Perceptual proxies for extracting averages in data visualizations

Lei Yuan – Indiana University

Steve Haroz – Sorbonne Université

Steven Franconeri – Northwestern University

ABSTRACT

Across science, education, and business, we process and communicate data visually. One bedrock finding in data visualization research is a hierarchy of precision for perceptual encodings of data, e.g., that encoding data with Cartesian positions allows more precise comparisons than encoding with sizes. But his hierarchy has only been tested for single value comparisons, under the assumption that those lessons would extrapolate to multi-value comparisons. We show that when comparing averages across multiple data points, even for *pairs* of data points, these differences vanish. Viewers instead compare values using surprisingly primitive perceptual cues, e.g., the summed area of bars in a bar graph. These results highlight a critical need to study a broader constellation of visual cues that mediate the patterns that we can see in data, across visualization types and tasks.

Keywords: Graph Comprehension, Visual Perception, Magnitude Perception, Data visualization

INTRODUCTION

Graphs and other data visualizations allow us to efficiently extract statistics, patterns, and relations in data (Cleveland, 1985; Kosslyn, 2006; Pinker, 1990; Szafir et al., 2016; Tufte, 2001), and are an essential tool for scientific understanding, discovery, and public communication (Friendly, 2008; Huff, 2010). But the power of these visualizations is constrained by the limitations and idiosyncrasies of the human visual system. Categorical color perception may lead to the introduction of unintended boundaries in otherwise continuous values (Rogowitz, Treinish, & Bryson, 1996). Visual biases towards the centers of objects can bias our memory of a bar graph's value as closer to the center of the bar, even when the value is indicated by the top (Newman & Scholl, 2012). Visual working memory capacity can limit the complexity of graphed relationships that we can extract at a time, even within a small number of data points (Halford, Baker, McCredden, & Bain, 2005)

Among these constraints on visual processing of data, one bedrock finding is a hierarchy of judgment precision among visual features that encode data values (Cleveland & McGill, 1984; Heer & Bostock, 2010). For example, as illustrated on Figure 1, encoding data as a dot plot or bar graph allows the visual system to rely on the most precise feature of *spatial position*, while encoding data as a stacked bar graph (or otherwise unaligned bars) only allows the use of *length*, which is a less precise encoding. A treemap or pie graph relies on *area* (Kosara & Skau, 2016), which is even less precise than length.

The precision of these encodings was determined for comparisons between single values, under the assumption that what is true for comparing single values is also true for multiple values. Graph comprehension and data visualization researchers often decompose graph reading into "elementary perceptual processes", such as detecting the orientation of a line, the height of a bar or the size of an angle (Larkin & Simon, 1987; Simkin & Hastie, 1987). But a decomposition approach may not scale when a task demands integration among multiple values, or extraction of general trends from the data (Carswell, 1992; Ratwani, Trafton, & Boehm-Davis, 2008).

Consider the example in Figure 2, in which we seek to compare average salaries for male and female employees at a small company. If there are equal numbers of male and female employees, as in the left graph, it is easier to see that male salary is higher on average. A decomposition approach suggests that an average could be constructed from the spatial position of the top of each bar. But this bar graph allows encoding data in at least three redundant ways: the positions of the tops, the lengths of the bars, and even the areas of the bars—a judgment might be made using any of those dimensions. You may not even extract a true average, because in the left display, you can use the sum of the salaries (the total area taken up by all bars of each category) as a proxy for that average. The right graph allows us to tease these possibilities apart, because even though the viewer still seeks to compare *average* salaries, there are more female employees. Because the number of bars is unequal, using the

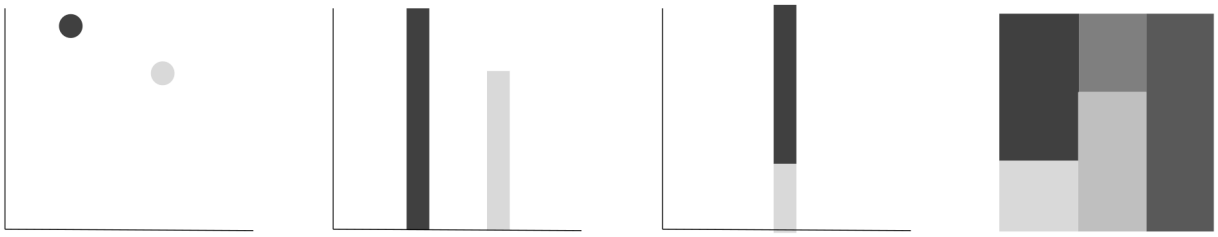

Figure 1. For a single-value comparison, the dot plot provides encoding of the data value by spatial position in the y-dimension. The bar graph provides both spatial position (the top of the bar) and a redundant encoding by the bar's length. The stacked bar graph provides only a length encoding, because the bases of the bars are misaligned. A treemap (Shneiderman, 1992) provides an area encoding, which affords the least precise comparison of data values.

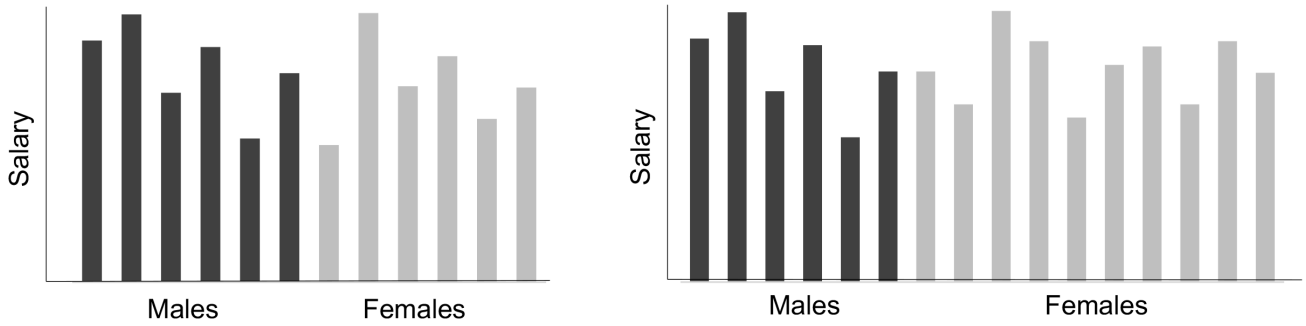

Figure 2. Sample graph requiring comparison of multiple values. Magnitude information is carried by the positions of the bar tops, their lengths, and even their summed area. While we assume that aggregate operations, such as comparing the averages of these groups, rely on precise position information, here we show that summed area is more likely the perceptual proxy, as revealed by low performance in the right graph, where relative area is no longer strongly correlated with relative average value between the two sets.

sum as a proxy for the average is no longer possible – and you may feel (and our data show) that comparing average salaries is indeed much harder.

Our hypothesis is that the advantage of position encoding from previous research (Cleveland & McGill, 1984; Simkin & Hastie, 1987) is available for single-value comparisons—but for comparisons involving larger sets, the visual system fails to use the most precise feature (position) and uses a less precise feature (length or even the *area* of the bars). Further, the visual system may not be able to extract averages from multiple bar values, it may be restricted (or at least, irresistibly tempted) to considering *sums* only.

We tested these predictions by presenting participants with graphs that have spatial position information (dot plots), both spatial position and spatial extent information (normal bar graphs), and only spatial extent information (misaligned bar graphs), and measured their discrimination threshold for choosing the set with the higher average value (Fig. 3). We found that when reading simple bar graphs that require comparing only two values, judgments were similarly precise for normal bar graphs and dot graphs (compared to misaligned bar graphs), suggesting that the visual system relied on spatial position information. However, for bar graphs involving comparison between two sets of values, performance was equally imprecise across bar graphs and misaligned bar graphs, suggesting that only the length or area of the bars was available to extract the data values. For judgments across sets of bars with unequal set sizes, performance plummeted, which is consistent with a surprising account that *summed* area (or length) is the key feature that underlies average value comparisons between sets with multiple values.

EXPERIMENT 1

Participants judged which one of two variables had a higher value (single-value comparison, i.e., 1vs1) or which one of two groups of variables had a higher mean value (multi-value comparison, i.e., 2vs2 or 6vs6). To test the critical visual units that people selectively attend to when reading bar graphs, we presented participants with three graph types— dot plots

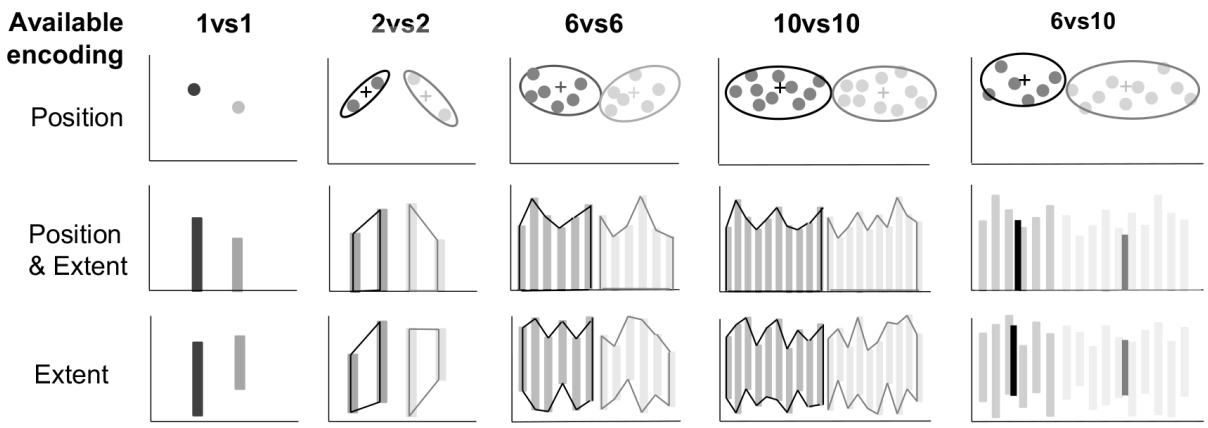

Figure 3. Stimuli used in the current study and hypothesized underlying perceptual processes. 1vs1, 2vs2 and 6vs6 were used in Experiment 1. 6vs6, 10vs10 and 6vs10 were used in Experiment 2. For single-value comparison (1vs1), spatial position encoding appears to be engaged when reading bar graphs that afford both spatial position and extent information. In contrast, for multi-value comparison (2vs2, 6vs6, 10vs10), spatial extent encoding appears to be engaged, suggesting that participants relied on the total summed area of all bars. This spatial extent strategy forbids precise judgment between datasets that have different numbers of items (6vs10). In contrast, encoding the centroid of multiple spatial positions (upper right display) may lead to more precise judgments.

(presenting only spatial position information), typical bar graphs (presenting both spatial position and spatial extent information), and misaligned bar graphs (presenting only spatial extent information, simulating the comparison of parts of a stacked bar graph). Thus, the design was a comparison type (3) X graph type (3) within-subjects design. The prediction was that single-value comparison and multi-value comparison involved the use of different visual features, resulting in an interaction between comparison type and graph type.

PARTICIPANTS

37 people participated in this experiment through Amazon Mechanical Turk. All lived in a predominantly English-speaking country and had a 90% or higher acceptance rate of past HITs. The mean duration for completing the study was under 20 minutes. Participants were compensated $5 for their participation.

STIMULI AND APPARATUS

The experiment was run via a Javascript website built on the d3.js library (Bostock, Ogievetsky, & Heer, 2011), which generated the stimuli as SVG vector graphics. All sizes are reported in pixels; for smaller screens, the stimuli were proportionally scaled down to fit on the screen.

Each graph was centered on the screen, and consisted of two axes (identical for all conditions), and two (or more) bars (in the normal bar graphs and the misaligned bar graphs condition), or two (or more) dots (in the dot graphs condition). The graph size was 960 x 570 pixels. The bars and circles were 14 pixels wide and had 28 pixels of space between each other. The two individuals or two groups of bars (or dots) were denoted by different colors (i.e., red and blue), and the locations of the colors were counterbalanced.

We used a staircase procedure to adaptively measure Just Noticeable Differences (JND), a measure of discrimination threshold. A small JND means a participant can detect small difference between two groups, reflecting higher perceptual discriminability. During the staircase procedure, if participants correctly (or incorrectly) answered the previous trial, the difficulty of the next trial increased (or decreased) by one level. The level of difficulty was defined by the difference between the means of the Y-axis values of each group. The Y-axis value was defined by the height (the distance between the X-axis to the top of the bars or dots) in the normal bar graphs and dot graphs conditions; for the misaligned bar graphs conditions, the value was defined by the length of the bars. There were 12 levels of staircase, ranging from the easiest to the hardest (in pixels): 60, 45, 34, 25, 18, 14, 11, 8, 6, 4, 2, and 1. The first trial of the experiment started with the easiest level of 60. The staircase had a 1up-1down format, which means that each correct or incorrect response would move the staircase value up or down by one level. The staircase was truncated (García-Pérez, 1998) by ceiling (60) and floor (1) levels. A separate staircase was run for each of the 9 conditions.

The mean Y-axis value for the smaller group was a random number between 200 and 300 pixels. The mean value for the larger group was calculated by adding the staircase level to the smaller group's mean. For 6vs6 comparisons, to rule out

the possibility that participants used statistical information other than the mean in their judgment, we randomized which set had the larger mean with which set had the larger standard deviation, largest single value, and smallest single value. For 2vs2, we randomized which side had the larger standard deviation. We chose two fixed standard deviations: 90 and 120 pixels. The individual Y-axis values were generated by randomly sampling numbers from a normal distribution repeatedly until the desired set properties were achieved (to be within 1%). The individual Y-axis values were used as height information to create the bars in the normal bar graphs condition, dot positions in the dot graphs condition, and length in the misaligned bar graphs condition. For the misaligned bar graphs condition, the bars were misaligned by offsetting each bar upwards by a random value between 0 and 150 pixels.

PROCEDURE

Participants were introduced to the experiment as a graph-reading task. They were told that the Y-axis represented the time it took for a red and a blue team to finish a car racing game. Sometimes there was only one car within each team; sometimes there were multiple cars. Participants were told to judge which team on average took longer. For the misaligned bar graph condition, they were told explicitly that the length of the bars represented the time values.

Each trial began with a fixation at the center of the display for 500ms. A graph then appeared at the center of the screen until participants made a judgment about which side (left or right) had a higher value or a higher mean value by pressing either the left or right key on the keyboard. Participants were given 2 seconds to respond; if no response was provided within the time limit, the current trial was marked as incorrect and participants were reminded to respond quickly next time.

There were 25 trials in the 1vs1 conditions, and 49 trials in the 2vs2 and 6vs6 conditions.

The design of the experiment—3 set size comparisons and 3 types of graph—results in 9 unique conditions, which were randomly presented in each block of 9 trials. Each block was followed by a break that the participant could end by pressing a key. A new block of trials was then be shown in a random order, and these blocks continued until the end of the experiment. Once the 1vs1 conditions were complete, only 2vs2 and 6vs6 trials were shown. The experiment included a total of ([25 trials for 1vs1] + [49 trials for 2vs2] + [49 trials for 6vs6]) × [3 graph types] = 369 trials. There was an additional set of 2 practice trials at the beginning of the experiment for all set sizes and graph types with the highest staircase value (i.e. 60 pixels difference between two groups). Participants were given feedback at practice to make sure that they understand the instruction.

RESULTS

JND (Just Noticeable Difference) was calculated based on the last half of the trials[1]. A comparison type (1vs1, 2vs2, 6vs6) × graph type (bar vs. dot vs. misaligned bar) repeated-measure ANOVA was performed. Most relevant to our hypothesis, there was a significant interaction between comparison type and graph type, F (4,144) = 5.49, p = .0004, η_p^2 = .13 (Fig. 4). Because we were particularly interested in whether viewers selectively attend to the spatial position or spatial extent information when reading normal bar graphs, we focused our analysis on two planned comparisons: normal bar graphs vs. misaligned bar graph, and normal bar graphs vs. dot graphs. For 1vs1 comparison, performance on the misaligned bar graphs (M = 6.6, SD = 4.28) was significantly worse than the normal bar graphs (M = 1.65, SD = .74), t (36) = 7.4, p < .001, d = 1.22, but there was no significant difference between normal bar graphs and the dot graphs (M = 1.75, SD = 1.29), t (36) = .57, p =.6, d = .09. These results indicate that participants had relied on the spatial position information when comparing individual values on normal bar graphs. In contrast, for 2vs2 comparison, performance on the dot graphs (M = 9.35, SD = 8.35) was marginally significantly worse than the normal bar graphs (M = 7.06, SD = 5.12), t (36) = 1.95, p = .059, d = .32, but there was no significant performance difference between the normal bar graphs and the misaligned bar graphs (M = 7.55, SD = 5.74), t (36) = .5, p = .6, d = .08. For 6vs6 comparisons, there was no significant difference among conditions, ps > .3.

The ANOVA analysis also revealed a main effect of comparison type, F (2, 72) = 43.1, p < .0001, η_p^2 = .55. Comparison between single values (M = 3.35, SD = 1.7) was significantly better than that comparison between pairs of values (M = 7.99, SD = 5.22), t (36) = 6.8, p < .0001, d = 1.11, which was significantly better than that comparison between two groups of values (M = 11.11, SD = 6.72), t (36) = 3.7, p = .0008, d = .6. We did not find a main effect of graph type, F (2, 72) = 1.45, p = .24, η_p^2 = .038.

[1] *We ran a Monte Carlo simulation of 20,000 experiments with 37 subjects making random responses, and we computed JNDs from these simulations. 95% of these simulations found JNDs between 21 and 31. Because all of our JND values were well below this range, we can safely assume that all performance levels were better than chance.*

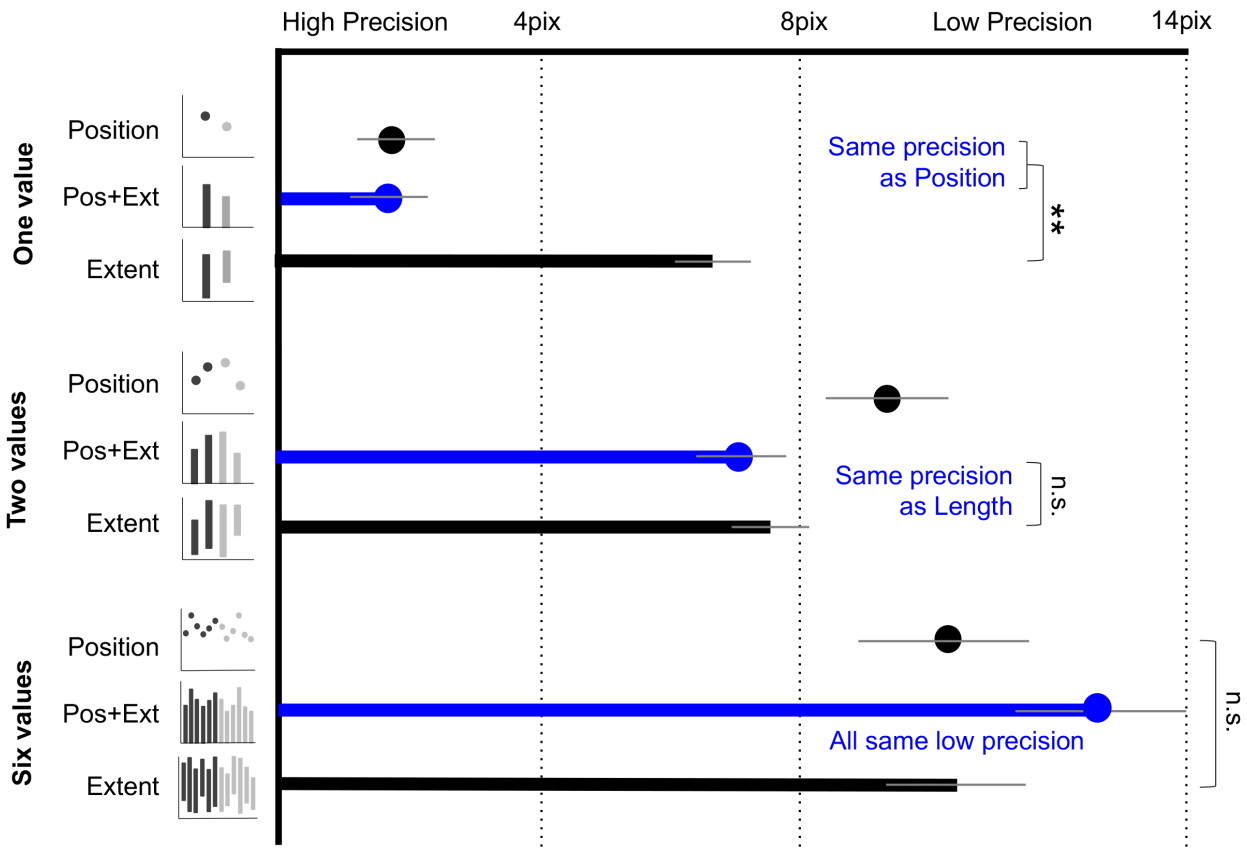

Figure 4. Experiment 1 results: JND by comparison types (1vs1, 2vs2 or 6vs6) and visual features (Extent, Position + Extent, Position). Error bars depict within-subject standard error. *** p <.001 ** p <.01 * p < .05

DISCUSSION

Comparison between single values on bar graphs appears to rely on different visual features than comparison of averages of sets. When comparing two individual values, normal bar graphs showed similar performance as dot graphs (spatial position), and both were better than the misaligned bar graphs (spatial extent). Consistent with previous findings (Cleveland & McGill, 1984; Simkin & Hastie, 1987), this result suggests that participants relied on the spatial position when judging two individual values on bar graphs. However, the pattern was different for multi-value comparison on bar graphs. For comparisons between two pairs of values (2vs2), performance on the normal bar graphs was no better than the misaligned bar graphs (spatial extent), and there was a trend showing that performance on the dot graphs (spatial position) was different from that of the normal bar graph. These results raise the possibility that participants had relied on the spatial extent of the bars rather than their spatial position, on both the normal bar graph and misaligned bar graph conditions.

For the multi-value comparisons, performance was surprisingly similar across conditions. One might assume that participants make these judgments by leveraging information about *average* spatial positions or spatial extents (length or area). But while it might seem counterintuitive, it is possible that participants treat entire set of bars as a single unit, summing the area of its shape envelope instead of the averaging the length of its bars. If so, this strategy should suffer for displays with an unequal number of bars between the groups. We tested this prediction in the second experiment.

EXPERIMENT 2

We again presented the same data in three different formats—normal bar graphs, dot graphs and misaligned bar graphs. There were two comparison types—equal numbers of items in two groups (6vs6 and 10vs10) or unequal numbers of items in two groups (6vs10). Since the large set size comparisons were more challenging, we also added small set size conditions (1vs1 and 2vs2) and used them as base-line measures to exclude outliers (defined as 3 standard deviations from the mean).

This experiment's preregistration, data, and experiment materials are available at https://osf.io/vhgdn/.

Reported analyses that were not preregistered are marked as exploratory.

PARTICIPANTS

41 people participated in this experiment through Amazon Mechanical Turk. One participant was dropped based on a priori exclusion criteria (JNDs in the 1vs1 or 2vs2 conditions were 3 standard deviations from the mean).

STIMULI AND PROCEDURES

The staircases began at 60 and had a 3 down / 1 up design: 3 correct answers in a row multiplied the staircase value by 0.75, and 1 incorrect answer multiplied the staircase value by 1.11. These constants yielded a staircase that would converge on a JND equivalent to 68% accuracy (García-Pérez, 1998). The stimuli could be generated for staircase values from 0 to 75. If the staircase went above 75, a "virtual" staircase could still reach higher values, but the stimuli values were capped at 75. A Monte Carlo simulation of this staircase design showed that random responses were very unlikely (less than a 2.5% chance) to yield JNDs below 66, which allows us to measure higher JNDs than previous experiment could.

The experiment comprised 3 blocks. In each block, 16 trials of one condition were presented consecutively. Following a break, this process was repeated until all conditions were seen. The second and third blocks were the same as the first, but the order of the conditions was randomized each time. The staircase of each condition in the second and third block was a continuation of the previous bock's staircase. The experiment included a total of (5 set sizes) × (3 graph types) × (3 blocks) × (16 trials) = 720 trials per subject. Feedback was given for the first 16 trials for all conditions (the easiest part of the staircase) to ensure that subjects understood the task.

RESULTS

An ANOVA of set size (6vs6, 10vs10) X graph type (bar, dot, misaligned bar) did not detect a significant main effect of set size, F (1,39) = 2.68, p = .11, η_p^2 = .06, graph type, F (2,78) = .17, p = .85, η_p^2 = .005, nor interaction between them F (2,78) = 1.09, p = .34, η_p^2 = .027. Thus, in the analysis below, we combined these 6vs6 and 10vs10 conditions into the *same set size* comparison condition; 6vs10 was treated as the *different set size* comparison condition. Most critical for the present argument, the ANOVA revealed a robust main effect of comparison type (same vs. different set size), F (1,39) = 66.1, p < .001, η_p^2 =.63. As expected, participants performed better when there was equal number of items for both groups (M = 26.48, SD = 16.25) than when there was unequal numbers of items (M = 37.97, SD = 17.24). The ANOVA did not detect a significant main effect of graph type[2], F (2,78) = 1.66, p = .2, η_p^2 = .04.

Interestingly, an exploratory analysis showed that the dot (position-only) condition appeared to be less impaired by the different set-size comparison condition. A comparison type (same vs. different set size) X graph type (3) repeated-measure ANOVA revealed a significant interaction between comparison type and graph type, F (2,78) = 3.15, p = 0.048, η_p^2 = .076 (Fig. 5). When there were same numbers of items for both groups (same set size) [2],there was no significant performance difference between dot graphs (M = 27.06, SD = 16.86) and normal bar graphs (M = 26.15, SD = 13.76), t (39) = .44, p = .7, d = .07, or between misaligned bar graphs (M = 26.22, SD = 11.63) and normal bar graphs, t (39) = .06, p = 1, d = .009. However, when there was unequal numbers of items between groups (different set sizes, a planned comparison between dot graphs and the normal bar graphs indicates that performance was significantly better on the dot graphs (M = 34.65, SD = 17.12) than the normal bar graphs (M = 41.08, SD = 16.83), t (39) = 2.6, p = .01, d = .41. In contrast, there was no significant performance difference between the normal bar graphs and the misaligned bar graphs (M = 38.17, SD = 17.18), t (39) = 1.2, p = .2, d = .19. It is possible that, for the dots in the 6 vs. 10 condition, a center a center-of-mass proxy might be used for the dots, as opposed to a summed area proxy for the bars.

DISCUSSION

When comparing two groups of data points that have different set sizes, performance was equally reduced for both normal bar graphs and misaligned bar graphs (compared to dot graphs), suggesting that the summed area served as a proxy for comparisons of 'average' across a set of values.

There was also a trend suggesting that the dot (position-only) condition appeared to be less impaired by the different set-size comparison condition. If participants were also able to rely on a center-of-mass proxy for the dot graphs, that strategy might be less impaired by the different set sizes of the two groups. But we might expect performance to remain impaired by the irrelevant signal of the size of the dot group, for the same reason that area would interfere with decisions about bars or misaligned bars. Groups of 10 are larger in area (and also more numerous) than groups of 6, which may serve

[2] *While these tests were not included in the preregistration and do not affect our hypotheses, we report them here for the interested reader.*

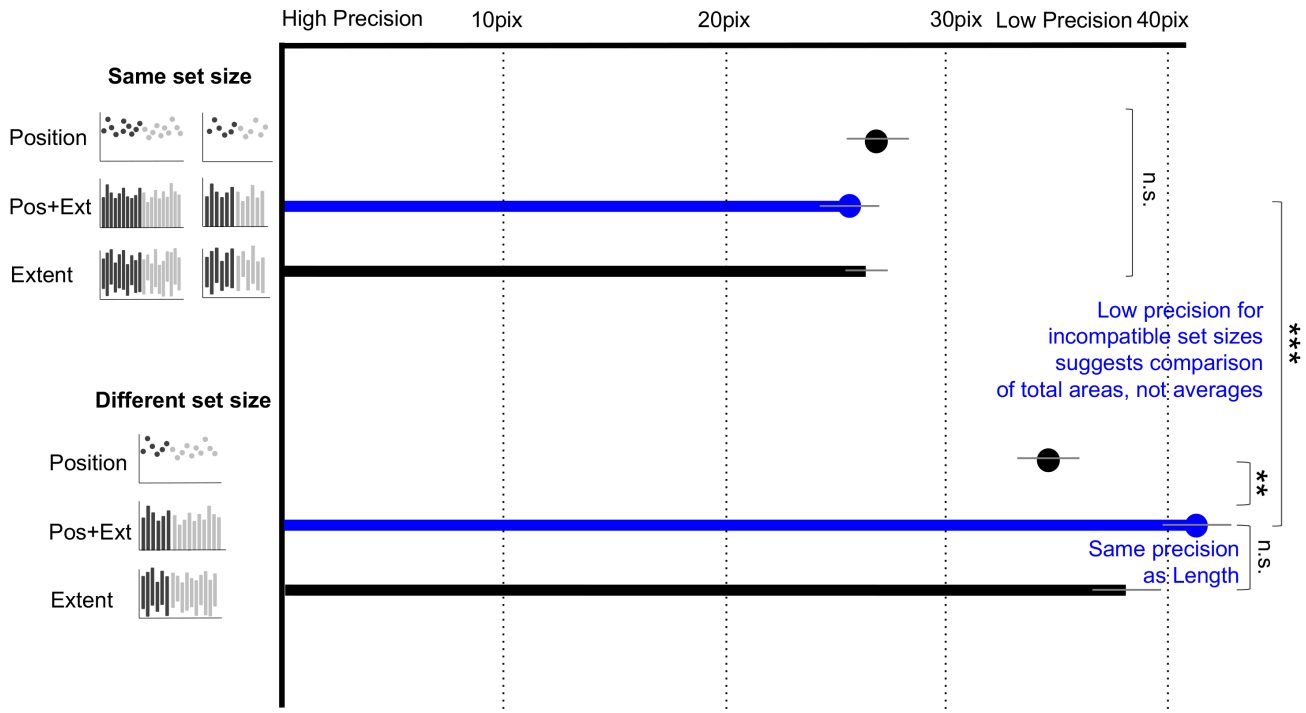

Figure 5. Experiment 2 results: JND by comparison types—same set size (6vs6, 10vs10) or different set size (6vs10)—and visual features (Extent, Position + Extent, Position). Error bars depict within-subject standard error. *** p <.001 ** p <.01 * p < .05

as an incongruent signal in some trials and a congruent signal in others, adding bias to the decision process (we thank an anonymous reviewer for suggesting this analysis). Indeed, an exploratory analysis showed that participants were more accurate in their response when the group with the higher average value was also the group with more dots (M = 76%, SD = 14%), compared to when the group with the higher average value was the group with less bars (M = 56%, SD = 16%), t (39) = 5, p = .00001. This same effect was also present in the bar and misaligned bar conditions (ts > 4.6, ps < .0001). These results suggest that irrelevant information about number or area biases decisions about average length or position.

If the different set size bar condition was also affected by interference from its area, perhaps that interference would be easier to avoid if the area of the bars were less salient (we again thank an anonymous reviewer for this suggestion). We conducted two additional experiments (see supplemental materials) where we reduced the salience of the areas in some or all conditions. In the first, we changed all shapes (dots, bars, misaligned bars) to single-pixel outlines, and in the second, we depicted dots as filled, while changing the other two to outlines, which approximately equates the amount of 'ink' of the area across conditions. In both of these experiments, misaligned bar still showed the worst performance, but bars improved to the same level of dots. It is possible that decreasing the salience of the area of the bars helped participants relatively inhibit their area, and focus on their tops. While further work is needed on the topic before issuing prescriptions (for instance, a within-participant comparison of filled vs. unfilled bars), we find it intriguing that such design changes to bar graph might help people perceive visualized data, and the statistics among them, more accurately.

GENERAL DISCUSSION

The precision of data extraction is a core perceptual constraint in understanding graphs and other data visualizations. The hierarchy of encoding precision of single-value comparisons (Cleveland & McGill, 1984) is typically assumed to extrapolate to more complex displays. The present results suggest that it does not, and that multi-value comparisons can rely on surprisingly different perceptual proxies for the desired statistics. When comparing two individual values on bar graphs, viewers seemed to rely largely on the spatial position information of the bars. But when comparing the mean values of large groups of data values on both typical and misaligned bar graphs, viewers appeared rely on a counterintuitive feature, the summed area of the bars. We find this result striking because the height information in the bar graphs—the tops of the bars—is available, and can be used more effectively to construct more precise averages in the dot plot condition, possibly via a center-of-mass judgment.

Why did participants appear to rely on *sums* instead of averages? That result is particularly surprising given extensive evidence that the visual system can compute visual statistics across basic features (Alvarez, 2011; Ariely, 2001; Chong & Treisman, 2003; Haberman & Whitney, 2012), including the *average* of spatial positions (Alvarez & Oliva, 2008), or the sizes of circles (Chong & Treisman, 2003), even across groups of different set sizes (Chong & Treisman, 2005).

Such efficient extraction of statistics from spatial positions, lengths, and areas is thought to support pattern extraction from data visualizations (Gleicher, Correll, Nothelfer, & Franconeri, 2013; Szafir et al., 2016). If people can extract averages across positions or lengths, why not in the present displays? This incongruent result may provide a case of use-inspired basic research (Shneiderman, 2015), leading visual cognition researchers to find the limiting factors to such statistical extraction. In the present displays, the density, proximity, bottom- or center-alignment, or filled vs. unfilled texture of the objects may prevent extraction of average value, and it will be important for future research to outline the power and limits of these abilities in real-world relevant displays.

Why did participants rely on *extents*—the areas or lengths of the objects—instead of the spatial positions of their tops? The visual system is biased to deal with whole objects, rather than the features or parts of the objects (Duncan & John, 1984; Egly, Rafal, Driver, & Starrveveld, 1994). When asked to track multiple moving objects, observers have extreme difficulty tracking just one end of a moving bar, reflexively snapping their attention to the whole bar instead (Scholl, Pylyshyn, & Feldman, 2001), and a similar reflex may lead viewers to bias recall of values plotted in bar graphs toward the center of the bar, instead of the top edge (Newman & Scholl, 2012). In the current study, although spatial position might be the most relevant feature on a bar graph for comparing the averages between groups of values, viewers seem to be incapable of ignoring the spatial extent (length or area) information, and select the whole bars as the unit of attention (Scholl, 2001). The supplemental experiments reported in Experiment 2 gives an initial hint that design changes might mitigate this impairment.

The established hierarchy of perceptual precision for single-value comparison does not extrapolate to judgments across larger sets of data. These findings have direct implications for graph designers and educators who aim to produce accurate and unbiased visual interpretation of graphs. For example, they point to advantages for the dot plot when a graph involves group comparisons, and for explicitly representing the mean value when reading a graph that requires making judgment about the mean value. More broadly, these findings demand new work on the constrains of our visual system's ability to extract statistical information from data displays and other artificial worlds, as well as new theories of information encoding from data displays of more realistic complexity.

REFERENCES

Alvarez, G. a. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003

Alvarez, G., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392–8. https://doi.org/10.1111/j.1467-9280.2008.02098.x

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science*, *12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309. https://doi.org/10.1109/TVCG.2011.185

Carswell, C. M. (1992). Choosing specifiers: an evaluation of the basic tasks model of graphical perception. *Human Factors*, *34*(5), 535–554.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404. https://doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: computing the average size in perceptual groups. *Vision Research*, *45*(7), 891–900. https://doi.org/10.1016/j.visres.2004.10.004

Cleveland, W. S. (1985). *The elements of graphing data*. Wadsworth Advanced Books and Software Monterey, CA.

Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, *79*(387), 531–554. https://doi.org/10.1080/01621459.1984.10478080

Duncan, J., & John. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*(4), 501–517. https://doi.org/10.1037/0096-3445.113.4.501

Egly, R., Rafal, R., Driver, J., & Starrveveld, Y. (1994). Covert

Orienting in the Split Brain Reveals Hemispheric Specialization for Object-Based Attention. *Psychological Science*, *5*(6), 380–383. https://doi.org/10.1111/j.1467-9280.1994.tb00289.x

Friendly, M. (2008). A Brief History of Data Visualization. *Handbook of Data Visualization*, 15–56 (1–43). https://doi.org/10.1007/978-3-540-33037-0_2

García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research*, *38*(12), 1861–1881. https://doi.org/10.1016/S0042-6989(97)00340-4

Gleicher, M., Correll, M., Nothelfer, C., & Franconeri, S. (2013). Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2316–25. https://doi.org/10.1109/TVCG.2013.183

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. M. Wolfe & L. Robertson (Eds.), *From Perception to Consciousness: Searching with Anne Treisman*. (pp. 393–404). Oxford University Press. NY: New York.

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science*, *16*(1), 70–76. https://doi.org/10.1111/j.0956-7976.2005.00782.x

Heer, J., & Bostock, M. (2010). Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. *Proceedings of the 28th Annual Chi Conference on Human Factors in Computing Systems*, 203–212. https://doi.org/10.1145/1753326.1753357

Huff, D. (2010). *How to lie with statistics*. WW Norton & Company. NY: New York.

Kosara, R., & Skau, D. (2016). Judgment Error in Pie Chart Variations. *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization: Short Papers*, 91–95. https://doi.org/10.2312/EUROVISSHORT.20161167

Kosslyn, S. M. (2006). *Graph Design for the Eye and Mind*. Oxford University Press. NY: New York.

Larkin, J. H., & Simon, H. A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, *11*(1), 65–100. https://doi.org/10.1111/j.1551-6708.1987.tb00863.x

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, *19*, 601–607. https://doi.org/10.3758/s13423-012-0247-5

Pinker, S. (1990). A theory of graph comprehension. *Artificial Intelligence and the Future of Testing*, 73–126. https://doi.org/10.1145/2046684.2046699

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology. Applied*, *14*(1), 36–49. https://doi.org/10.1037/1076-898X.14.1.36

Rogowitz, B. E., Treinish, L. A., & Bryson, S. (1996). How Not to Lie with Visualization. *Computers in Physics*, *10*(3), 268. https://doi.org/10.1063/1.4822401

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*(1–2), 1–46. https://doi.org/10.1016/S0010-0277(00)00152-9

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition*, *80*(1–2), 159–177. https://doi.org/10.1016/S0010-0277(00)00157-8

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, *11*(1), 92–99. https://doi.org/10.1145/102377.115768

Shneiderman, B. (2015). *The New ABCs of Research: Achieving Breakthrough Collaborations*. *Oxford University Press* (Vol. 1). Oxford University Press. NY: New York. https://doi.org/10.1093/acprof:oso/9780198758839.001.0001

Simkin, D., & Hastie, R. (1987). An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association*, *82*(398), 454–465. https://doi.org/10.1080/01621459.1987.10478448

Szafir, D. A., Haroz, S., Gleicher, M., Franconeri, S., M., C. L., S., Y., & M., H. (2016). Four types of ensemble coding in data visualizations. *Journal of Vision*, *16*(5), 11. https://doi.org/10.1167/16.5.11

Tufte, E. R. (2001). *The visual display of quantitative information*. Graphics Press. CT: Cheshire.

SUPPLEMENTAL MATERIAL

Why do people rely on the summed area when comparing averages across two groups on a bar graph instead of relying on the spatial position information? According to an object-based-attention hypothesis, the visual system is biased to deal with entire objects rather than the features or parts of the objects, so that viewers would have difficulty ignoring the spatial extent (length or area) information, and select the whole bars as the unit of attention. In contrast, by the total-pixels-of-ink hypothesis, this result stems from the fact that the color of the bars in a bar graph occupies considerable pixels of ink, which may distract viewers from attending to the spatial position information.

To test these hypotheses, we ran two additional experiments. Both experiments' preregistration, data, and source code are available at https://osf.io/vhgdn/ In the first experiment, we used unfilled bars and dots (see Fig. 1) while all the other aspects of the experiment were identical to Experiment 2. This design eliminated the pixel disparity between the bars and dots but preserved the object-hood of the bars. According to the object-based-attention hypothesis, we should still find an advantage of the dot graph than the bar graph when there was an unequal number of items between two groups (the 10vs6 condition). However, according to the total-pixel-of-ink hypothesis, we would not find such an advantage for the dot graphs, because the total amount of ink was comparable between the bar graphs and the dot graphs.

EXPERIMENT S1: OUTLINES

44 people participated in this experiment through Amazon Mechanical Turk. 4 participants were dropped based on the same a priori exclusion criteria as Experiment 2(JNDs in the 1vs1 or 2vs2 conditions were 3 standard deviations from the mean).

An ANOVA of set size (6vs6, 10vs10) X graph type (bar, dot, misaligned bar) did not detect a significant main effect of set size, graph type, nor interaction between them; thus, we combined these 6vs6 and 10vs10 conditions into the same set size comparison condition and 6vs10 was treated as the different set size comparison condition. An ANOVA of set size (same vs. different) X graph type (bar, dot, misaligned bar) revealed a robust main effect of comparison type (same vs. different set size), $F(1,39) = 57.8$, $p < .000001$. As expected, participants performed better when there was an equal number of items for both groups ($M = 23$, $SD = 12$) than when there were unequal numbers of items ($M = 38$, $SD = 18$). The ANOVA did not detect a significant interaction between set size and graph type, $F(2,78) = 2.21$, $p = .12$. When there was unequal numbers of items between groups (different set sizes, a planned comparison between dot graphs and the normal bar graphs did not detect a significant difference between the dot graphs ($M = 34.67$, $SD = 16.71$) than the normal bar graphs ($M = 37.27$, $SD = 14.89$), $t(39) = .22$, $p = .8$. There was also no significant performance

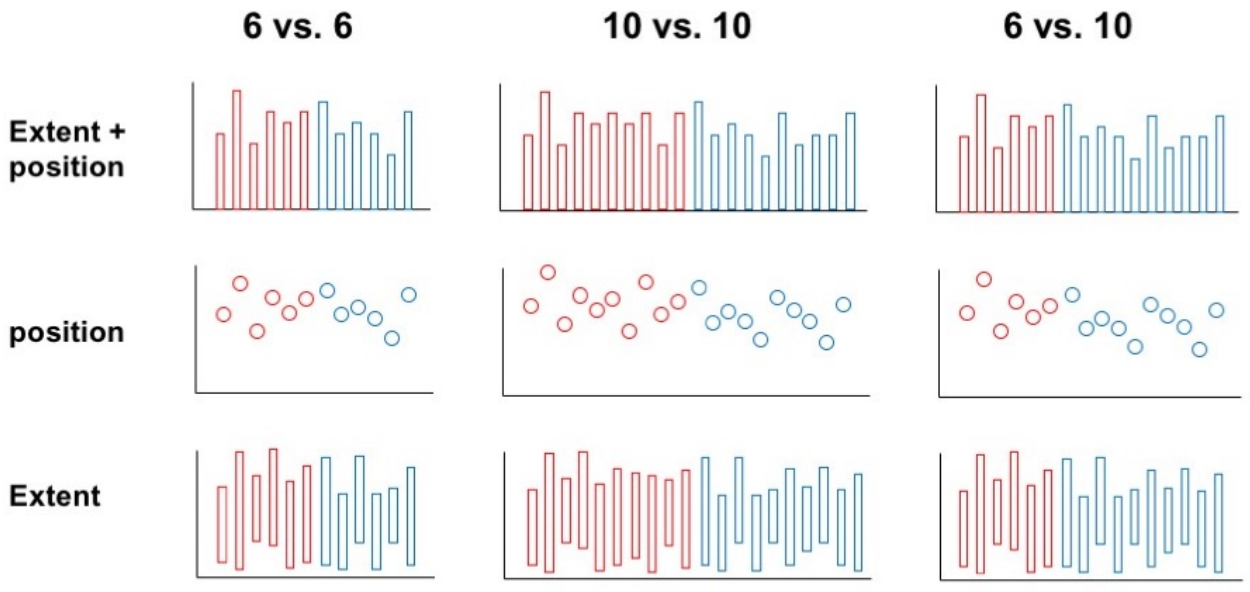

Figure S1. Illustration for the stimuli used in the first experiment

difference between the normal bar graphs and the misaligned bar graphs (M = 40.53, SD = 18.25), t (39) = 1.1, p = .3.

EXPERIMENT S2: EQUATED PIXELS

In the second additional experiment, we equated the number of pixels in the bar graphs and dot graphs (see Fig. 2) such that the width of the bars were 2 pixels and the radius of the dots was 12.6 pixels (because the average height of the bars cannot be known a priori in the staircase procedure, an estimate of 250 pixels was used based on the results from Experiment 2). All other aspects of the experiment were identical to Experiment 2. This design equaled the number of pixels between the bars and dots while preserved the objecthood of the bars. If the object-based-attention hypothesis is correct, then we would still find an advantage of the dot graph than the bar graph when there was an unequal number of items between two groups (the 10vs6 condition). In contrast, if the total-pixel-of-ink hypothesis is correct, then we would not find such an advantage for the dot graphs.

49 people participated in this experiment through Amazon Mechanical Turk. 9 participants were dropped based on the same a priori exclusion criteria as Experiment 2(JNDs in the 1vs1 or 2vs2 conditions were 3 standard deviations from the mean).

An ANOVA of set size (6vs6, 10vs10) X graph type (bar, dot, misaligned bar) did not detect a significant main effect of set size, graph type, nor interaction between them; thus, we combined these 6vs6 and 10vs10 conditions into the same set size comparison condition and 6vs10 was treated as the different set size comparison condition. An ANOVA of set size (same vs. different) X graph type (bar, dot, misaligned bar) revealed a robust main effect of comparison type (same vs. different set size), F (1,39) = 59.8, p < .000001. As expected, participants performed better when there was an equal number of items for both groups (M = 24.75, SD = 13.39) than when there were unequal numbers of items (M = 37.19, SD = 19.43). The ANOVA did not detect a significant interaction between set size and graph type, F (2,78) = 0.7, p = .5. When there was unequal numbers of items between groups (different set sizes, a planned comparison between dot graphs and the normal bar graphs did not detect a significant difference between the dot graphs (M = 34.87, SD = 15.68) than the normal bar graphs (M = 36.76, SD = 19.16), t (39) = .6, p = .6. There was also no significant performance difference between the normal bar graphs and the misaligned bar graphs (M = 39.93, SD = 19.55), t (39) = .97, p = .3.

In both of these additional experiments, the result failed to find an advantage of dot graphs in the 10vs6 condition, consistent with a total-pixel-of-ink explanation. Although more research is needed, this result suggests that reducing the pixels of ink on a bar graph may indeed help viewers to select the spatial position information of the bars.

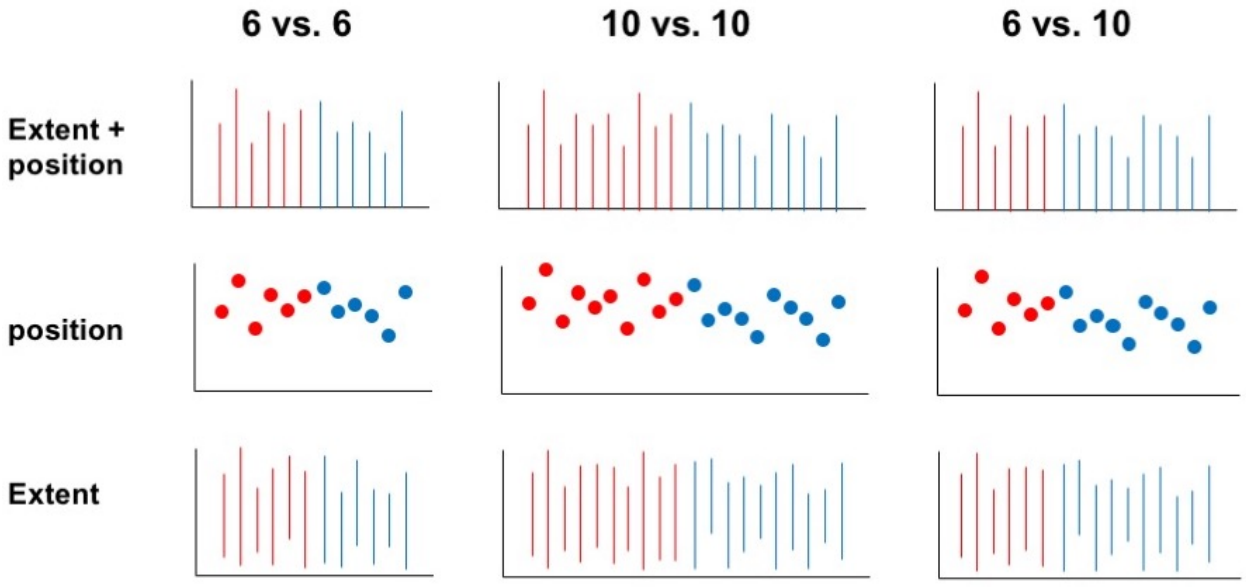

Figure S2. Illustration for the stimuli used in the second experiment