between Q1 (mean = .068) and Q4 (mean = -1.26), $t(15) = 2.29, p = .04$). Lower values (<25%)are overestimated and the higher values (>75%) are underestimated.

Experiment 1 Discussion Adding context to a bar reveals the full range of possible values (0-100%), such that height of the bar now reflects the percentage of a greater whole. We found that the integration of the graphed value with a stacked bar invoked a perceptual bias away from 50% (Figure 7, right column). In Experiments 1a and 1b, we constrained the re-draw location to a horizontal translation, since position on an aligned scale produces the most precise encoding when measuring verbal reports of ratios [7]. Participants showed increasing unsigned error for higher values, and a signed error showed a repulsion around 50% (see Figure 7, Experiment 1).

Spatial Translation Experiment 2 moves from a constrained task of redrawing the bar across from where it originally appeared, to a task where redraw location was less predictable. The goals of Experiment 2 are to conceptually replicate Experiment 1 (test the role of reference bars and integrated bars on error and perceptual bias) and to study whether the observed patterns generalize beyond the simple horizontal translation and into non-aligned baselines.

Experiment 2a is conducted in the laboratory, on a fixed set of machines to afford a more controlled set-up (e.g., monitor and display sizes held constant, viewing distant held constant). Experiment 2b is a between-subjects design, which helps to establish that the observed perceptual biases in the Integrated condition are not influenced by other condition blocks (e.g., to rule out learning effects, fatigue). The between-subjects design has the added benefit of testing on a remote population, so as to support the generalizability of our findings.

3.3 Experiment 2: Bar translation and context

Experiment 2a Design Experiment 2a employed a within-subjects design, where all participants observed 99 trials per condition (representing 1-99% of the ratio space). The conditions were blocked, and presented in counter-balanced order between participants to avoid possible biasing or learning effects in the responses.

The task for Experiment 2a is similar to the procedure in Experiment 1. Participants see a bar on the screen for 0.5 seconds, a blank screen for 0.5 sec, and then are shown a small line to indicate where to reproduce the stimulus in one of eight cued locations (Figure 6).

Experiment 2a Results Eighteen undergraduate participants performed the task in exchange for partial course credit. All 18 met the inclusion criterion of <10% median unsigned error over the all of their trials. All but 2.1% of trials were viable (unsigned error within 3 standard deviations of the mean).

In Experiment 2a, there was a significant difference between the groups' unsigned error over all of the presented values, $F(2,16) = 30.556$, $p < .001$, $\eta^2 = .275$. The proportion corrected error revealed a difference between conditions, $F(2,16) = 5.74$, $p = .01$, $\eta^2 = .120$, and there was also a difference between the groups' signed error, $F(2,16) = 4.226$, $p = .034$.

To test for categorical repulsion, we conducted a set of Bonferroni-corrected t-tests for signed error corresponding to the Quartile 2 of

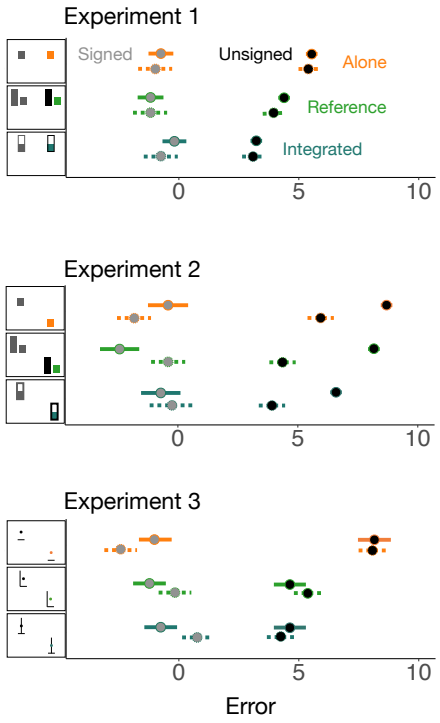

Fig. 8. The mean error for all presented values (0-100%). The solid line indicates Experiment A; the dashed line indicates Experiment B. The context conditions, alone, reference and integrated, are shown by differently colored marks. Mean unsigned error is shown by light circles, and mean signed error is shown by dark circles. Error bars reflect Fisher's Least Significant Difference. Notice that the integrated condition has lower unsigned error than the reference and alone conditions, over the full range of presented values for all three experiments.

presented values (25-49%) and the Quartile 3 of presented values (51-75%). The Alone condition showed no difference between Q2 (mean = -2.11) and (Q3 mean = -1.99), $t(13) = .24, D = .060, p = .816$. The Reference condition also showed no difference between Q2 (mean = -2.33) and (Q3 mean = -.725), $t(15) = 1.02, D = .378, p = .322$. The t-tests conducted on the signed error for the Integrated condition, however, did show a significant difference between Q2 (mean = -2.64) and Q3 (mean = 1.26), $t(13) = -6.68, D = 2.573, p < .001$. As shown in Figure 7, the values approaching 50% are underestimated and the values exceeding 50% are overestimated.

Experiment 2b Design Experiment 2b participants were assigned to one of three possible between-subjects conditions (alone, reference or integrated). The between-subjects design allowed us to gather enough trials to cover the full response space, and while keeping full experiment time to a reasonable duration (approximately 20 minutes) for an online task.

Participants saw the full range of values representing integers from 1 .. 99, but only for the graph type of the condition that they were assigned (alone, reference or integrated).

Experiment 2b Results There were 60 participants recruited from Amazon Mechanical Turk. Seven workers failed to meet the inclusion criteria of a median unsigned error <10% and were excluded from subsequent analyses. Three additional workers were excluded for repeatedly responding so rapidly (<500ms) that the logging system failed.

As in Experiment 2a, Experiment 2b showed group differences in unsigned error through a between subjects ANOVA, $F(2,47) = 6.69$, $p = 0.003$, $\eta^2 = 0.22$ but not in proportion corrected error, $F(2,47) = .35, p = .71$, or signed error, $F(2,47) = 2.04$, $p = 0.142$.

Paired t-tests show differences between Q2 (mean = -.401) and Q3 (mean -3.24) for the Alone condition; such that the error becomes

more negative as the presented value increases, $t(21.7) = 1.66, D = .629, p = .010$. There's no detectable difference in the Reference condition, between Q2 (mean = .204) and Q3 mean (-3.24), $t(31.4) = .92, D = .307 p = .364$. As with the above experiments, the Integrated condition error is lower in Q2 (mean = -1.13) than Q3 (mean = .363), $t(34) = -2.83, D = .942, p = .008$. See Figure 7 to observe the error patterns, which provide further evidence for categorical repulsion around 50%. Signed error switches from underestimating presented values in Quartile 2 (presented values from 25-49%) to overestimating values in Quartile 3 (51-75%) of possible ratio values.

We compared Q1 and Q4 with a set of exploratory t-tests find whether small values (Q1) were overestimated relative to large values (Q4). In the Alone condition, small values (Q1 mean = .897) were overestimated relative to than high values (Q4 mean = -4.38), $(t(13.5) = 2.85, p = .013)$. The same pattern is observed in the Reference condition: Q1 mean = 2.6, Q4 mean = -2.71 $(t(34) = 7.66, p < .001)$. In the Integrated condition, as well, low values (Q1 mean = 1.77) are overestimated and high values (Q4 mean = -1.62) are underestimated, $(t(31.6) = 5.15, p < .001)$.

Experiment 2b asked participants to redraw the bar across from, vertically from, or diagonally from the original stimulus. Given earlier reports that position along a non-common baseline is associated with lower perceptual precision [6], we expected the bars drawn horizontally from the original stimulus (common baseline, mean = -.6, 95% C.I. = [-1.2, -0.04]) would be more precisely reproduced. This was not the case. A linear mixed effects model was fit to the proportion-corrected error, where the best predictors were the presented value, and the resulting model output was subjected to post hoc comparisons. The common baseline (horizontal redraw) was no different from the diagonal redraw location (z = -1.96, p =.12; mean = -1.9, C.I. [-2.4, -1.2]) or from the vertical redraw location (z =.36, p =.93, mean = -2.7, C.I. = [-3.4,-2.1]). It appears that the common baseline has far less impact on participants' ability to redraw bars from memory than it does on part-to-whole judgment tasks.

Experiment 2 Discussion In Experiment 2, we also prompted participants to redraw the presented value in different quadrants of the screen. It was expected that redrawing the item diagonally or vertically from where it was originally presented would increase error, but we found that was not the case, which could challenge the importance of a common baseline in the ranking of basic perceptual tasks [6].

Experiment 2 replicated and extended Experiment 1. We found again that context (the integration condition) invoked an apparent categorical bias. It sampled the full range of 0-100% in both the laboratory and online environments, which allows us to observe error over the full range of ratio values (Figure 7), improving confidence in earlier conclusions that a) larger bars are underestimated, b) error is approximately 10% of the presented value and c) categorical repulsion was present with the addition of context that conveys the full range of possible values.

3.4 Experiment 3: Position translation and context

The experiments to this point have tested the error and perceptual biases observed in different kinds of bar graphs and at different sampling rates. To isolate position encodings, we shifted from bars to simple dots. This reduces the likelihood that the error patterns and biases we observed above arose due to area and/or length, which also changed with the height of a bar. Aside from the switch from bar encodings to dot encodings, Experiment 3 was structured identically to Experiment 2.

Experiment 3a Methods Experiment 2a employed a within-subjects design, where all participants observed 99 trials per condition (representing 1-99% of the ratio space). The conditions were blocked, and presented in counter-balanced order between participants to avoid learning effects.

The task was the same as the above experiments: participants redrew a presented stimulus in a different location from where it was originally presented.

Experiment 3a Results Twenty-three undergraduate participants performed the task in Experiment 3a in exchange for partial course credit. All 23 met the overall inclusion criterion (median unsigned error <10% over the whole experiment). All but 0.6% of their trials were analyzed (they met our inclusion criterion of being within 3 standard deviations of the mean unsigned error).

To test whether the group means differed, errors (absolute and signed) were each subjected to an ANOVA (Figure 8). Due to the similarity between the tests, the familywise corrected alpha is .05/2 = 0.025 experiments.

The ANOVA on unsigned error revealed a difference between groups in Experiment 3a, $F(2,38) = 22.188$, p <0.001, η^2 = 0.363. No difference was observable in the proportion corrected error data, $F(2,38) = 1.00, p = .38$, but there was a difference for Experiment 3a, $F(2,38)$ =8.35, p =0.001, η^2 = .211.

T-tests on the median signed error values show differences between all three of the Alone, Reference, and Integrated conditions (Table 1), but in different directions. The signed error in the alone condition appears more negative as the presented values increase, consistent with underestimation of higher values. The reference and integrated conditions, however, show a different direction of change, consistent with categorical repulsion at 50%, since the signed error in associated with Quartile 2 values is less than the signed error in the Quartile 3 values (see Figure 7).

The remaining presented values, for ratio values of <25% and > 75%, were compared using a set of t-tests. The Alone condition showed a higher mean error in Q1 (mean = 2.19) than Q4 (mean = -9.41), $t(21) = 5.85, p < .001$. The Reference condition also showed the same pattern, Q1 (mean = 1.54) was greater than Q4 (mean = -1.53), $t(20) = -6.14, p < .001$; the Integrated condition Q1 (mean = 1.65) was greater than Q4 (mean = -2.02), $t(21) = -3.86, p = .001$. Lower values are subject to overestimation while higher values are underestimated.

Experiment 3b Methods Experiment 3b was identical to Experiment 2b, except that it tested dot encodings instead of bars. As in Experiment 2b, participants were recruited from Amazon Mechanical Turk and saw only one of three possible conditions because this experiment employed a between subjects design (Figure 4, dot stimuli.)

Experiment 3b Results Fifty-nine participants were recruited for Experiment 3b from Mechanical Turk. 54 of them met the inclusion criterion of <10% median unsigned error over all trials. All but 2% of the remaining trials met the additional criterion of error <3 SD of the mean.

The unsigned error data from Experiment 3b was subjected to an ANOVA, which showed difference between conditions, $F(2,51) = 18.948$, p <0.001, η^2 = 0.426 (Figure 8). No difference was observed for the proportion corrected error, $F(2,51) = 2.58, p = .09$, but there was a difference in signed error, F(2,51) = 11.36, p <0.001, η^2 = 0.308.

Table 1 shows the results of the t-tests run on the signed error. The pattern mirrors the results observed in Experiment 3A, but the between-subjects design has less statistical power. We find a significant difference in the Integrated condition, where signed errors approaching for presented values approaching 50% are lower than errors for presented values exceeding 50%.

In Experiment 3b, we constrained the possible re-draw locations to quadrants of the screen such that participants could draw across from, vertically from, or diagonally from the original bar. The experiment design was the same as Experiment 2b, where we tested the idea that non-common baselines would yield worse recall. In Experiment 2b, however, the stimuli were full bars. Experiment 3b reproduction accuracy and bias for *position* per se, because the dot plot isolates positon from length and area encodings. A linear mixed effects model fit the participants' proportional error with condition (alone, reference, or integrated) and redraw location as predictors. Post hoc tests showed no difference when re-drawing the graph horizontally from the original, on the same baseline versus vertically (z = -.05, p =.998); horizontally versus diagonally (z =.476, p =.883); or vertically versus diagonally

Table 1. Experiment 3: differences in signed error between presented values Q2:25-49% and Q3:51-75% for the alone, reference and integrated dot charts.

Condition	Q2 Mean	Q3 Mean	t	D	p
3A: Alone	-.760	-5.20	4.728	.866	<.001
3A: Ref.	-1.18	1.51	-6.14	1.58	<.001
3A: Int.	-.983	.784	-3.86	1.26	.001
3B: Alone	.011	-4.30	3.080	1.036	.009
3B: Ref.	-1.39	1.40	-3.612	.1.33	.002
3B: Int.	-2.18	3.42	-7.461	2.23	<.001

($z = .422$, $p = .906$). This post-hoc finding suggests that a position on a common baseline may not be as critical outside of a verbal ratio judgment task.

Experiment 3 Discussion Experiment 3 shows the same patterns of errors as Experiments 1 and 2 (see Figure 7). Unsigned error is higher with the dot-only encoding and response, suggesting some advantage of the bars beyond the position encoding itself. The height of the mark (dot or bar) relative to a range of invokes a bias to underestimate values approaching 50% overestimate values exceeding 50%.

Additionally, the pattern of low value overestimation and high value underestimation was observed again. In the Alone condition, presented values up to 25% were overestimated (mean = 2.19) compared to presented values exceeding 75% (mean = -9.41) , $t(21) = 5.85, p < .001$. Reference Q1 (mean = 1.54) presented values were overestimated compared to Q4 (mean = -1.53), $t(20) = -6.14, p < .001$. The same was observed in the Integrated condition: Q1 (mean = 1.65) was greater than Q4 (mean = -2.02), $t(20) = 3.5, p = .002$.

Fig. 9. The median proportion corrected error averaged over all trials in the experiments. Proportion corrected error is the unsigned error divided by the presented value, as shown in Figure 7. The Error bars reflect the interquartile ranges. The grey band represents the interquartile range of proportion corrected errors observed in a replication [14] of [6]. The error observed in the current studies is comparable to the error observed in Cleveland and McGill's most precise encodings.

4 Categorical Perception Can Amplify Precision

The overall error through Experiments 1-3 is lower in the integrated condition: on average, for all of the presented values 1-99%, the participants are more accurate (Figure 8). But this reduced error comes at the cost of bias. When the graphed value is integrated into a stacked bar (e.g., column 3; Figure 7), the error is also biased around 50% such that values below are underestimated and values above are overestimated.

More generally, categorization is the gateway for humans to increase their capacity for metric values, by dividing an otherwise continuous value into smaller and more numerous discrete categories. The 50% mark of a nearby bar can be treated like a gridline, such that categorizing a value as being above or below that mark provides a single bit of information about the value. When colors are presented as a continuous rainbow, human observers see them as a graded collection of ordered categories. The difference between the bluest blue and the greenest green feels larger than the difference between the value on the blue-green boundary and the value on the green-yellow boundary even if each pair is objectively the same distance in colorspace, because our perceptions of those values are differentially impacted by their category membership [19, 28]. In data visualization, multi-hue color heatmaps allow finer resolution of the number of values in the data, though with the caveat that those boundaries create perceived categories where none exist in the data [31, 3].

A seminal paper in cognitive psychology, 'The magical number seven, plus or minus two: Some limits on our capacity for processing information' [24] proposes that higher perceptual precision primarily stems from being able to discriminate among more categories. People are able to categorize positions, lengths, loudness, and even levels saltiness of water into 5-9 categories (hence the 'magical' number 7). But in some cases, like the position of a dot in a square, or color discrimination, people can beat that limit substantially. Miller suggested that being able to divide a whole into more categories, like the four quadrants of the square, or breaking hue down into "reds" and "blues", is the key to amplifying human ability to categorize with more precision. The position of a dot on a horizontal line is likely aided by categorizing the line into "left half" or "right half", and potentially even quarters as well (see Figure 2). In the present experiments, participants likely developed a natural category boundary of the halfway point of an reference bar or axis. We found that this boundary improved precision, but at the cost of biasing perception by either repulsing estimates away from that boundary, or perhaps attracting estimates toward category prototypes (25%, 75%).

5 Limitations and Open Questions

Relative to earlier reports [6, 14, 34], one advantage of our study design is that we trade out verbal translation error, but this comes at the cost of introducing motor and memory error (and bias). Verbal representations of numbers and ratios could be even *more* prone to categorical biases. Past work shows that suppressing verbal coding (by having someone repeat numbers or nonsense syllables) can suppress categorical perception, by reducing the influence of categorical linguistic descriptions [30]. For example, according to this past work, in the current task, saying "just above halfway" to yourself while remembering the position of a bar should amplify the categorical repulsion effect, but performing the same task (e.g., recreating a previously seen bar graph) while the verbal system is overloaded (e.g., a dual-task of repeating nonsense syllables) should reduce it.

Response biases arise with the introduction of context (see Figure 7, integrated condition, rightmost column). Although the absolute error is lower overall when the value is integrated with its axes, this advantage trades off with the introduction of systematic bias for values around 50%. Values between 25-49% are underestimated and values between 51-75% are overestimated. This bias might inflate the difference between values on the opposite sides of that category boundary. For example, if the approval numbers for two political candidates are 48% and 52%, displaying the data with a reference to the full range of values may cause those two numbers to appear more different than 4%. Because the present studies only test reproductions of individual values, future work would need to verify this bias in a comparison of differences between two values – but the examples in Figure 3 suggest that the bias would appear.

Future studies could document error patterns associated with recreating values from 1-99% for other visual data encoding variables [2] including area, orientation, intensity, etc. Would their percent error value be substantially less than the 10% observed here? Would they also experience repulsion from added context from nearby values or examples in a legend? We anticipate a similar consequence of context on these other visual variables: that indications of a full range of values make it possible to perceive categorically which can increase

precision but also introduce bias.

This paper constrains the graphical perception task to a simple magnitude recreation task, relying upon the method of adjustment to capture participants' perception. In this method of adjustment paradigm, we observe no large differences in error when participant draw vertical bars horizontally from the initial presented value (common baseline) versus when they draw diagonally or vertically (non-common baseline) from the original (bar or dot) graph (Figure 9). This is contrary to previous reports of data showing that position on a common baseline is more precise. We expected to observe the same baseline advantage in the horizontal redraw conditions. Our failure to replicate this advantage may be due to a) the method of adjustment versus the part-to-whole judgement b) relying on memory to draw stimuli eliminating the common baseline advantage or c) another unaccounted-for variable that could impact how we rank the visual variables.

6 Considering Bias in Visualization Design

The present data show that marks in the presence of context cues are subject to bias, due to the influence of categorical perception. The results from this work may be a useful first step in providing debiasing guidelines for designers.

When choosing a visualization type, practitioners balance multiple constraints, including visualization affordances (e.g. a bar graph for absolute values, a pie graph for percentages), aesthetic considerations, and space efficiency. The biases observed here can be included as one of these constraints. If precision is critical, practitioners may wish to avoid presenting nearly-equal values in a stacked bar graph, since values around the 50/50% mark are vulnerable to bias. Instead, those data may be better presented in reduced-context environments, such as stand-alone bars, where each mark is less susceptible to bias. Alternatively, if the important values are high (e.g. 95%) ratio values, then using a stacked bar graph may reduce error. In this scenario, the 100% mark (the full range marker) serves to reinitialize the percept. Functionally, in a stacked bar graph, 5% and 95% are both 5% away from a starting point. The perception of the displayed value can be informed by the end point of the range, in contrast to a lone bar, where the perception of the displayed value can rely only on the start point of the range (0%).

We observe reliable patterns in response errors. A single bar is overestimated when it represents a small value, but is underestimated when it represents a large value. Responses are also subject to categorical repulsion when visual context is present to indicate the full range (0-100%) of possible values. The repulsion can be observed in the signed error, shown in the right panel of Figure 7. An up arrow means that the error in Quartile 2 (25-49%) was less than the error in Quartile 3 (51-75%). A down error means the opposite: error was lower for Quartile 3 than Quartile 2 (consistent with underestimation of increasing values). For every experiment, the context (integrated condition) invokes categorical perception.

When designing data displays and prioritizing precision, the number of tick marks is an important consideration. Too few, and the observer is unable to find a visual reference to help interpret a position encoding. But as more are added, marks will be (on average) closer to ticks, causing increasing repulsion of their remembered positions. As references (0% and 100%) improve accuracy in the present data, adding grid lines and tick marks likely similarly increases precision, which may be worth the price of bias (Figure 7).

Conclusion Visualizations using position encodings (e.g., bar charts or dot plots) are said to be the most precise. While position may be the most precise visual variable to encode data, studies have shown that position encodings can also invite bias [5, 40]. The present experiments show that the precision of position is modulated by local context, for example that reference marks "repulsed" reports of data value away from the 50% mark (Figure 7). These observed bias patterns presents a trade-off with precision: while the reference and integrated conditions invite bias, the presence of that bias appears to be due to the influence of perceptual categories that, in the end, decrease the overall error observed. As reference marks, containment, and other visual components are pervasive features in visualization design, the results of these experiments and the methodologies used to explore them lay needed groundwork to further explore how categorical perception and biases shape how people process visualizations.

Acknowledgments

Thanks to Evan Anderson, Cristina Ceja, Elsie Lee and Chase Stokes for their comments on an early version of the paper and Enrico Bertini for his feedback on the experiment design. This work was supported in part by grant #IIS-1901485 from the National Science Foundation.

References

[1] O. Amir, I. Biederman, S. B. Herald, M. P. Shah, and T. H. Mintz. Greater sensitivity to nonaccidental than metric shape properties in preschool children. *Vision Research*, 97:83–88, 2014.

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

[3] D. Borland and R. M. T. Ii. Rainbow color map (still) considered harmful. *IEEE computer graphics and applications*, 27(2):14–17, 2007.

[4] M. H. Bornstein and N. O. Korda. Discrimination and matching within and between hues measured by reaction times: Some implications for categorical perception and levels of information processing. *Psychological research*, 46(3):207–222, 1984.

[5] C. Ceja, C. McColeman, X. C, and S. Franconeri. Truth or square: Aspect ratio biases recall of position encodings. *IEEE Transactions on Visualization and Computer Graphics*, (2020).

[6] 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.

[7] W. S. Cleveland and R. McGill. An experiment in graphical perception. *International Journal of Man-Machine Studies*, 25(5):491–500, 1986.

[8] N. L. Etcoff and J. J. Magee. Categorical perception of facial expressions. *Cognition*, 44(3):227–240, 1992.

[9] M. Fabbri, J. Cancellieri, and V. Natale. The a theory of magnitude (atom) model in temporal perception and reproduction tasks. *Acta Psychologica*, 139(1):111–123, 2012.

[10] M. W. Fang, M. W. Becker, and T. Liu. Attention to colors induces surround suppression at category boundaries. *Scientific reports*, 9(1):1443, 2019.

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

[12] R. L. Goldstone and A. T. Hendrickson. Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):69–78, 2010.

[13] 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.

[14] 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.

[15] J. Hollands and B. P. Dyre. Bias in proportion judgments: the cyclical power model. *Psychological review*, 107(3):500, 2000.

[16] 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.

[17] 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, 2016.

[18] S. M. Kosslyn, G. L. Murphy, M. E. Bemesderfer, and K. J. Feinstein. Category and continuum in mental comparisons. *Journal of Experimental Psychology: General*, 106(4):341, 1977.

[19] A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358, 1957.

[20] A. Lovett and S. L. Franconeri. Topological relations between objects are categorically coded. *Psychological science*, 28(10):1408–1418, 2017.

[21] G. Lupyan. Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in psychology*, 3:54, 2012.

[22] K. S. MacKain, C. T. Best, and W. Strange. Categorical perception of english/r/and/l/by japanese bilinguals. *Applied Psycholinguistics*, 2(4):369–390, 1981.

[23] M. Metfessel. A proposal for quantitative reporting of comparative judgments. *The Journal of psychology*, 24(2):229–235, 1947.

[24] 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.

[25] 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, 2019.

[26] F. Müller-Lyer. Zur lehre von den optischen täuschungen. *Über Kontrast und Konfluxion. Zeitschrififiir Psychologie und Physiologie der Sinnesorgane, IX*, pages 1–16, 1896.

[27] N. Newcombe, J. Huttenlocher, E. Sandberg, E. Lie, and S. Johnson. What do misestimations and asymmetries in spatial judgement indicate about spatial representation? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4):986, 1999.

[28] T. Regier and P. Kay. Language, thought, and color: Whorf was half right. *Trends in cognitive sciences*, 13(10):439–446, 2009.

[29] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010.

[30] D. Roberson and J. Davidoff. The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition*, 28(6):977–986, 2000.

[31] B. E. Rogowitz and L. A. Treinish. Data visualization: the end of the rainbow. *IEEE spectrum*, 35(12):52–59, 1998.

[32] K. B. Schloss, F. C. Fortenbaugh, and S. E. Palmer. The configural shape illusion. *Journal of vision*, 14(8):23–23, 2014.

[33] D. S. Schwarzkopf, C. Song, and G. Rees. The surface area of human v1 predicts the subjective experience of object size. *Nature neuroscience*, 14(1):28, 2011.

[34] D. Skau and R. Kosara. Arcs, angles, or areas: Individual data encodings in pie and donut charts. In *Computer Graphics Forum*, volume 35, pages 121–130. Wiley Online Library, 2016.

[35] I. Spence. Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4):683, 1990.

[36] I. Spence and P. Krizel. Children's perception of proportion in graphs. *Child Development*, 65(4):1193–1213, 1994.

[37] S. S. Stevens. On the psychophysical law. *Psychological review*, 64(3):153, 1957.

[38] V. S. Störmer and G. A. Alvarez. Feature-based attention elicits surround suppression in feature space. *Current Biology*, 24(17):1985–1988, 2014.

[39] J. Talbot, V. Setlur, and A. Anand. Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics*, 20(12):2152–2160, 2014.

[40] 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*, 2019.

[41] F. Yang, L. T. Harrison, R. A. Rensink, S. L. Franconeri, and R. Chang. Correlation judgment and visualization features: A comparative study. *IEEE transactions on visualization and computer graphics*, 25(3):1474–1488, 2019.