The Weighted Average Illusion: Biases in Perceived Mean Position in Scatterplots

Matt-Ian Hong, Jessica K. Witt, and Danielle Albers Szafir *Member, IEEE*

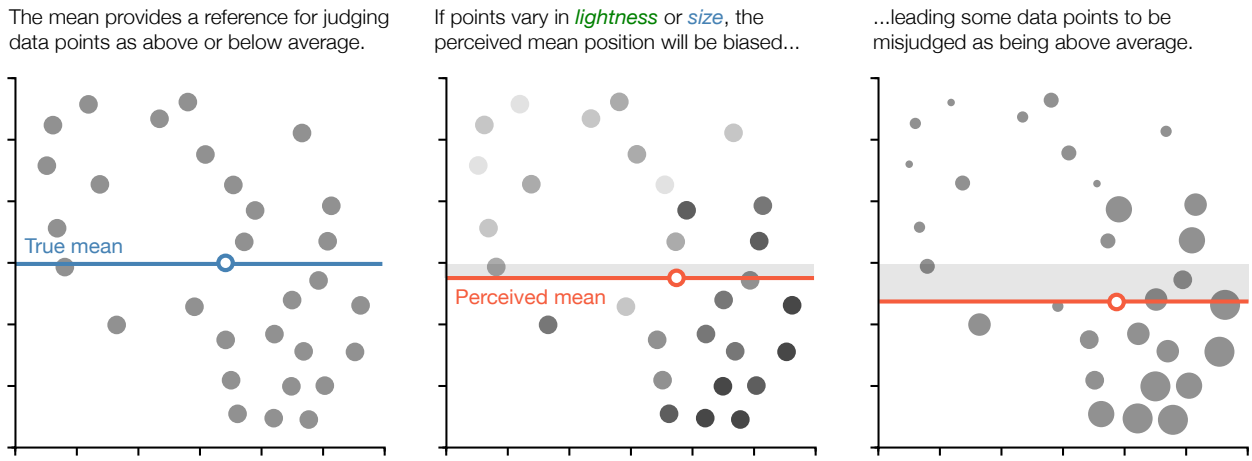

Fig. 1. We measured how integrating size or lightness into scatterplots can systematically bias the perceived mean of those points. We model how individual marks contribute to this bias, allowing us to predict differences in the actual mean (blue) and where people see the mean (orange). This misinterpretation can inhibit sensemaking and decision making. For instance, points in the grey boxes above will be incorrectly perceived as having higher-than average y-values. Under this illusion, graph readers' mean estimates can be increasingly biased as a function of size or lightness ranges mapped onto the data.

Abstract—Scatterplots can encode a third dimension by using additional channels like size or color (e.g. bubble charts). We explore a potential misinterpretation of trivariate scatterplots, which we call the *weighted average illusion*, where locations of larger and darker points are given more weight toward x- and y-mean estimates. This systematic bias is sensitive to a designer's choice of size or lightness ranges mapped onto the data. In this paper, we quantify this bias against varying size/lightness ranges and data correlations. We discuss possible explanations for its cause by measuring attention given to individual data points using a vision science technique called the centroid method. Our work illustrates how ensemble processing mechanisms and mental shortcuts can significantly distort visual summaries of data, and can lead to misjudgments like the demonstrated weighted average illusion.

Index Terms—Human-Subjects Quantitative Studies, Perception & Cognition, Scatterplots, Feature-Based Attention, Bias.

1 Introduction

Effective visualizations like scatterplots communicate data by leveraging fast and accurate visual processes. Scatterplots map two data dimensions to position [8, 77], a precise channel for comparing values [2, 14, 51]. They also leverage our ability to summarize sets of point marks through *ensemble processing* mechanisms [6, 35, 87, 93]. Ensemble processing helps readers easily intuit how data points are distributed, allowing judgments about summary statistics such as correlations [39, 72], position means [34, 92], and clusters [1].

While ensemble perception is mostly beneficial for data comprehension, it has limitations that can also interfere with visual communication. Recent work in vision science suggests that visual channels corresponding to common design elements, like size or color, may systematically bias our abilities to estimate the mean position of a small collection of point marks [26, 74, 83]. These limitations could manifest in common visualization techniques like line charts—which lead to underestimation of means [96] and overestimation of trends [90])—and scatterplots.

Our work focuses on a systematic bias in estimating the mean position of data points in a scatterplot. This task is commonly used to assess differences across groups of data points [34, 77]. For example, Rosling used multiclass scatterplots for Gapminder [75] to compare mean GDP per capita and mean mortality rates between different geographic regions. Using small multiples scatterplots like Parlapiano [65], people can assess trends in per capita income and life expectancy by comparing means and variances across time [53]. Mean position can also provide a reference point: a survey [43] published by The New York Times used a scatterplot to ask readers to decide which venues should reopen first during the COVID-19 pandemic. Readers' decision criteria [81] are likely to be determined based on the perceived x- and y-means, as illustrated in Hessney et al. [44] and in Figure 1.

We conducted a crowdsourced study with 130 participants (§4.1) to measure biases in mean position estimates in scatterplots where point marks varied in their size or lightness. Our study illustrates the difficulty of computing the true position means of scatterplots. People's enduring tendency is to compute biased position means, where some data points weigh more toward the average depending on their size or lightness values. However, this is not a valid task in scatterplots, since weighted position means shift with the range of size or lightness

- *Matt-Ian Hong and Danielle Albers Szafir are with the ATLAS Institute, University of Colorado Boulder. Email: matt.hong@colorado.edu, danielle.szafir@colorado.edu.*
- *Jessica K. Witt is with Colorado State University. E-mail: jessica.witt@coloradostate.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

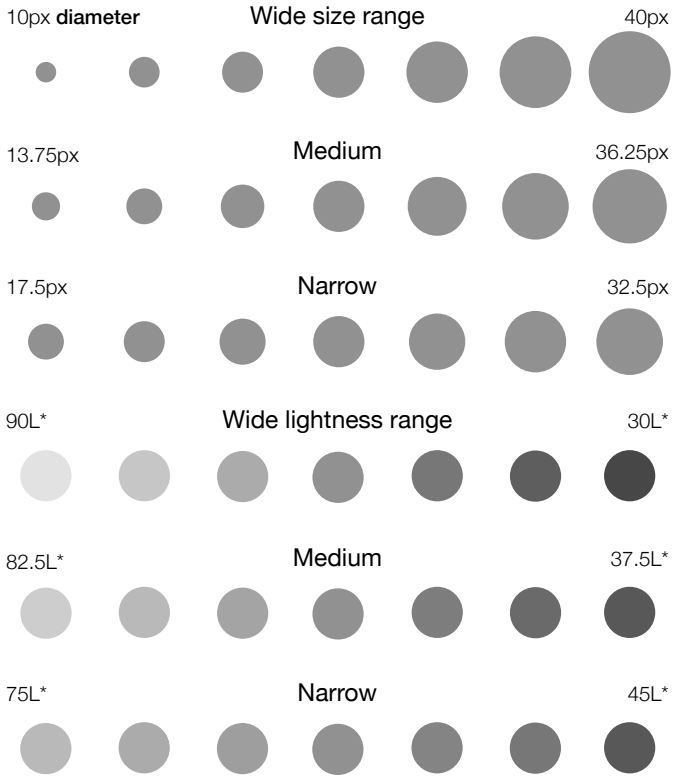

Fig. 2. We mapped the third dimension in trivariate scatterplots to one of three size or lightness ranges (shown here to scale). Steps in pixel diameter and L* values are evenly spaced, with the midpoints of all ranges being identical. Every scatterplot stimulus contained seven different mark sizes or lightness.

mapped onto the data (see Figures 1 and 2), as opposed to the data itself. We label biased judgments occurring when weighted means are used to estimate position means of scatterplot marks as the *weighted average illusion*.

Our demonstration of this misjudgment and the ensuing bias adds to the growing literature on visualization biases [24, 90, 96]. To guide our discussion of causal factors in §7.2, we incorporate a statistical model of feature-based attention (FBA) [12] known as the *centroid method* [83] which quantifies attention given to individual data marks (§6.2). In addition, our models lead to hypotheses about other strategies people might utilize, such as selective attention [63] (§7.2.2) and spatial segmentation [31], that may help practitioners understand how similar biases may arise in other designs (§7.2.3).

Contributions: Our primary contribution is quantifying the bias that may arise under the weighted average illusion in trivariate scatterplots, as a function of design choices and data patterns (Figure 6), even after training. We also demonstrate how modeling techniques in visual cognition like the centroid method can enable causal interpretations of visualization studies. These models allow practitioners to predict when biases may arise and avoid potential misjudgments. Our findings create opportunities at the intersection of visualization and vision science by hypothesizing possible mental shortcuts used in visualization interpretation and decision making.

2 Background

Honest communication in visualization means avoiding deceptions and biased interpretations of data. Understanding bias has a long tradition in visualization research [18, 59, 64, 86, 94], and one recent study demonstrates possible priming effects on position means [96]. We suspect that position mean biases could also be caused by limits on our visual system. We build on knowledge and techniques from vision science to model bias as a function of visual elements in scatterplots.

2.1 Visualization Biases

Visualizations provide a powerful yet imperfect means for communicating data. Subtle design choices may bias conclusions drawn from data [18, 59]. For example, encoding data using a bubble's area rather than its diameter can inflate perceived data differences [64]. Other biases may emerge from choices made in the data processing pipeline. For example, changing the rendering order of scatterplot points can distort our perception of distributions [59]. And cognitive biases often do not stem from the visualizations themselves, but from imperfections in an analyst's sensemaking process (see Dimara et al. [24] for a survey).

Design guidelines can provide proactive strategies for avoiding common biases [86]. For example, designers are encouraged to avoid rainbow colormaps, partly because they can cause "banding" biases that lead people to group marks that share similar hues [9, 70]. Intelligent design systems, such as visualization linters [46, 58], can leverage these guidelines to identify potentially misleading practices.

However, we lack formal models of biases that can help us reason about potential design trade-offs. While controlled experiments [64] substantiate the harmfulness of distorting aspect ratios and truncating y-axes, this latter (much maligned) practice can be helpful for interpretation depending on the data patterns and communication goals [17, 94]. If magnitudes of bias could be modeled as a function of visual elements, designers can make better judgments about these trade-offs to enhance the overall effectiveness of a visualization while limiting bias [73].

While bias can emerge as an artifact of explicit design choices, other biases may be more subtle. Xiong et al. [96] showed that position means can be biased in conventional plots: people systematically overestimated the mean using bar charts, and underestimated it using line charts. Statistical patterns in scatterplots can be distorted by geometric scaling [92]. Visualizations optimized for one task may fail for many others [67], and such failures may lead to bias: even the most honest charts might mislead the reader.

2.2 Scatterplot Perception

Scatterplots are amongst the most commonly used [76] and scrutinized [77] visualization techniques. They primarily map data to both x- and y-positions, yet encompass a diverse range of design variations [77] and are applicable to a broad array of tasks, from comparing individual values [2, 14, 16] to summary statistical tasks such as clustering [78], averaging [34], and correlation [39, 72].

Their variations [77] frequently communicate more than two data dimensions. The ways these additional dimensions are depicted can impact the scatterplot's usability, making it easier or harder to estimate trends [19] or compute differences across groups of data points [10, 34]. Computational models of these trade-offs exist for optimizing scatterplot designs for different tasks [60].

Such trade-offs in scatterplot designs become more complex as a function of visual channels used to encode data. The term *separability* refers to the ease with which our visual system can process one visual encoding without interference from another encoding [32, 62, 91]. For example, color and position encodings are separable, since our perception of position is robust to changes in color, and vice versa. But when visualizing data using red hue values for one measure and blue hue values for another, people will struggle to process each measure independently [91].

While position is generally considered separable from all other channels, evidence suggests that position may be integral with some channels for more complex visualization tasks [19, 51]. For example, position is integral with motion in outlier detection in multivariate scatterplots [89]. Recent studies provide formal models of separability across color, shape, and size for comparing data values [22, 80, 85], but we lack formal models for the separability of position with other channels in perceiving means and distributions.

A scatterplot's ability to support averaging and other ensemble tasks may also be affected by the underlying data distributions [51, 90]. While one study showed that averaging in scatterplots is generally robust to the size of data and a variety of tertiary encodings [34], positional outliers may bias trend perceptions [19], and performance on position summary tasks can vary across geometric scales [92]. In this study, we model

the separability of position and two common visual channels—size and lightness—for the position mean task across varying data distributions.

2.3 Ensemble Perception

Visualization tasks often involve extracting statistical summaries from data, such as mean position and size of data points; detecting color, size, or position outliers; spatial or feature-based segmentation; and judging correlation [87]. These tasks rely on our ability to rapidly summarize sets of visual elements through *ensemble processing* [6, 35, 87, 93] at a glance, even without focused attention to locations of individual objects [15, 21, 29, 45, 72, 95]. While ensemble perception is mostly beneficial for data comprehension, it may be susceptible to bias. Although mark color should be separable from position, Sun et al. [84] suggests color lightness will systematically bias position averaging. These biases can lead to illusory misjudgments when viewing dot plots.

Previous work [96] studied the impact of higher-level cognitive influences on position mean biases. Our work draws on vision science techniques to model bias in terms of visual elements using the regression models of the *centroid method* [83]. These models have been used to explain systematically biased mean estimates as a function of object lightness [83, 84], hue [82], size [74], orientation [48], and texture [83] in small sets of objects. However, these studies aim to isolate specific mechanisms, so stimuli are presented in subsecond durations. Only twelve or fewer targets are displayed, and no additional context such as labels or axes could divert attention. We extend their methods to realistic visualization scenarios to understand how biased data interpretations may arise as a function of design choices.

3 Hypotheses

Graph readers can use scatterplot means to characterize a range of values, compare classes in multiclass scatterplots [34], compare distributions over time or other facets [7, 53], or create decision thresholds [44, 81]. Studies in vision science showed that common visual variables like lightness or size can interfere with people's abilities to estimate mean properties for a small collection of objects [26, 74, 83]. However, these studies sought to measure perceptual mechanisms using flash tasks (i.e., stimuli were displayed for 500ms) and asking people to average fewer than 16 randomly distributed objects. These interference effects may not hold in visualization contexts, where people see a large number of marks, including the axes and ticks, for longer periods of time. To measure potential systematic bias in scatterplot averaging when adding lightness and size, we asked participants to indicate the average position of all data points in scatterplots where an irrelevant data dimension was mapped to either channel. Based on prior studies, we hypothesize:

***H1:* Scatterplot means will always be pulled towards locations of dark or larger points.**
Vision science studies demonstrate that locations of more salient (larger or darker) marks pull the perceived mean among small sets of objects. We expected these results to extend when reading trivariate scatterplots over the course of several seconds.

***H2:* Increasing correlations between the irrelevant channel and position will increase bias.**
Increased correlation between position and either size or lightness will cause dark or large data points to group together. If locations of dark or large data points pull mean position, we will see significantly biased mean responses directed towards regions where these points are located. Kim & Heer [51] demonstrated increased error rates with scatterplots with size variance, suggesting that size may have a stronger biasing effect than lightness.

***H3:* Widening the encoding range of the irrelevant channel will amplify bias.**
Increasing the visual difference among data points increases the differences in contrast across a scatterplot (e.g., the large marks would become larger, and the small marks smaller). If variations in contrast already shift the perceived position mean, increasing this difference would increase the ensuing bias.

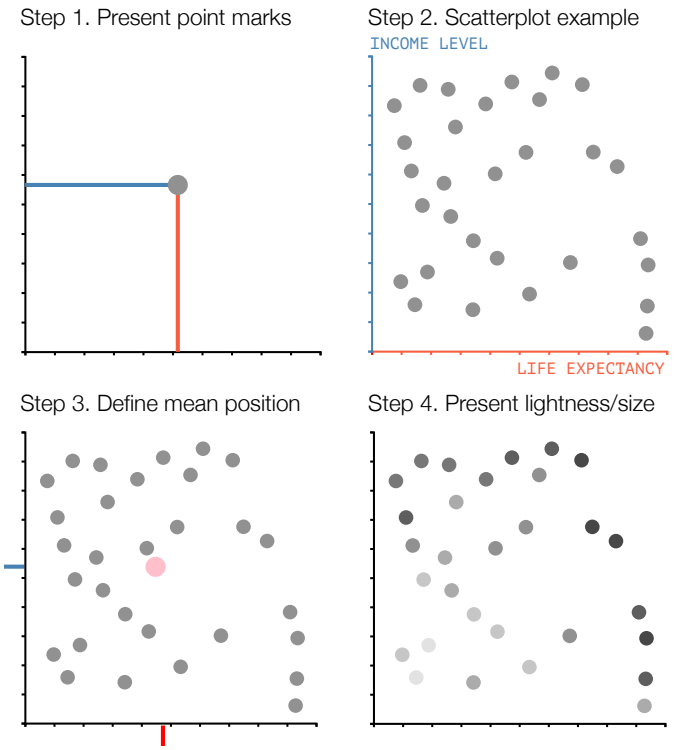

Fig. 3. To ensure comprehension, we stepped participants through the target task, steadily building in more information. The tutorial images above were presented serially with the following instructions: 1. Scatterplots reduce information about two measures into a single data point. 2. On this scatterplot, 30 countries are presented in terms of their life expectancies and income levels. 3. Therefore, the pink dot alone represents both the average life expectancy and the average income of the 30 countries shown. 4. Each point on a scatterplot can also depict a third variable, such as unemployment rate. After viewing the tutorial, participants were instructed that: "In the following study, you are asked to estimate and click on the average position of all points (i.e., average life expectancy and average income of all countries) on each scatterplot."

4 Methods

We investigated the effects of size and color on mean position estimation for trivariate scatterplots in two separate experiments: one investigating size and the other lightness. Each experiment was a 3 (encoding ranges) $\times$ 3 (correlations) within-subjects design. We measured performance using the vector between the reported means and the true means. We provide anonymized data and our study infrastructure at `https://osf.io/h8ft3/?view_only=9564278544b4411c82610c73daed8c00`.

4.1 Stimuli Generation

Our stimuli consisted of 500 $\times$ 500 pixel scatterplots generated using D3 (Figure 3). Each scatterplot was rendered on two orthogonal black axes with unlabeled tick marks every 50 pixels.

To generate the x- and y-data, we used Poisson disk sampling [50] to produce 30 uniquely distributed point grids, with minimum distance between the boundaries of any two points set at 8 pixels. This methodology is similar to Gleicher et al. [34]. Each dataset always contained 30 marks, with the number of points selected in piloting.

For each of the above datasets, we generated additional data for size and lightness to satisfy each of three spatial correlation levels: no correlation, low correlation, and high correlation. In the no correlation condition, the position of points had no correlation with the size or lightness data, providing a random distribution of the distractor encodings. In the low correlation condition, both the x- and y-positions of points were correlated with size or lightness by $\rho = 0.4 \pm 0.05$; in the high correlation condition, position was correlated with size or lightness

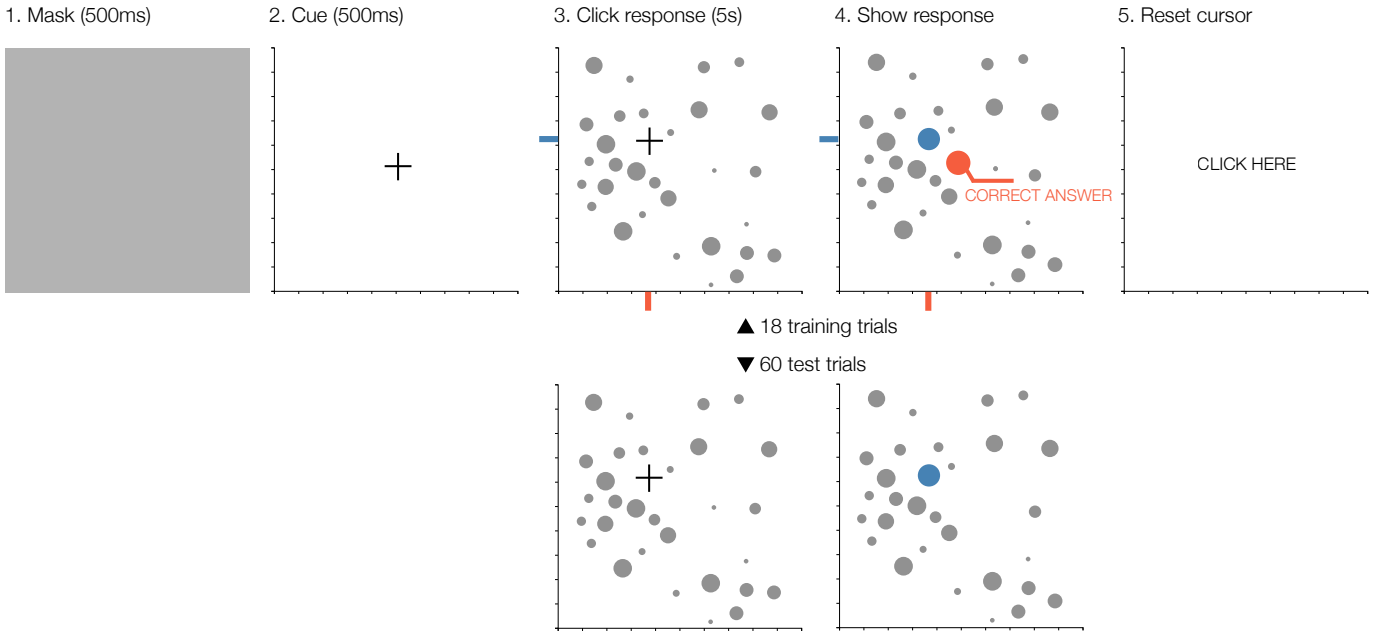

Fig. 4. The formal study consisted of five steps. A grey mask (1) would be replaced by a fixation cross (2) to guide participant attention. Participants would then see a scatterplot (3) and report the perceived mean position by clicking on the corresponding location in the scatterplot, with the cross moving with the cursor. We would visually indicate the reported value (4) and then ask participants to click on a link at the center of the screen (5) to recenter their mouse before the next trial.

by $\rho = 0.8 \pm 0.05$. We generated four datasets for each direction of correlation—either increasing or decreasing, along either the positive or negative diagonal—for each of the 30 point grids. These correlations between the third measure and the two position measures created a visual gradient (from light-to-dark, or small-to-large) along one of these four diagonals (Figure 1, right). This process led to 12 variations for each of the 30 scatterplot stimuli.

The encoding range for each distractor variable (size or lightness) corresponded to one of three levels: $45L^*$ to $75L^*$, $37.5L^*$ to $82.5L^*$, and $30L^*$ to $90L^*$ for lightness and 17.5px to 32.5px, 13.75px to 36.25px, and 10px to 40px for size, specified in terms of mark diameter (Figure 2). The linear ranges were sampled at seven evenly spaced values and shared the same middle value (i.e., the same size or lightness) to focus on he effect of range width rather than its midpoint. We chose the lightness ranges based on conventional L^* distributions in ColorBrewer [40], where the darkest colors are typically near 30L*, and the brightest colors are typically near 90L*. The ranges of size were restricted to those that would prevent occlusion given the 8 pixel limit on distance between points and align with default ranges found in commercial systems. In both size and lightness experiments, the narrowest range spanned half the width of the largest range.

Each participant saw 60 trials: 54 test trials varying in size or lightness and 6 control trials with no size or lightness variation to provide a baseline error rate. The test trials consisted of six trials for each combination of correlation (none, low, high) and encoding range (narrow, medium, wide). Data for each trial was randomly sampled from the set of pre-generated datasets with the correct corresponding correlation level. We presented the 60 trials in a random serial order.

4.2 Procedure

We conducted a Mechanical Turk experiment consisting of four phases: 1. informed consent, 2. instructions and tutorial, 3. formal trials, and 4. demographics. Participants were first provided with a consent form containing basic information about the data to be collected in the study. After providing consent, each participant passed an online Ishihara plate test to screen for color vision deficiencies [13].

After completing the screening, participants moved to a tutorial to ensure proper task understanding. The tutorial walked participants through the design of a trivariate scatterplot, first describing mark position, then demonstrating mean mark position, and finally introducing an additional channel using unemployment rate as an example measure. The tutorial, their accompanying figures, and text annotations were iterated through extensive piloting. During in-person piloting, we debriefed participants on our research goals, and confirmed that no participant interpreted the additional third measure as a factor that should affect the mean value of scatterplots. Figure 3 summarizes this aspect of the experiment.

After the tutorial, participants were instructed to "*Click on the average position of all points*" on each scatterplot, and this instruction persisted throughout both training and test trials. Training consisted of 18 trials where participants received immediate feedback by being presented the true mean position (Figure 4) after their click response. Prior work [83] utilized such feedback to collect more consistent estimates of mean position. During these practice trials, cursor movements were animated by moving reference lines along the x- and y-axis, reinforcing the idea that the participant was to average the x- and y-positions of the data points.

During the test trials, the interactive cursor guides and feedback were removed. Participants saw each of their 60 scatterplots in a random serial order. Each trial began with a 500ms gray mask followed by a fixation cross symbol to cue the participant that the trial was about to begin. After another 500ms, the scatterplot was rendered and participants had five seconds to click on the mean position of points in the scatterplot, with the time limit determined in piloting. A pink dot appeared briefly at the clicked location, signaling the trial's end. Participants could complete the task after the scatterplot was hidden, but delayed responses prompted an alert to encourage the participant to respond within the allotted time. Before beginning the next trial, participants had to move the cursor back to the center of the stimulus by clicking a link in the middle of the scatterplot to reset their cursor position, minimizing potential motor bias.

We interspersed four engagement checks in the formal study to assess honest participation. On average, 7.5 trials passed before the first engagement check, and then 15 trials passed before each successive check. During these engagement checks, a single data point was shown in one of the four quadrants, and participants were removed from the

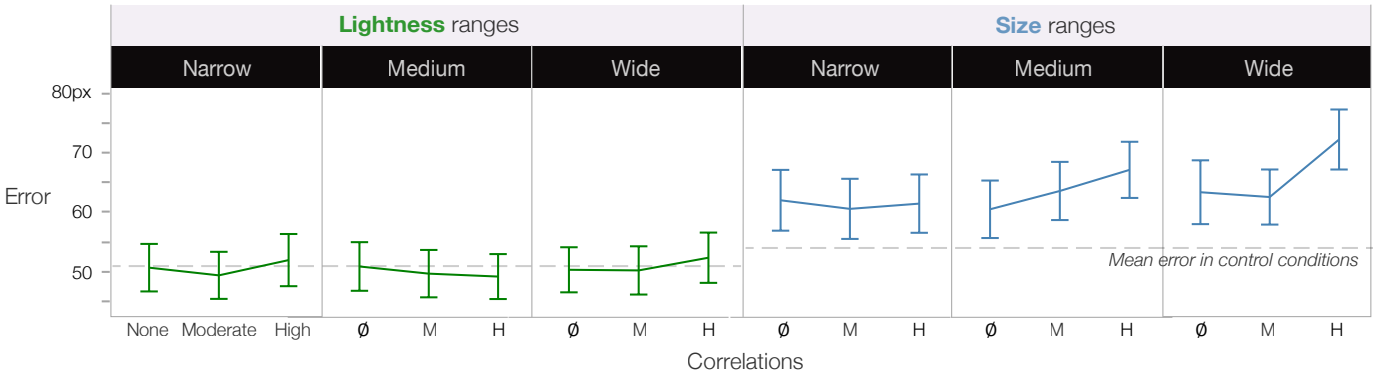

Fig. 5. Mean error magnitude across encoding type, encoding range, and correlations. Note the y-axis represents pixel values, and our stimuli size was 500*px* by 500*px*. Using lightness (green lines) did not significantly increase error over the baseline. In the size experiment (blue lines), errors were significantly higher than the baseline condition for all ranges. Error bars represent 95% confidence intervals.

analysis if they failed two or more checks. After completing all trials, participants completed a demographic survey and were compensated for their participation.

4.3 Measures & Analysis

Error was measured as the magnitude of the error vector between the true mean and the reported mean. We additionally computed the directional bias in responses by projecting the error vectors to the direction of the correlation gradient.

We analyzed these measures using a two-factor (encoding range, and correlation level) ANCOVA with trial order, direction of the correlation gradient, the specific datasets, and interparticipant variation as random covariates. We examined both primary effects and first-order interaction effects and used Tukey's Honest Significant Difference (HSD) test ($\alpha = .05$) with Bonferroni correction for post-hoc comparisons. These analyses excluded the control scatterplots (those with no lightness or size differences) as those trials reflect no measurable correlations or encoding ranges. However, we used Dunnett's Method to identify where error or bias significantly differed from baseline performance measured in these trials.

4.4 Participants

We recruited 174 subjects with US IP addresses from Amazon's Mechanical Turk. We excluded 22 participants with limited or no cursor movements, flagged by back-to-back clicks on the same pixel position, and another 22 participants who failed the engagement trials, leaving us with a final sample size of 130. These participants were between 20 and 71 years of age ($\mu = 37.3, \sigma = 10.5$). 108 (83.1%) participants used mouse clicks and 21 (16.1%) participants used a trackpad. One participant used a touchscreen. Participants were compensated $1.75 for their time. On average, the experiment took 7.8 minutes.

Crowdsourcing platforms exchange some control for ecological validity: sizes and colors may be affected by the display and environment used by each participant. However, this variety reflects visualization viewing in practice and has been shown to produce reliable models in past visualization research [42, 51, 52, 80, 85].

5 Results

We report inferential statistics, means, and 95% bootstrapped confidence intervals (means $\pm$ 95% confidence intervals) for relevant effects in accordance with guidelines for transparent statistical communication [25]. Means and confidence intervals are reported in pixel values: our stimuli size was 500*px* by 500*px*.

5.1 Error

Figure 5 summarizes error rates across the experimental conditions. We did not find any significant differences of error in the lightness experiment. In the size experiment, there was a significant interaction effect of correlations and size range on error rates ($F(4,63) = 3.02, p <$ .02). Tukey's HSD reveals that error rates rose dramatically in the high-correlations, high-size range condition ($\mu = 70.2px \pm 5.0$).

We used Dunnett's Method to compare errors in the control condition and the three experimental conditions for both experiments. We found no significant differences between the control condition and the three lightness conditions. However, each size condition introduced significantly greater error rates than the control condition (narrow size ranges: $p < .05$; middle size ranges: $p < .005$; large size ranges $p < .0001$).

5.2 Bias

While the previous section showed rates of *precision* that were comparable across conditions, *bias* in participants' responses may still increase, making error rates a less reliable measure in evaluating position mean perception.

To illustrate, assume that participant responses in the control condition were normally and randomly distributed around the true mean, and that size or lightness pulled these responses northeast. In this case, some responses (i.e., those already northeast of the true mean) are pulled further away from the true mean, potentially increasing the error rates. However, some responses (i.e., those southwest of the true mean) are simultaneously pulled closer to the true mean; these displacements can effectively balance out the above shifts in error rates.

Correlations between the distractor encoding and position created a visual gradient (from light-to-dark, or small-to-large) along one of the four diagonals. We measured bias as a function of the amount of signed error displaced along each scatterplot's direction of gradient. Overall, position means were biased toward the direction of increasing size or darkness, with the magnitude of the bias increasing with correlation between the distractor encoding and position. Figure 6 summarizes the results.

Lightness: While we did not find a significant effect of increasing lightness range alone ($F(2,67) = 0.77, p < .5$), bias in the perceived mean increased as the correlation between position and lightness increased ($F(2,67) = 4.51, p < .02$). Highly correlated scatterplots with a wide lightness range ($\mu = 22.2px \pm 5.4$) were significantly more biased than all other conditions except the other highly correlated scatterplots (narrow range: $\mu = 17.1px \pm 5.4$; middle range: $\mu = 15.8px \pm 5.4$). The highly correlated conditions with narrow and middle ranges displayed significantly higher bias than uncorrelated data with moderate ($\mu = 7.3px \pm 5.4$) and wide ($\mu = 17.1px \pm 5.5$) lightness ranges.

Size: While we did not find a significant effect of increasing size range alone ($F(2,63) = 0.47, p < .7$), there was a significant interaction effect between increasing correlations and increasing size levels ($F(4,63) = 2.44, p < .05$). Increasing correlation between size and position led to significant increases in bias ($F(2,63) = 15.22, p < .0001$). Specifically, in the high correlations conditions, middle ($\mu = 32.91px \pm 9.6$) and wide size ranges ($\mu = 36.0px \pm 9.7$) were

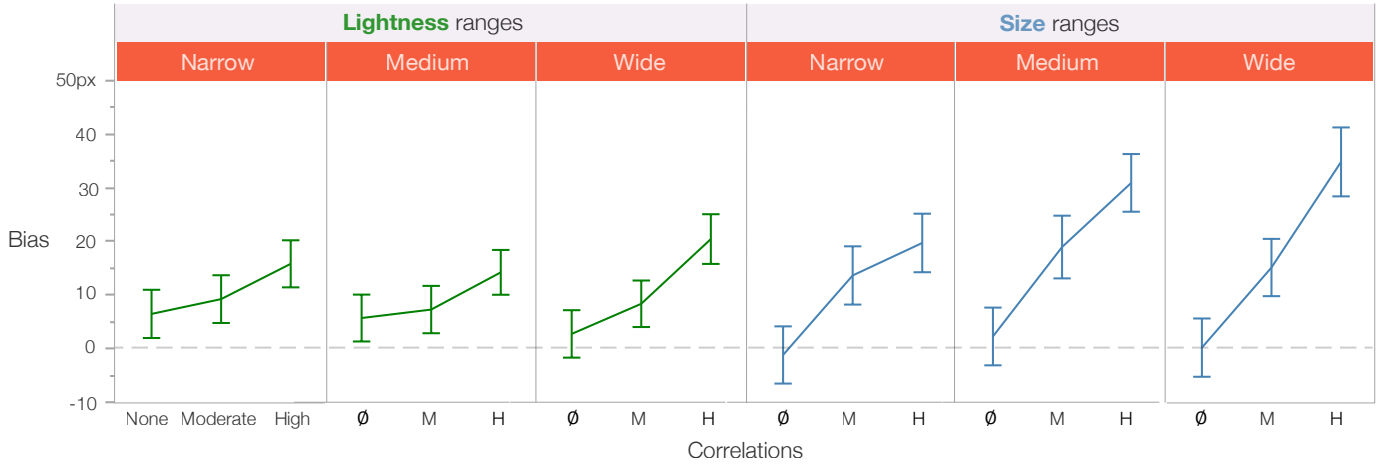

Fig. 6. Mean bias across encoding type, encoding range, and correlations. Note the y-axis represents pixel values, and our stimuli size was 500*px* by 500*px*. Errors in both lightness (green) and size (blue) exhibited systematic biases towards locations of larger or darker points. These effects were amplified by increased correlations. Error bars represent 95% CIs.

significantly more biased than the narrow size range ($\mu = 20.5px \pm 9.7$). Within each size range condition, if correlations were greater, bias was greater. The exception to this pattern was the narrow size range, where increasing correlations from low ($\mu = 15.2px \pm 9.7$) to high ($\mu = 20.5px \pm 9.7$) did not significantly affect bias.

6 Modeling Feature-Based Attention

Mean position estimates were biased toward locations of larger and (to a lesser extent) darker points, creating a *weighted average illusion* giving more weight to areas with larger or darker marks. However, our ANCOVA results do not give an account of *why* the bias emerges. What elements of a scatterplot's design might be biasing position mean perception?

While our results raise many phenomenological questions (see §7.2 for a discussion), it will be helpful for visualization practice to build a more predictive model of this bias as a function of visual elements and data distributions. With such a model, practitioners can better reason about design trade-offs by predicting magnitudes of bias across designs and distributions. This motivates our models of observed bias using the centroid method [83]. This technique uses linear regression to measure how much of the observed errors in participants' position mean responses is attributable to varying attention given to mark features, such as its specific size or lightness.

Our primary goal with this analysis was to model bias in the no-correlation conditions, where size and lightness were randomly distributed. In such cases, we could not measure bias as signed errors along directions of the correlation gradient as there was no correlation gradient to project against. Models of this condition can provide a baseline rate of bias for any given trivariate scatterplot. Since it is less likely that random scatterplots contain global features like a texture density gradient, which is a potentially confounding factor [37], this model will help us answer whether people attend to data points with certain features differently toward the mean.

6.1 Feature-Based Attention

Feature-based attention is an attention mechanism that operates in a parallel, distributed manner across space [4]. It does so by modulating the gain of regions in the visual cortex selective for a feature (e.g., a given color or shape) [79]. When performing a visual summary task, such as comparing average values in multiclass scatterplots [34], feature-based attention helps viewers attend to each set of point marks separately before comparing their average values.

Feature-based attention is usually considered a selective attention mechanism in visualization [87]. However, the distribution of attention given to objects that vary across a continuous channel, such as size or lightness, need not be selective. Attention can be automatically attuned to one type of objects relatively more than others (e.g., large marks over small marks), while still being distributed across all objects to execute a summary task like position averaging [83]. In other words, feature-based attention may automatically "weigh" certain kinds of marks relatively more than others when summarizing data.

The centroid method is a linear regression model of feature-based attention that models such weights distributed across a set of point marks. Much like eye tracking, the centroid computation provides a behavioral marker that act as a proxy for studying attention. However, unlike eye-tracking, the centroid method allows us to model attention being distributed across multiple locations in parallel by estimating the weight given to certain kinds of marks based on their visual features (e.g., size or lightness levels).

6.2 The Centroid Method

The centroid method [83] defines a weight function $w(\tau)$, where $\sum^{\tau} w(\tau) = 1$, as the weight given to each item of type τ. This weight function w is called the *attention filter*, a model of feature-based attention [12]—our ability to attend to items with specific visual features more, or less, than other items—which may cause the bias towards larger and darker points. This attention filter corresponds to the visual 'weight' of individual marks. We discuss the role of feature-based attention in interpreting our results in §7.2.1.

We model a participant's mean estimate $(R_{t,x}, R_{t,y})$ on scatterplot trial t as

$$R_{t,x} = V\mu_{t,x} + (1-V)x_{default} + Q_{t,x}$$
$$R_{t,y} = V\mu_{t,y} + (1-V)y_{default} + Q_{t,y}$$

where $\mu_{t,x}$ and $\mu_{t,x}$ are coordinates of the true mean position of the points; V, where $0 \leq V \leq 1$, is the *Data-Drivenness* parameter—a measure of how much participants depended on a default location $(x_{default}, y_{default})$ (e.g., the center of the graph) rather than true point positions to compute mean position—and $Q_{t,x}$ and $Q_{t,y}$ are independent and normally distributed random response errors.

We can define $\mu_{t,x}$ and $\mu_{t,x}$ as the weighted sums of mark coordinates divided by the sum of all weights:

$$\mu_{t,x} = \frac{\sum_{i=1}^{N_{stims}} w(\tau_{t,i})x_{t,i}}{\sum_{i=1}^{N_{stims}} w(\tau_{t,i})}, \mu_{t,y} = \frac{\sum_{i=1}^{N_{stims}} w(\tau_{t,i})y_{t,i}}{\sum_{i=1}^{N_{stims}} w(\tau_{t,i})}$$

where $\tau_{t,i}$ is the item type of mark i in trial t, w is the attention filter (the visual 'weight' of each mark), and $x_{t,i}$ and $y_{t,i}$ are coordinates of mark i in trial t.

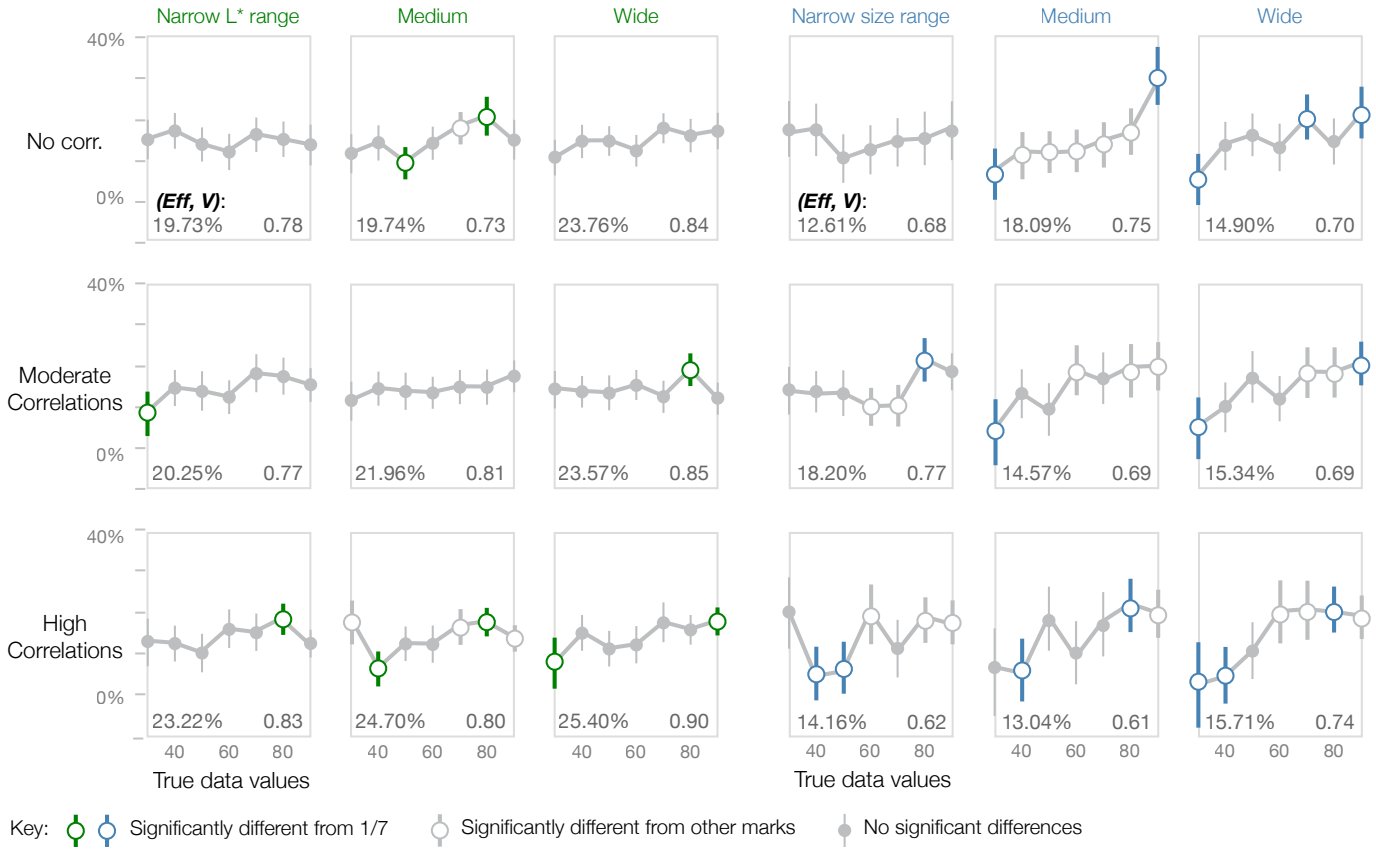

Fig. 7. Weights describing the contribution of classes of marks to the observed bias derived using the centroid method (means and 95% confidence intervals). White marks indicate points that have significantly different weights and highlighted marks indicate where weights differed significantly from equal weighting. An upward slope indicates that people give greater weight to larger or darker marks.

The process for generating point and interval estimates for w is outlined in Appendix 1 and 2 of Sun et al. [83]. The point estimation finds the maximum likelihood estimate of w via linear regression. The interval estimation uses Fieller's theorem [30] for calculating a 95% confidence interval for the ratio of two means. The interval estimates give the expected variance of visual weights given to each mark across the population.

6.3 Results

Our stimuli contained marks of seven different sizes or lightnesses, which we used to represent our mark categories τ. If participants weighed all marks equally regardless of size or lightness, the weight for any mark would be $1/7$ (14.29%). This threshold provides a baseline for determining whether participants weighed certain marks more heavily than others. Figure 7 summarizes the weight distributions across conditions. The leftmost values within each plot correspond to the smallest or brightest marks in each scatterplot, and the rightmost values correspond to the largest or darkest marks in each scatterplot.

We found that even when size had no correlation with position (Figure 7, top row, blue), people weighed the smallest marks significantly less than and the largest marks significantly more for moderate and wide size ranges. When data was uncorrelated and the encoding ranges were narrow (top row, 1st and 4th columns), participants weighed marks roughly equally.

We found that these weights varied significantly along with variations in range widths and correlations. In general, people weighed marks more heavily as they became darker or larger, correlated with the corresponding increase in bias toward locations of those marks from §5.2. We also found evidence that these weights, rather than a default response like clicking in the center of the graph, explain the errors we found in our data. The Data-Drivenness (V) of click responses were 81.09% for lightness, and 69.46% for size, comparable to those found by Sun et. al. [83] in in-person laboratory studies.

The centroid method allows us to compute how much each mark type contributes to the perceived means. We can use these weights to predict where people are likely to see the mean on a scatterplot (Fig. 1). While we focused on lightness and size, this weighting approach has been used with other visual features like orientation [48] and hue [82]. Future work could similarly extend the approach to other kinds of visualizations, such as bar charts or line charts, where position is redundantly coded using height or orientation. Further, the centroid method offers explicit insight into how visualization designs direct attention, which we explore in §7.2.1.

7 Discussion

We explored the relationship between size, lightness, and positional means in trivariate scatterplots. Our results show that the perceived mean of a scatterplot is biased towards larger or darker points (***H1***). We have labeled this bias the weighted average illusion because the bias can be explained by asymmetries in weights we assign to marks based on irrelevant properties like size and color. We found that these effects were robust to training and increased as the structure in the data and range of size or lightness increased (§5). Bias always increased as correlations between position and the third data dimension increased. This bias was directed toward areas of larger or darker points and, in the strongest conditions, caused people to misread the average by 35 pixels, supporting ***H2***. Widening size ranges also affected bias as correlations increased, partially supporting ***H3***. While these effects were stronger for size than lightness, they demonstrate the predictability of this bias as a function of data patterns and design choices (Figure 6).

7.1 Design Implications

Our results indicate that, despite classical guidelines that position is robust to other encodings [91], adding additional data to scatterplots can interfere with and even bias people's abilities to reason over position. These biases likely extend to other visualizations using both size and position to communicate data, such as Augmented Stripplots [71]. While we explore this phenomena in the context of mean judgments, bias may occur in other summary tasks that rely on similar visual processes, like variance estimation and correlation. Designers can use the observed bias and weights to predict when this bias might occur and use alternative design strategies, like providing reference lines or explicit summary values, to support critical summary tasks.

Size is a more precise channel for communicating data than lightness [14], but our study indicates that size interferes with position summaries more significantly. Increasing the range of sizes in visualization communicates more precise data differences but introduces significantly more bias into scatterplot position means. Further, bias in bubble charts may be much greater in practice than demonstrated in our results where our tested ranges were limited to prevent occlusion. Error rates in the lightness experiment were often comparable to those in the control condition. Designers should consider these trade-offs given their target audience, datasets, and the visualization tasks at hand: if assessing global properties, such as means or variances, are more important than comparing values, lightness may provide a more robust channel. If interpreting individual values is more critical, size may be preferable.

Using the centroid paradigm, designers can run exploratory self-studies to quantify the degree to which readers can summarize relevant data when distracting categorical (e.g., hue palettes [62, 82]) or continuous channels (e.g., size or lightness) are present. When the relevant and distractor channels may not seem separable, weights modeled using the centroid method can be used to estimate a baseline rate of bias. For example, by simply computing the average of mark locations weighed according to our attention filters presented in Figure 7, designers can predict which data points will be seen as being above or below average. This is how the predicted perceived means in Figure 1 (open circles) were computed. Developing a more extensive set of models covering a larger range of designs could provide significant predictive power for predicting bias in information design.

Both current and prior work suggest that readers can be trained to mitigate this bias in small collections of objects. While our results exhibited a consistent bias, there can be significant individual differences in how much weight people give to each mark during averaging that can be reduced with sufficient training [26]. In pilot studies, we found that without training, people exhibited a wider range of bias magnitudes but these biases converged over time. We account for this in our study with our training phase, where participants receive feedback for 18 trials before entering the formal study. However, these findings suggest the possibility that certain kinds of biases can be mitigated by training.

7.2 Potential Basis for the Weighted Average Illusion

By modeling how different elements of a design may bias mean estimates, we can generate new hypotheses for visualization psychology that allow us to anticipate *why* this bias might arise and *when* it may impact other visualization types or tasks. Many ensemble coding strategies, proxies, and heuristics can support or combine to help people compute visual summary statistics. We expand on what the centroid method results (§6.2) tell designers about attention in trivariate scatterplots and propose two possible mental shortcuts that people could be using to compute mean positions: 1. sampling only a subset of marks that could be held in working memory (Figure 8-1), and 2. comparing texture densities between spatial segments of the data (Figure 8-2).

7.2.1 Local Basis: Feature-Based Attention

The position mean is a reliable behavioral marker for studying feature-based attention in vision science [83] and information visualization [34]. Feature-based attention implies that attention is unevenly distributed across space, skewing towards marks with certain visual features [12]. The attention filters modeled with the centroid method quantify this skew [83]. When reading visualizations, we rely on feature-based attention to, for example, make judgments about different data categories [34] or search for relevant data [38].

Figure 8-1. Subsampling

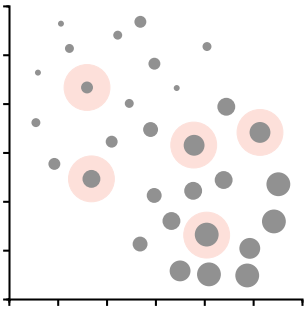

Figure 8-2. Density strategy

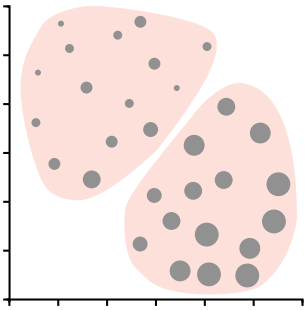

Fig. 8. Potential strategies employed by participants. (Left) Participants could have been focusing on a small subset of point marks, shortcutting the need to attend to all points. This would bias the mean position if darker or larger marks are more likely to be focused on. (Right) Participants could have segmented highly correlated scatterplots into sets of smaller points and sets of larger points, then computing the midpoint of those segments weighted by estimates of density in each segment. This strategy amounts to asking the question "How many points are here, as opposed to there?" instead of directly computing the mean position. Although the segment with larger marks might seem to contain more data, both segments contain 15 marks.

Although we did not use eye tracking in our study (as discussed in §7.3), prior work already confirm that dark and large marks are more salient than light and small marks [41]. This implies that participants' eye movements may have reflexively saccaded to the locations of dark or large objects as soon as each scatterplot was presented. Such behavior might a priming effect on computing the mean position of scatterplot marks, causing feature-based attention to be distributed more towards larger or darker marks and potentially leading to the skewed attention filters modeled in Figure 7.

However, if feature-based attention was the only factor at play, we would see the same attention filters in each correlation condition, varying only as a function of increasing size or lightness ranges. Instead, we found that the attention filters varied with increasing correlations in the data as well, indicating there is more to the observed bias than simple pop-out effects. People may alternatively be using subsampling or density-driven strategies (Figure 8). We discuss these strategies below as both potential explanations for the observed bias as well as opportunities to inform future designs.

7.2.2 Hyperlocal Basis: The Subsampling Strategy

Myczek and Simons [63] demonstrated that subsampling four objects from a larger collection can account for performance on size averaging. However, other studies show that people may need significantly more information to make sense of eight or more objects [5], far fewer than the number of marks in a typical visualization, with later work estimating that the number of samples must be closer to the square root of the number of marks [93]. Using the centroid method, we can compute the minimum portion of marks, called the *Efficiency*, or *Eff*, required to converge onto our models in §6.2 (w, V, $x_{default}$ and $y_{default}$). In other words, this measure of efficiency quantifies the minimum number of objects that would have to be used in computing position mean to achieve the same level of performance achieved by the participants.

Given the residual sum of squares $SS_{Residual}$ derived using the centroid method, the standard error $\widehat{\sigma}$ provides an unbiased estimate of the standard deviations of the random response errors $Q_{t,x}$ and $Q_{t,x}$:

$$\widehat{\sigma} = \sqrt{\frac{SS_{\text{Residual}}}{df}}$$

We assume as in prior work [5] that all error in $\widehat{\sigma}$ was due to the number of items unattended to by participants. To compute *Eff*, we use the technique from Sun et al. [83] to calculate the variance of the difference between participants' responses and the predicted responses using w, V, $x_{default}$, and $y_{default}$ after removing N marks. Starting by deleting only $N = 1$ mark from each stimulus scatterplot, this variance is iteratively computed for each possible number of marks that can be deleted until a single mark is remaining ($0 < N < 30$), terminating once the variance is greater than $\widehat{\sigma}$.

Averaging over the experimental conditions, our models in Figure 7 could have been achieved by attending to as little as 22.48% of marks (around 7 marks) in the lightness experiment, and 15.18% of marks (around 5 marks) in the size experiment, aligning with the square root of the number of marks predicted in Whitney & Leib [93]. These results may imply that people attend to a small number of marks when summarizing trivariate scatterplots. Our measured bias could arise if dark or large marks capture more attention and are disproportionately represented in participants' subsamples. If people are subsampling marks, then we anticipate that this bias may arise in other visualizations that use discrete marks, such as bar charts. This bias may be reduced by using continuous representations like kernel-density estimation [28].

7.2.3 Global Basis: The Density Strategy

Attention filters in §6.2 varied as a function of both encoding ranges and correlations, suggesting that clustering darker and larger marks together may introduce other factors that increase the observed bias. One factor may be that the density across different spatial segments also pulls the perceived mean.

We use the term *density* to refer to any of three mechanisms possibly involved numerosity estimation in a group of marks: subitization (when discriminating 10 or fewer data points), estimation (for more than 10 data points), and low spatial frequency features such as texture density or contrast energy [68] (for review, see Picon et al. [66]). People can estimate the densities of overlapping dot textures of different colors in parallel [36] and visualization techniques leverage density to deal with challenges like overdraw [28, 56].

When a scatterplot depicts a third data dimension and this dimension is correlated with position, the resulting clusters will share the same visual features. For example, a bubblechart will have clusters of mostly large and mostly small dots. When clusters emerge, the visual system may immediately segment the scene and then estimate the density of the resulting segments [31].

Some participants might have computed the midpoint of those centroid estimates, weighted by estimates of density in each segment. Prior work confirmed that large or dark points clustering together within an area creates illusions of higher point density [20,27,33,47,54,61,66,88]. If people use the positions and densities of spatial segments to estimate the overall mean, these illusions will bias the mean position towards clusters that contain large or dark marks. This strategy would imply that visualizations that mitigate spatial clustering will mitigate the observed bias.

7.3 Limitations & Future Work

We investigated two common visual channels—size and color—that are frequently used to encode additional information in scatterplots. However, the simplicity of scatterplots affords a large space of potential designs that may offer different perceptual trade-offs [77]. Future work might consider these alternatives to understand the robustness of mean position perception across scatterplot designs and how this bias may influence a broader range of ensemble tasks like correlation and variance estimation. For example, the centroid method employed here could be used to understand biases introduced by a range of visual variables for both categorical and numeric data. Further, we measured weights along a small set of size and lightness ranges. While these weights allow us to reasonably estimate the probable location of a perceived mean over these ranges, we do not have sufficient data for a fully predictive model. Modeling weights across a wider range of factors may enable robust bias predictions across a range of encoding lengths and designs.

Our experimental stimuli were generated according to Poisson disk sampling, creating random dot textures that adhere to the assumptions of the centroid method. While these textures contain a range of clusters and structures, real datasets often have additional statistical features that can cause data points to more strongly cluster and to form a larger range of cluster shapes. These factors could introduce additional biases in realistic scatterplots and may limit the generalizability of the computed weights. Future work should investigate how the visual structure of scatterplots might shift this bias. Additionally, our data generation approach led to scatterplots where the direction of correlation was slightly displaced from the true diagonal. Although these displacements are small and randomly distributed, since we define bias as the magnitude of the error vectors projected onto the diagonals, the bias illustrated in Figure 6 may underestimate the true bias.

While our crowdsourcing study led to consistent results, the attention filters derived in §5.2 may vary across individuals. For example, certain people can discount the contribution of peripheral objects toward the mean [26]. These individual differences may shift the ways people reason about means, especially for visualizations targeting data within specific areas of expertise [49]. Disciplines may exaggerate bias by introducing semantic factors such as context or risk that add mental "weight" to critical data. Future work might leverage the centroid method to model these individual differences under varying data contexts to the influence of disciplinary knowledge and other factors on bias.

Lastly, we identified two potential strategies that might explain the observed bias. Given that feature-based attention operates prior to voluntary eye movements [57] and constraints of the COVID-19 pandemic, we did not incorporate eye tracking in our study. However, eye-tracking may provide further insight into these strategies. Future work should combine the centroid computation and eye tracking, both state-of-the-art behavioral markers for studying visuospatial attention, to investigate the subsampling and density hypotheses.

8 Conclusion

Our abilities to estimate summary statistics from scatterplots may be sensitive to attributes of the scatterplot design and underlying data [51]. We conducted a crowdsourced experiment investigating how irrelevant encodings, the ranges of those encodings, and relationhips between position and other dimensions may shift position averaging. We found that the perceived mean position in a scatterplot is biased in the direction of larger, and to a lesser extent, darker marks. We model the contribution of different elements of a visualization design towards this bias using the centroid method, a statistical technique from vision science, offering designers a way to predict weights people give to a given mark.

Our results raise new opportunities at the intersection of visualization and vision science: we elucidate a systematic bias that gives insight into feature-based attention in visualization as well as a variety of ensemble strategies that can be employed in visualization interpretation and decision making. For designers, modeling bias as a function of design choices and data patterns allows designers to avoid misleading practices and experienced graph readers to notice and correct for potential bias.

Research in visualization biases tend to develop in parallel with research in visualization literacy, since biases often arise when novice readers are not familiar with a visualization [55]. The weighted average illusion adds to the growing literature of visualization biases [3, 11, 23, 24, 69, 90, 96] that may affect interpretations of even the most familiar visualizations. In our pilot studies with visualization researchers, even visualization experts had a hard time avoiding this bias, often naming justifications for weighing point marks differently, such as: "What if the size channel represented population size?" We note that this illusion is a misjudgment regardless of the measure lightness or size represents, since weighted averages change with the ranges of lightness or size a designer chooses to map the data onto as opposed to the data itself. By surfacing these biases, we can both promote honest communication in visualization practice and guide the development of data literacy for everyone.

REFERENCES

[1] M. M. Abbas, M. Aupetit, M. Sedlmair, and H. Bensmail. Clustme: A visual quality measure for ranking monochrome scatterplots based on cluster patterns. In *Computer Graphics Forum*, vol. 38, pp. 225–236. Wiley Online Library, 2019.

[2] D. Albers, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 551–560, 2014.

[3] E. C. Alexander, C. C. Chang, M. Shimabukuro, S. Franconeri, C. Collins, and M. Gleicher. Perceptual Biases in Font Size as a Data Encoding. *IEEE Transactions on Visualization and Computer Graphics*, 24(8):2397–2410, 2018. doi: 10.1109/TVCG.2017.2723397

[4] G. A. Alvarez. Representing multiple objects as an ensemble enhances visual cognition, 2011. doi: 10.1016/j.tics.2011.01.003

[5] G. A. Alvarez and A. Oliva. The representation of simple ensemble visual features outside the focus of attention: Research article. *Psychological Science*, 19(4):392–398, 2008. doi: 10.1111/j.1467-9280.2008.02098.x

[6] D. Ariely. Seeing sets: Representation by statistical properties. *Psychological science*, 12(2):157–162, 2001.

[7] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps (Translated by William J. Berg)*. University of Wisconsin Press, 1983.

[8] E. Bertini, M. Correll, and S. Franconeri. Why Shouldn't All Charts Be Scatter Plots? Beyond Precision-Driven Visualizations. pp. 206–210, 2021. doi: 10.1109/vis47514.2020.00048

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

[10] D. Burlinson, K. Subramanian, and P. Goolkasian. Open vs. closed shapes: New perceptual categories? *IEEE transactions on visualization and computer graphics*, 24(1):574–583, 2017.

[11] A. Calero Valdez, M. Ziefle, and M. Sedlmair. Studying Biases in Visualization Research: Framework and Methods. *Cognitive Biases in Visualizations*, pp. 13–27, 2018. doi: 10.1007/978-3-319-95831-6_2

[12] M. Carrasco. Visual attention: The past 25 years. *Vision Research*, 51(13):1484–1525, July 2011. doi: 10.1016/j.visres.2011.04.012

[13] J. Clark. The ishihara test for color blindness. *American Journal of Physiological Optics*, 1924.

[14] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods, 1984. doi: 10.1080/01621459.1984.10478080

[15] M. A. Cohen, D. C. Dennett, and N. Kanwisher. Ensemble Perception, Summary Statistics, and Perceptual Awareness: A Response, sep 2016. doi: 10.1016/j.tics.2016.06.007

[16] M. Correll, D. Albers, S. Franconeri, and M. Gleicher. Comparing averages in time series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1095–1104, 2012.

[17] M. Correll, E. Bertini, and S. Franconeri. Truncating the y-axis: Threat or menace? In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020.

[18] M. Correll and J. Heer. Black hat visualization. In *Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVe), IEEE VIS*, 2017.

[19] M. Correll and J. Heer. Regression by eye: Estimating trends in bivariate visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1387–1396, 2017.

[20] S. C. Dakin, M. S. Tibber, J. A. Greenwood, F. A. Kingdom, and M. J. Morgan. A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49):19552–19557, 2011. doi: 10.1073/pnas.1113195108

[21] S. C. Dakin and R. Watt. The computation of orientation statistics from visual texture. *Vision research*, 37(22):3181–3192, 1997.

[22] Ç. Demiralp, M. S. Bernstein, and J. Heer. Learning perceptual kernels for visualization design. *IEEE transactions on visualization and computer graphics*, 20(12):1933–1942, 2014.

[23] E. Dimara, G. Bailly, A. Bezerianos, and S. Franconeri. Mitigating the Attraction Effect with Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):850–860, 2019. doi: 10.1109/TVCG.2018.2865233

[24] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic. A task-based taxonomy of cognitive biases for information visualization. *IEEE transactions on visualization and computer graphics*, 2018.

[25] P. Dragicevic. Fair statistical communication in hci. In *Modern statistical methods for HCI*, pp. 291–330. Springer, 2016.

[26] S. A. Drew, C. F. Chubb, and G. Sperling. Precise attention filters for Weber contrast derived from centroid estimations. *Journal of Vision*, 10(10), 2010. doi: 10.1167/10.10.20

[27] F. H. Durgin. Texture Density Adaptation and the Perceived Numerosity and Distribution of Texture. *Journal of Experimental Psychology: Human Perception and Performance*, 21(1):149–169, 1995. doi: 10.1037/0096-1523.21.1.149

[28] P. H. Eilers and J. J. Goeman. Enhancing scatterplots with smoothed densities. *Bioinformatics*, 20(5):623–628, 2004.

[29] L. Feigenson. Objects, Sets, and Ensembles. *Space, Time and Number in the Brain*, pp. 13–22, 2011. doi: 10.1016/B978-0-12-385948-8.00002-5

[30] E. C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):175–185, 1954.

[31] S. L. Franconeri, D. K. Bemis, and G. A. Alvarez. Number estimation relies on a set of segmented objects. *Cognition*, 113(1):1–13, oct 2009. doi: 10.1016/j.cognition.2009.07.002

[32] W. R. Garner. Interaction of stimulus dimensions in concept and choice processes. *Cognitive psychology*, 8(1):98–123, 1976.

[33] T. Gebuis and B. Reynvoet. The interplay between nonsymbolic number and its continuous visual properties. *Journal of Experimental Psychology: General*, 141(4):642–648, 2012. doi: 10.1037/a0026218

[34] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2316–2325, 2013. doi: 10.1109/TVCG.2013.183

[35] J. Haberman and D. Whitney. Ensemble perception: Summarizing the scene and broadening the limits of visual processing. *From perception to consciousness: Searching with Anne Treisman*, pp. 339–349, 2012.

[36] J. Halberda, S. F. Sires, and L. Feigenson. Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, 17(7):572–576, 2006. doi: 10.1111/j.1467-9280.2006.01746.x

[37] S. Haroz and D. Whitney. How Capacity Limits of Attention Influence Information Visualization Effectiveness. 18(12):2402–2410, 2012.

[38] S. Haroz and D. Whitney. How capacity limits of attention influence information visualization effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2402–2410, 2012.

[39] 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.

[40] M. Harrower and C. A. Brewer. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, 40(1):27–37, 2003. doi: 10.1179/000870403235002042

[41] C. G. Healey and J. T. Enns. Attention and Visual Memory in Visualization and Computer Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, jul 2012. doi: 10.1109/TVCG.2011.127

[42] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 203–212, 2010.

[43] S. Hessney, R. Peck, and K. Dickensheets. What's going on in this graph? — easing lockdowns. *The New York Times*, Dec 2020.

[44] S. Hessney, R. Peck, and A. Fetter. What's going on in this graph? — april 17, 2019. *The New York Times*, Apr 2019.

[45] S. Hochstein, M. Pavlovskaya, Y. S. Bonneh, and N. Soroker. Comparing set summary statistics and outlier pop out in vision. *Journal of Vision*, 18(13):12–12, 2018.

[46] A. K. Hopkins, M. Correll, and A. Satyanarayan. Visualint: Sketchy in situ annotations of chart construction errors. In *Computer Graphics Forum*, vol. 39, pp. 219–228. Wiley Online Library, 2020.

[47] F. Hurewitz, R. Gelman, and B. Schnitzer. Sometimes area counts more than number. *Proceedings of the National Academy of Sciences of the United States of America*, 103(51):19599–19604, 2006. doi: 10.1073/pnas.0609485103

[48] M. Inverso, P. Sun, C. Chubb, C. E. Wright, and G. Sperling. Evidence against global attention filters selective for absolute bar-orientation in human vision. *Attention, Perception, & Psychophysics*, 78(1):293–308, 2016.

[49] K. A. Kastens, T. F. Shipley, A. P. Boone, and F. Straccia. What geoscience experts and novices look at, and what they see, when viewing data visualizations. *Journal of Astronomy & Earth Sciences Education*, 3(1):27–58, 2016.

[50] kchapelier. poisson-disk-sampling.js.

[51] Y. Kim and J. Heer. Assessing Effects of Task and Data Distribution on the Effectiveness of Visual Encodings. Technical Report 3, 2018. doi: 10.

1111/cgf.13409

[52] R. Kosara and C. Ziemkiewicz. Do mechanical turks dream of square pie charts? In *Proceedings of the 3rd BELIV'10 Workshop: Beyond time and errors: Novel evaluation methods for information visualization*, pp. 63–70, 2010.

[53] R. S. Kramer, C. G. Telfer, and A. Towler. Visual comparison of two data sets: Do people use the means and the variability? 2017.

[54] Q. Lei and A. Reeves. When the weaker conquer: A contrast-dependent illusion of visual numerosity. *Journal of Vision*, 18(7):1–16, jul 2018. doi: 10.1167/18.7.8

[55] H. Mansoor and L. Harrison. Data Visualization Literacy and Visualization Biases: Cases for Merging Parallel Threads. *Cognitive Biases in Visualizations*, pp. 87–96, 2018. doi: 10.1007/978-3-319-95831-6_7

[56] A. Mayorga and M. Gleicher. Splatterplots: Overcoming overdraw in scatter plots. *IEEE transactions on visualization and computer graphics*, 19(9):1526–1538, 2013.

[57] J. A. Mazer. Spatial Attention, Feature-Based Attention, and Saccades: Three Sides of One Coin? *Biological Psychiatry*, 69(12):1147–1152, jun 2011. doi: 10.1016/j.biopsych.2011.03.014

[58] A. McNutt and G. Kindlmann. Linting for visualization: Towards a practical automated visualization guidance system. In *VisGuides: 2nd Workshop on the Creation, Curation, Critique and Conditioning of Principles and Guidelines in Visualization*, 2018.

[59] A. McNutt, G. Kindlmann, and M. Correll. Surfacing visualization mirages. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2020.

[60] L. Micallef, G. Palmas, A. Oulasvirta, and T. Weinkauf. Towards perceptual optimization of the visual design of scatterplots. *IEEE transactions on visualization and computer graphics*, 23(6):1588–1599, 2017.

[61] M. J. Morgan, S. Raphael, M. S. Tibber, and S. C. Dakin. A texture-processing model of the 'visual sense of number'. *Proceedings of the Royal Society B: Biological Sciences*, 281(1790):1–9, 2014. doi: 10.1098/rspb.2014.1137

[62] T. Munzner. Marks and Channels. In *Visualization Analysis and Design*, pp. 95–116. A K Peters/CRC Press, jul 2018. doi: 10.1201/b17511-5

[63] K. Myczek and D. J. Simons. Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception and Psychophysics*, 70(5):772–788, 2008. doi: 10.3758/PP.70.5.772

[64] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini. How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1469–1478, 2015.

[65] A. Parlapiano. Where income is higher, life spans are longer. *The New York Times*, Mar 2014.

[66] E. Picon, D. Dramkin, and D. Odic. Visual illusions help reveal the primitives of number perception. *Journal of Experimental Psychology: General*, 148(10):1675–1687, 2019. doi: 10.1037/xge0000553

[67] S. Pinker. Chapter 4: A Theory of Graph Comprehension. In *Artificial intelligence and the future of testing*. 1990.

[68] A. Pomè, G. Anobile, G. M. Cicchini, and D. C. Burr. Different reaction-times for subitizing, estimation, and texture. *Journal of Vision*, 19(6):1–9, jun 2019. doi: 10.1167/19.6.14

[69] M. Procopio, A. Mosca, C. E. Scheidegger, E. Wu, and R. Chang. Impact of Cognitive Biases on Progressive Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2626(c):1–18, 2021. doi: 10.1109/TVCG.2021.3051013

[70] P. S. Quinan, L. M. Padilla, S. H. Creem-Regehr, and M. Meyer. Examining implicit discretization in spectral schemes. *Computer Graphics Forum*, 38(3):363–374, 2019. doi: 10.1111/cgf.13695

[71] R. A. Rensink. On the prospects for a science of visualization. In *Handbook of human centric visualization*, pp. 147–175. Springer, 2014.

[72] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010. doi: 10.1111/j.1467-8659.2009.01694.x

[73] J. Ritchie, D. Wigdor, and F. Chevalier. A lie reveals the truth: Quasimodes for task-aligned data presentation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2019.

[74] L. M. Rodriguez-Cintron, C. E. Wright, C. Chubb, and G. Sperling. How can observers use perceived size? Centroid versus meansize judgments. *Journal of Vision*, 19(3), mar 2019. doi: 10.1167/19.3.3

[75] H. Rosling and Z. Zhang. Health advocacy with gapminder animated statistics. *Journal of epidemiology and global health*, 1(1):11–14, 2011.

[76] D. M. Russell. Simple is good: Observations of visualization use amongst the big data digerati. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 7–12, 2016.

[77] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, Data, and Designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):402–412, jan 2018. doi: 10.1109/TVCG.2017.2744184

[78] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A Taxonomy of Visual Cluster Separation Factors. *Computer Graphics Forum*, 31(3pt4):1335–1344, 2012. doi: 10.1111/j.1467-8659.2012.03125.x

[79] J. T. Serences, S. Saproo, M. Scolari, T. Ho, and L. T. Muftuler. Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *NeuroImage*, 44(1):223–231, 2009. doi: 10.1016/j.neuroimage.2008.07.043

[80] S. Smart and D. A. Szafir. Measuring the separability of shape, size, and color in scatterplots. In *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–14. Association for Computing Machinery, New York, New York, USA, may 2019. doi: 10.1145/3290605.3300899

[81] T. R. Stewart. Uncertainty, Judgment, and Error in Prediction. Technical report, 2000.

[82] P. Sun, C. Chubb, C. E. Wright, and G. Sperling. Human attention filters for single colors. *Proceedings of the National Academy of Sciences*, 113(43):E6712–E6720, 2016.

[83] P. Sun, C. Chubb, C. E. Wright, and G. Sperling. The centroid paradigm: Quantifying feature-based attention in terms of attention filters. *Attention, Perception, and Psychophysics*, 78(2):474–515, feb 2016. doi: 10.3758/s13414-015-0978-2

[84] P. Sun, C. Chubb, C. E. Wright, and G. Sperling. High-capacity preconscious processing in concurrent groupings of colored dots. *Proceedings of the National Academy of Sciences*, 115(52):E12153–E12162, 2018.

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

[86] D. A. Szafir. The good, the bad, and the biased: five ways visualizations can mislead (and how to fix them). *interactions*, 25(4):26–33, 2018.

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

[88] M. S. Tibber, J. A. Greenwood, and S. C. Dakin. Number and density discrimination rely on a common metric: Similar psychophysical effects of size, contrast, and divided attention. *Journal of Vision*, 12(6):8–8, jun 2012. doi: 10.1167/12.6.8

[89] R. Veras and C. Collins. Saliency deficit and motion outlier detection in animated scatterplots. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[90] A. C. Warden, J. K. Witt, M. Fu, and M. Dodd. Overestimation of Variability in Ensembles of Line Orientation, Size, and Hue. *Journal of Vision*, 20(11):1240, oct 2020. doi: 10.1167/jov.20.11.1240

[91] C. Ware. *Information visualization: perception for design*. Morgan Kaufmann, 2019.

[92] Y. Wei, H. Mei, Y. Zhao, S. Zhou, B. Lin, H. Jiang, and W. Chen. Evaluating Perceptual Bias During Geometric Scaling of Scatterplots. 2019. doi: 10.1109/TVCG.2019.2934208

[93] D. Whitney and A. Yamanashi Leib. Ensemble perception. *Annual review of psychology*, 69:105–129, 2018.

[94] J. K. Witt. Graph construction: Setting the range of the y-axis. *Meta-Psychology*, 3, 2019.

[95] J. K. Witt. The perceptual experience of variability in line orientation is greatly exaggerated. *Journal of Experimental Psychology: Human Perception and Performance*, 45(8):1083, 2019.

[96] C. Xiong, C. R. Ceja, C. J. Ludwig, and S. Franconeri. Biased Average Position Estimates in Line and Bar Graphs: Underestimation, Overestimation, and Perceptual Pull. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):301–310, jul 2020. doi: 10.1109/TVCG.2019.2934400