Average Estimates in Line Graphs Are Biased Toward Areas of Higher Variability

Dominik Moritz, Lace M. Padilla, Francis Nguyen, and Steven L. Franconeri

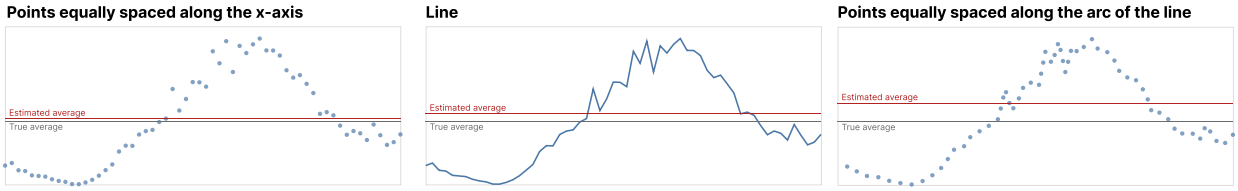

Fig. 1: Demonstration of the bias toward variability for three mark types showing the same data. The red line shows the mean estimated averages across all participants in our second experiment. The line chart (center) shows a bias of the estimated average toward higher variability in the higher y-values. The bias is smallest when the data is shown as points equally spaced along the x-axis (left). The bias in line charts is in the same direction as the bias of estimates of points sampled at equal intervals along the arc of the line (right).

Abstract— We investigate *variability overweighting*, a previously undocumented bias in line graphs, where estimates of average value are biased toward areas of higher variability in that line. We found this effect across two preregistered experiments with 140 and 420 participants. These experiments also show that the bias is reduced when using a dot encoding of the same series. We can model the bias with the average of the data series and the average of the points drawn along the line. This bias might arise because higher variability leads to stronger weighting in the average calculation, either due to the longer line segments (even though those segments contain the same number of data values) or line segments with higher variability being otherwise more visually salient. Understanding and predicting this bias is important for visualization design guidelines, recommendation systems, and tool builders, as the bias can adversely affect estimates of averages and trends.

Index Terms—bias, lines graph, ensemble perception, average

1 Introduction

Since William Playfair invented line graphs in 1786 [23], they have become one of the most common data visualization types. Designers use line graphs to visualize stocks, sensor data, machine learning metrics, and human vitals (e.g., heart rate). Line graphs show a continuous variable's change over another continuous variable, typically time, as the changing position of a line mark.

We generally assume that visualizations, especially of effective visual encoding channels such as position, are perceived not perfectly but without bias [5]. The popularity of line graphs may be because the visual encoding of time series as the position of a line is considered effective relative to other visual encoding channels, such as hue, depending on the task. However, designers should be cognizant of perceptual biases that can lead to misinterpretation of visualizations [14, 29]. For example, prior work demonstrates that the background color can bias the perception of the color of marks [28], and continuous rainbow color maps are perceived as discrete categories [17, 25].

There may be unexplored biases in line charts as well. When drawing a line, the length of the line drawn varies not only with the duration of the visualized time series but also with the variability of the values (and the resulting variability of the line graph). For example, take two time series of regularly sampled values over the same duration. The first value may be constant while the second value oscillates. Both time series have the same number of values (the same duration), but in the visualization as a line graph, the second line has a longer overall length—we call this the *arc length* of the line. The arc length is the sum of the length of all line segments. Steeper line segments are longer than other line segments of the same length along x. The arc length of a line affects how much visual weight a line has (how much "ink" is needed to draw it) and how much it draws viewers' attention [30]. Within a single line, periods of the same length may have a longer or shorter arc length depending on how much the line goes up and down, which depends on the amount of variability in the visualized time series.

Estimates of average values may be biased by design features of the marks that draw viewers' attention, as found in prior work [13], and increased variability in visualized times series may capture attention. Our bottom-up attention is generally attracted to visual information that contrasts with its surroundings [30]. Marks can vary in contrast to the background and other elements, which dictates how capturing they are to our attention, referred to as *salience*. For example, areas of a line graph with high variability also have more ink (often in color) and more edges, creating high contrast with the background. Therefore, we hypothesize that average estimates in lines are biased toward areas of line graphs that have a longer relative arc length (i.e., that have a longer arc length for the same duration or that use more ink). Put differently, we hypothesize that increased variability in higher values increases the average estimate of a time series (and vice versa) in line graphs and that the bias is consistent with the salience of the line.

We tested this hypothesis in two experiments. Our first experiment showed that average estimates are biased toward the area of the line that visualizes more variable data. In the second experiment, we sought to understand the reasons for the observed bias. We hypothesized that average estimates in line graphs are consistent with the salience of a line. We, therefore, hypothesized that average estimates of points drawn along the arc of a line are more consistent with average estimates

- *Dominik Moritz is with Carnegie Mellon University. E-mail: domoritz@cmu.edu.*
- *Lace M. Padilla is with Northeastern University. E-mail: l.padilla@northeastern.edu.*
- *Steven L. Franconeri is with Northwestern University. E-Mail: franconeri@northwestern.edu.*
- *Francis Nguyen is with UBC. E-mail: frnguyen@cs.ubc.ca.*

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

on lines than points drawn at regular intervals along the time dimension of a graph. In other words, the bias to variability may decrease from a line graph encoding to a point encoding of the same data. The results of the second experiment confirm this hypothesis, and a demonstration of the findings is shown in Figure 1. We preregistered the experiments on the Open Science Framework (Experiment 1 and Experiment 2) and have made our study materials available at (OSF Link).

Although we are not the first to document that different time series have different arc lengths, and that arc length affects aggregates [12, 21], we experimentally show the effect and reveal how much average estimates in line graphs can be biased by variability in the time series. Understanding this bias is important because people often use line graphs (as one of the most common visualization types) to visually assess whether values are, on average, above or below critical thresholds or to estimate future trends. Our results show that variability in line graph biases the perception of averages and, therefore, conclusions people draw. While the variability affects audiences' perception, it may be an artifact of an irrelevant factor that should not affect their conclusions. We discuss these implications and potential designs that could reduce the observed bias.

2 Related Work

Line graphs are a common visualization— especially of time series data—in various domains [24], appearing in papers, reports, monitoring dashboards, and visual analysis systems. They are generally considered an effective visualization for time series data [31]. Mackinlay describes a visualization as *effective* when the information it conveys is more readily perceived than with other visualizations [18]. A visualization is always effective only with respect to a particular *task*. In this paper, the task is to estimate the average of a time series, which corresponds to "compute derived value" in Amar et al.'s popular low-level task taxonomy [3]. Whether a visualization is effective depends on choosing the right visual mark and effective visual encoding channels. Position is considered the most precise visual encoding channel [5].

However, research also demonstrates the limitations of line graphs for various tasks (e.g., [1,2,5,6,9,11,15,33]; reviewed in [24]). Several studies compared line charts to other encodings and found that the efficacy of line charts depends on the task [2,6,11,15]. Albers, Correll, and Gleicher found that line graphs are best suited for identifying the min, max, and range while less effective for average estimation [2] (see also [6]). Studies have shown that positional encoding may be a precise visual encoding channel, but it can produce systematic biases regarding how averages are perceived [33] and remembered [19]. Researchers investigated bias in composed displays with line and bar graphs [33]. When comparing two curved lines, work shows that the steepness of the lines causes a perceptual illusion making it challenging to estimate differences between the lines visually [5].

We are not the first to recognize that line graphs dedicate more visual weight to steeper lines. This effect becomes critical when summarizing large ensembles of line graphs. Heinrich and Weiskopf reduced the salience of steeper lines in density visualizations of parallel coordinate plots [12]. Moritz and Fisher aggregated line graphs to create density visualizations of large time series [21]. To avoid visual artifacts of steeper lines, they normalized each line by the arc length such that each time series contributes equally to the density visualization. Zhao et al. proposed an effective density computation and extended density visualizations with interactivity [34]. However, none of these works experimentally confirm that average position estimates in line graphs are biased due to the increased salience.

3 Experiments

In two experiments, we investigated the perception of average values in line graphs. The goal of the first experiment was to determine if the perception of averages is biased toward variability and whether we can predict this bias. To examine one possible source of the bias, in Experiment 2, we aimed to identify the contribution of the line encoding. To test this, we conducted a study comparing the bias of three mark types: 1) points equally spaced along the x-axis, or *Cartesian spaced*, 2) points equally spaced along the arc of the line, and 3) a line.

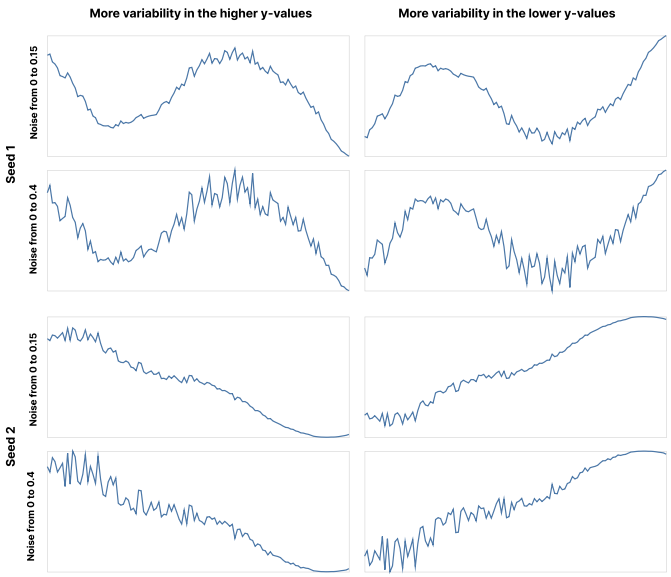

Fig. 2: First 8 of 48 stimuli for Experiment 1. The stimuli included two variability levels for each seed and conditions where the graphs were mirrored creating stimuli with variability in the higher and lower y-values.

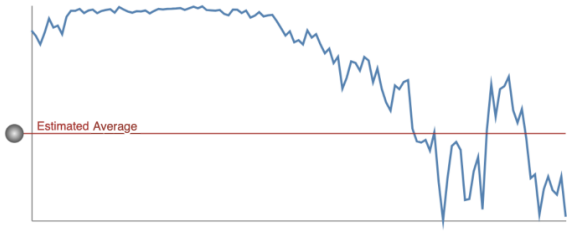

Fig. 3: Example stimulus. Participants can move the grabber up and down to where they estimate the line's average to be. Each series is shown as a 500px by 200px graph.

3.1 Experiment 1

The first experiment aimed to determine if there was a previously undocumented bias in perceptual line graph average estimation. We termed this potential bias *variability-overweighting*, which is when people believe that the average value of the data set is closer to the more variable data. For example, in Figure 1 center, if a participant were to indicate that the average value was located at the red line, they would be incorrect. In the figure, the true average is lower, but it could be the case that people are biased toward the more variable data.

We hypothesized that individuals would have a skewed perception of averages toward sets of values with higher variance. High variance increases the amount of ink in a line graph and, therefore, the visual saliency of the line. If, in a line graph, the variance correlates with the value plotted along the same axis (here y), then we expect people to estimate that most values are where the high-variance data is located.

To test if variability-overweighting occurs with line graphs, in Experiment 1, we showed participants line graphs of synthetic stock data that we modified to induce increased variability. We created stimuli that included more variability in the higher y-values and then reflected the stimuli to create graphs with more variability in the lower y-values (see Figure 2). We will refer to the stimuli with variability in the higher y-values as "variability upper" and those with variability in the lower y-values as "variability lower." Participants were tasked with estimating the average y-value of the stock data using a draggable line (Figure 3). We used a 2 (variability upper vs. lower) $\times$ 2 (more vs. less variability) within-subjects design for a total of 4 stimuli types of interest.

We generated the stimuli from 12 seeds to create 48 trials to ensure test-retest reliability. Creating images reflected vertically allowed us to test if variability-overweighting occurs similarly for higher or lower

areas on the y-axis. Participants were shown 48 images in a randomized order and estimated the average y-value for each image. We calculated each judgment's Euclidean distance estimation error by subtracting the actual average from the estimated average. The direction of the error was preserved, such that positive values indicated an overestimation of the average value, and negative values represented an underestimation.

3.2 Experiment 2

In Experiment 2, we aimed to determine if we could influence the degree of variability-overweighting by changing the mark type. To test this, we replicated Experiment 1, but we encoded the data using 1) Cartesian spaced points, 2) points equally spaced along the arc of the line, and 3) the same line encoding used in Experiment 1 (see Figure 4).

We hypothesize that by encoding the time series data as Cartesian spaced points, we can reduce the bias toward more variable data. Each data point is rendered as one point mark without a connection in this encoding. Therefore, two rendered points have the same salience regardless of their distance in y. In lines, the salience of the mark depends on the length of the arc. We also rendered points at equally spaced intervals along the arc (points along the arc) to simulate this behavior in our experiment. With points along arc, the average y-position of the points is heavily biased toward more variability since lines between neighboring points with more different values are longer. Therefore more points are along the arc of lines with more variability. In this encoding, a perfectly accurate viewer cannot estimate the true average of the underlying data series. We also included a line encoding to replicate Experiment 1.

As in Experiment 1, we showed participants graphs of synthetic time series. We then ask them to estimate the average using a draggable line. We used a 3 (point along x, point along arc, line) $\times$ 2 (variability upper vs. lower) $\times$ 2 (more variability vs. no variability) design.

We generated the stimuli types from 12 seeds to create 144 total trials. Participants viewed 48 images of one mark type from the 144 trials, and we calculated the error similarly to Experiment 1. We switched to a between-subject experiment to limit the number of graphs each participant saw and reduce the possibility of potential bias of viewing multiple graph types.

4 Stimuli Generation

We generated the stimuli for both experiments with the same process. This process used a simulation to generate realistically-looking line charts that we add linearly-interpolated noise to. We re-scaled the series to correct for a subtle yet important bias introduced by the noise. The code for our stimuli generation for Experiment 1 and Experiment 2 are available online and as supplemental material.

4.1 Experiment 1

The stimuli (with a sample shown in Figure 2) are line graphs of randomly generated series of numbers. Each series has 120 data points. A series is generated from a base series to which we add noise. We generated base series from a geometric Brownian motion stochastic process [32], a process used to generate realistic-looking stock data. We set $\mu = 0$ and $\sigma = 1$. We then applied a moving average over 30 points to smooth the base series. To get 120 data points in a series, we generate $120 + 30 = 150$ data points from geometric Brownian motion. We scale the base series, so all values are between zero and one. To the base series (consisting of data points base_i), we added uniform random noise (centered around 0). The amount of noise increases linearly with the value of the base series for positive y-alignment (linear interpolation between lowVariability and highVariability).

$$\begin{aligned} \text{noise}_i &= \text{lowVariability} \times (1 - \text{base}_i) + \text{highVariability} \times \text{base}_i \\ \text{dataPoint}_i &= \text{base}_i + (\text{rand}() - 0.5) \times \text{noise}_i \end{aligned} \tag{1}$$

Low variability series have less variability (0.15) than high variability series (0.4). We seeded the random number generator for the geometric Brownian motion stochastic process and noise to reproduce the same series. We curated the set of seeds to generate diverse line graphs with different shapes.

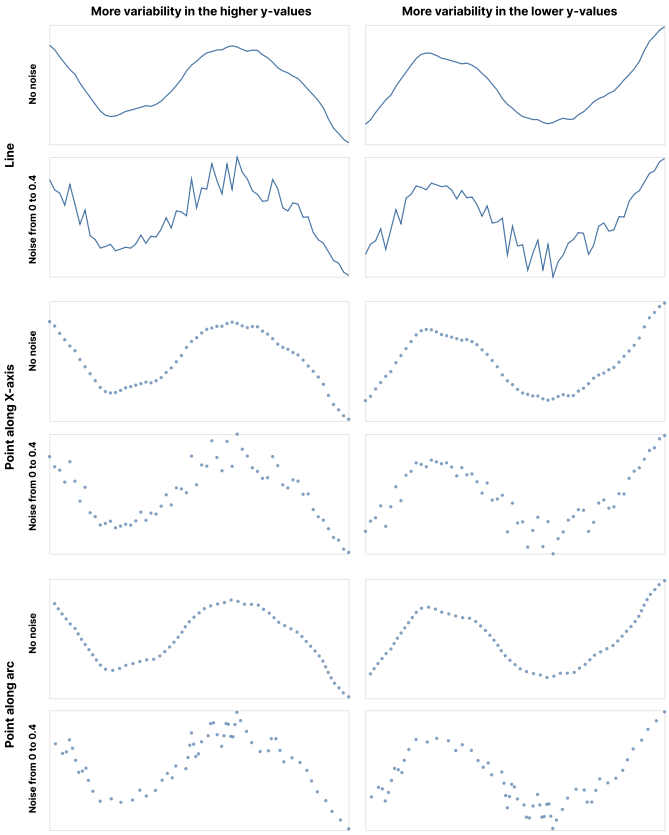

Fig. 4: First 8 of 144 stimuli for Experiment 2. For each seed, we included two variability levels, three mark types, and stimuli with variability in the higher and lower y-values.

We generated a series for negative y-alignment by mirroring the series vertically. We mirror each series to understand whether the average estimates may be biased toward higher and lower y-values and counter-balance this bias in our experiments to measure bias toward higher variability.

$$\text{mirroredDataPoint}_i = 1 - \text{dataPoint}_i \tag{2}$$

4.2 Scaling the averages

To allow for comparison across stimuli, we set the values between zero and one. We could naively scale the generated series to $[0, 1]$, but this would invalidate our experiment. To understand why, assume without loss of generality that the averages of the base series are around 0.5 and that the generation procedure adds noise to larger y-values (for not mirrored series). Therefore, the generated series are between 0 and ≥ 1 (with the exact amount depending on the noise). Therefore, if we rescaled the data to $[0, 1]$, we would push the average values of the series to lower y-values.

Let us assume our participants respond randomly or always estimate the average at 0.5. In both cases, we get the same result. Since the true averages are overall lower, we would find that estimates are biased to be higher than the average. We believe that every experiment that investigates bias should test its analysis with random responses. With random responses, any observed human bias should disappear.

To overcome the issue, we scale (multiply) the averages of the low-noise data to have the same averages as the high-noise data.

$$\text{scalingFactor} = \frac{\text{average}(\text{highNoiseData})}{\text{average}(\text{lowNoiseData})} \tag{3}$$

For mirrored series, we scale by $(1 - \text{average}(\text{highNoiseData}))/(1 - \text{average}(\text{lowNoiseData}))$. After this scaling, all stimuli with the same

seed (and mirroring) have the same average. If we simulate an experiment where participants always estimate 0.5, we find no bias toward higher-variable areas.

4.3 Experiment 2

We use the same stimuli generation procedure as in Experiment 1 and the same seeds. In addition to the line encoding, we created graphs with points sampled at equal distances along the x-axis and sampled along the arc of the lines. This design resulted in three stimuli types (lines, Cartesian spaced points, and arc spaced points). Like in Experiment 1, we used two—albeit different—levels of variability in the graphs (no additional variability, and 0.4). We chose no additional variability for the first level to understand the generated series' baseline bias. We used the same level of variability for the high variability series to replicate Experiment 1. The series had 60 data points and were scaled as in Experiment 1. Figure 4 shows example stimuli.

5 Design, Procedure, and Participants

5.1 Design

Experiment 1: We used a 2 (variability: .15 and .4) x 2 (variability upper vs. lower) within-subjects design. Average estimation error was collected as the dependent variable. This design resulted in a total of four trial types which were generated 12 times (48 total trials) to ensure test-retest reliability.

Experiment 2: We used a 2 (variability: 0 and .4) x 2 (variability: upper vs. lower) x 3 (mark type: line graph, Cartesian spaced points, arc spaced points) mix-design. The between-subjects measure was mark type, and the within-subjects conditions were variability 0/.4 and variability upper/lower. Variability upper/lower was used as a manipulation check. Mean estimation error was collected as the dependent variable. Each participant completed the task with graphs that included variability 0/.4 and variability upper/lower in a randomized order. The total number of trials was the same as in Experiment 1.

5.2 Procedure

In both experiments, participants completed this study online on their personal machines. After giving Institutional Review Board (IRB) approved consent to participate, individuals were given three types of instructions. The first set of instructions prompted participants to set their browser window to 100% zoom. The second set of instructions pertained to the task, which was:

> "***Experiment Instructions***. *Please read the following paragraphs carefully. You will be asked questions about the information in the paragraphs.*
> ***Scenario****: Assume that you are a stock market investor. You are investing your own money in stocks, and you want to determine the average price of a stock over time in order to pick the best investment.*
> ***Task****: In this experiment, you will be shown graphs of stock prices over a one-year period like the one below. Your task is to determine the average stock price for that year.*
> *What is the average stock price? (Click and drag the line to indicate the average stock price)*
> ***Response****: To indicate the average stock price, use your mouse to drag the line on the chart. Move the line to where you think the average stock price is for that year. You can readjust the line by clicking and dragging. Once you are happy with your judgment of the average stock price, click the next button.*"

The final set of instructions was an attention check, where participants were asked to fill in a blank with the word "stock". The sentence was, "*During this study, you will be asked to look at graphs of _____ prices.*" Following the instructions, participants completed 48 estimation judgments in a randomized order. They indicated their judgments using a horizontal slider that was superimposed on the stimuli (shown in Figure 3) to estimate the average data value in the graphs. The trials included text reminding the participants about the task. If participants failed to move the slider, they would be prompted to do so and restricted from progressing until they made their judgment. They received no feedback as to the accuracy of their judgments.

Following the main experiment, participants answered open ended questions about their strategy and what they thought the experiment was about. They also reported their gender and age.

5.3 Participants

Based on the effect size calculated from pilot data, a power analysis was conducted using G*Power, to determine an adequate sample size, and preregistered. At an alpha of 0.05, power of 0.95, 4 predictors, and an effect size of adjusted r-square of 0.13, the minimum number of participants needed is 132, which we rounded to 140. For Experiment 1, participants were 142 people from Amazon's Mechanical Turk, with participation criteria set to workers in the US who were 18 years of age or older. Participants demographics were 98 male and 44 female, with an average age of 39 (SD = 9).

For Experiment 2, participants were 420 (140 per between-subjects group) people from Amazon's Mechanical Turk. Of those who chose to answer, 46% identified as female, with an average age for the whole sample of 41 (SD = 11). IRB approval for this research was obtained from (removed for anonymization) University's IRB. Participants were paid in accordance with (removed for anonymization) minimum wage.

6 Results

To answer our primary analysis question, whether the perception of averages in lines is biased toward variability and whether we can manipulate and predict this bias, we will detail the results of the two experiments. In each experiment, we will begin with descriptive statistics about estimation error. We then show the results of statistical tests of our preregistered hypotheses. For all of the analyses, we did not remove any participants.

Following the preregistered analysis, we detail the thematic analysis of participants' strategies, including examining the variability-overweighting exhibited by participants who reported using the correct strategy. We also conducted a sensitivity analysis to determine if individuals who guessed the purpose of the study biased the results. We conclude the analysis with model comparisons that use the average along the arc to predict the observed biases.

6.0.1 Accuracy Calculation

We computed the error for each participant's estimates as the difference between the estimated average and the true average.

$$\text{Error} = \text{Estimated Average} - \text{Average} \tag{4}$$

We also calculated whether a participant overestimates the average. We specify that they overestimated when the estimated average is higher than the true average of the time series data.

$$\text{Overestimated} = \begin{cases} \text{Overestimated}, & \text{if Error} > 0 \\ \text{Underestimated}, & \text{otherwise} \end{cases} \tag{5}$$

To always have the high variability data at the higher y values, we also compute a normalized average as:

$$\text{Normalized Average} = \begin{cases} -\text{Average} + 1, & \text{if variability upper vs. lower} \\ \text{Average}, & \text{otherwise} \end{cases} \tag{6}$$

We similarly compute a normalized error where a positive error indicates an estimated average toward higher variability.

6.1 Experiment 1

Descriptive statistics. As a preliminary analysis of estimation error, we counted how many times participants over or underestimated the average. Of the 6816 responses, 3239 (48%) were overestimated, and 3577 (52%) were underestimated (see Figure 5). Participants generally underestimated averages.

Since our experimental data contains graphs with variability in the upper and lower y-values, we broke down these counts by this condition to determine whether the estimates are toward or away from the higher

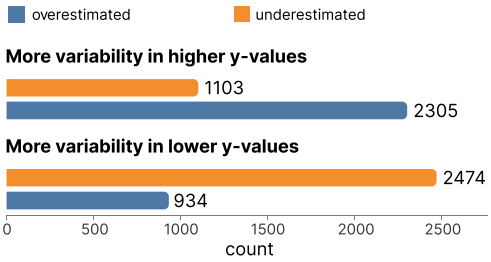

Fig. 5: Counts of the over- and under-estimations in Experiment 1, broken down by the variability upper vs. lower condition.

variability. As shown in Figure 5, for graphs where the variability was in the higher y values, participants overestimated 2305 (68%) trials, and they underestimated 1103 (32%). In the condition where the variability was in the lower y values, 934 (27%) were overestimated, and 2474 (73%) were underestimated. In both conditions, the estimation error was consistent with the variability.

Statistical tests of preregistered hypotheses. To determine if the findings from the descriptive statistics are robust, we conducted a statistical analysis of our preregistered hypotheses. We preregistered three hypotheses on the Open Science Framework for the first experiment[1].

H1 *"Estimation error will be significantly different than zero."*

H2 *"There will be significantly more estimation error for trials with higher variability compared to lower variability."*

H3 *"Estimation error will be observed in the direction of the increased variability (i.e., positive errors will be observed when the area of highest variability is above the average y-value and negative errors will occur when the highest variability is lower on the y-axis than the average.)"*

To test these hypotheses, we conducted the preregistered analysis, in which a linear regression model was fit to the data using the R function lmer [26] with restricted maximum likelihood estimation procedures [27]. Note that we used multi-level linear regression models to account for correlations between participants' responses instead of the more simplistic pre-registered linear regression models. Linear regression assumptions were tested and met. The model included *variability size* (.15 vs. .4), *variability position* (upper vs. lower), their interaction, and random intercepts for each participant to predict errors in participants' average estimations. The referents were .15, and variability in the upper y-values. The resultant model in R notation was: $Error \sim variabilitySize * variabilityPosition + (1|Id)$.

Test of H1: *estimation error will be significantly different than zero*. The results revealed a significant intercept of the model ($b = -0.036$, $t(6,811) = -6.6$, $p < .001$, 95% CI $[-.047, -.025]$), providing evidence that the absolute estimation error (3.6%) for the referent conditions was meaningfully different than zero (supporting H1). This effect can be seen in Figure 6 (left panel), which displays estimation errors for each condition, with none of the conditions overlapping zero.

Test of H2: *significantly more estimation error for trials with higher variability.* The results also revealed a significant main effect of variability ($b = -.022$, $t(6,811) = -6.5$, $p < .001$, 95% CI $[-.029, -.016]$). This effect can be seen in Figure 6, where there is a meaningful separation between the two variability types for the variability upper vs. lower conditions (denoted with H2). This finding supports H2, suggesting significantly more estimation error for trials with higher variability than lower variability.

Test of H3: *estimation error will occur in the direction of the increased variability.* There was also a significant interaction between *variability .15 vs .4* and *variability upper vs. lower* ($b = .034$, $t(6,811) = 7$, $p < .001$, 95% CI $[.025, .044]$). To unpack the interaction, we ran the same model as above but with the variability in the lower y-value graphs as the referent. This model yielded a significant effect of variability but in the opposite direction ($b =$ **.012**, $t(6,811) = 3.39$, $p = .001$, 95% CI $[.005, .018]$) compared to the prior model ($b =$ **-.022**). As seen in Figure 6, errors occurred in the direction of the increased variability, supporting H3. We found positive errors when the area of highest variability was above the average y-value and negative errors when the highest variability was lower on the y-axis than the average.

[1] In the original pre-registration, we used the term *noise* rather than *variability*. We updated the term here to be consistent.

6.2 Experiment 2

To examine one possible source of the variability-overweighting, in Experiment 2, our goal was to identify the contribution of the line encoding. We predicted that there would be an interaction between variability and the mark type, such that the effect of variability will be smaller for graphs with points spaced along the x-axis than graphs with points spaced along the arc and line graphs.

Descriptive statistics. Using the same methods as in Experiment 1, we counted how many times participants were biased toward variability (see Figure 7). Of the 6720 responses, 3724 (55%) for Point, 4226 (63%) for Line, and 4738 (71%) for Point Arc were biased toward variability.

Statistical tests of preregistered hypotheses. To test the reliability of the descriptive statistics, we preregistered two hypotheses on the Open Science Framework for the second experiment.

H4 *"There will be significantly more estimation error for trials with higher variability than no additional variability."*

H5 *"The least variability-overweighting will occur in graphs with points that are equally spaced along the x-axis."*

We used a multilevel model to fit the data using the *lmer* package [4] in R, which is appropriate for mixed designs with between- and within-subjects variables. The model used variability (0 and .4) to predict normalized estimation error (testing H4). We calculated the normalized estimation error for each condition using the absolute error (the error is computed in the same way as in Experiment 1) when the graph was vertically mirrored. We used *normalized error* rather than *error* and removed the *variability upper vs. lower* term to reduce the complexity of the model, which was preregistered. To evaluate H5, we also included an interaction term between mark type and variability and the necessary lower-order terms. Finally, we included random intercepts for each participant. The resultant model in R notation was: $NormalizedEstimationError \sim markType * variability + (1|Id)$. The referents of the model were the line mark and zero variability.

Test of H4: *significantly more estimation error for trials with higher variability.* Replicating Experiment 1, the model results revealed a main effect of variability ($b = 0.03$, $t(20,153) = 10.03$, $p < .001$, 95% CI $[.024, .036]$), indicating that graphs with more variability had greater estimation error. Figure 8 shows that, when collapsed across the mark types, estimation error increased by .03 from graphs with no additional variability to .4 variability (confirming H4). The meaningful increase in error with the more variable graphs can also be seen in Figure 6 (right panel), which shows the impact of variability on each mark type.

Test of H5: *least variability-overweighting in graphs with equally spaced points.* As shown in Figure 6 (right panel), Point (Cartesian spaced) had the smallest change in normalized error from charts with low to high variability (0 to .4). Our results revealed the change from low to high variability was meaningfully larger for Line vs. Point ($b = -.023$, $t(20,153) = -5.45$, $p < .001$, 95% CI $[-.031, -.015]$). The change in normalized error from low to high variability was also meaningfully larger for Point Arc vs. Point ($b = .05$, $t(20,153) = 10.74$, $p < .001$, 95% CI $[-.031, -.015]$). Point Arc showed the largest increase in the normalized error of 5%, followed by Line (3%), and then Point (.69%).

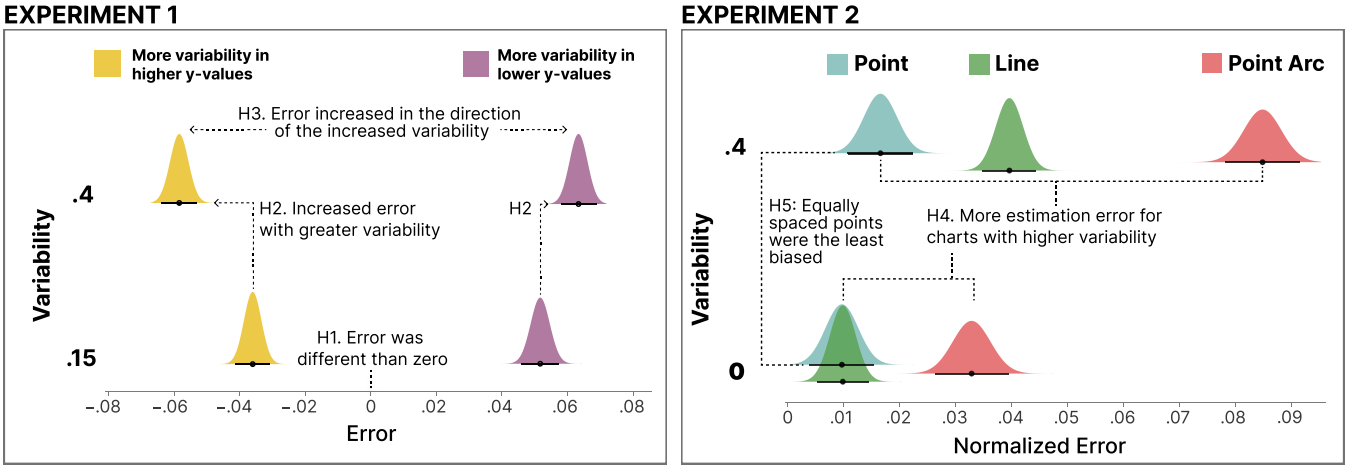

Fig. 6: Experiments 1 and 2 results, showing the impact of variability, variability upper vs. lower, and mark type on estimation error. The left panel details the findings of Experiment 1 with annotations describing confirmed hypotheses 1-3. The right panel shows Experiment 2 with annotations describing confirmed hypotheses 4-5. The black bars within the density plots show 95% CIs with a mean dot.

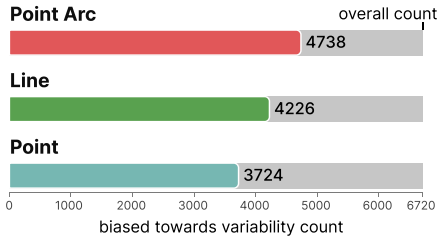

Fig. 7: Counts of the number of estimates that were biased toward variability in Experiment 2 for each mark type.

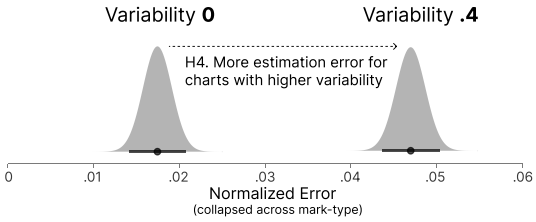

Fig. 8: The meaningful main effect of variability in Experiment 2, averaged over the mark types, including annotations describing confirmed H4. The black bars within the density plots show 95% CIs with a mean dot.

We conducted a follow-up regression analysis to determine if the small increase in normalized estimation error was meaningful for Point. This analysis revealed a meaningful but small bias for Point ($b = .007$, $t(6,718) = 2.24$, $p = .025$). In sum, these results support H5, indicating that the points equally spaced along the x-axis had the least bias (.69%), with the line (3%) and points along the arc showing greater bias (5%).

6.3 Open Responses for Experiments 1 and 2

After completing the estimation judgments, participants answered an open-ended question about their strategies in the task. The question was, "*We are very interested in how you made your decisions about the average stock price. Please list all the things you considered when making your judgments.*" Two raters read the responses and coded them based on the six most common strategies to analyze these data. The following sections report the six most frequent strategies and include example responses. In this analysis, we identify that a small proportion of the participants reported using the correct strategy. We conducted a follow-up analysis to determine if those who were consciously aware of the correct strategy displayed less variability overweighting.

Participants also reported their beliefs concerning the purpose of the experiment to determine if any participants intentionally biased their judgments. The question text included,"*What do you think the experiment was about?*" In the second part of this section, we report the proportion of participants who were aware of the purpose of the study. Then we conducted a sensitivity analysis to determine if the people who guessed the purpose of the study meaningfully impacted the findings.

6.3.1 Reported strategies

We identified six main strategies that participants used to estimate the average of the stock data. Using the strategy codes, we computed inter-rater reliability scores (*IRR*, Cohen's Kappa) [8] for the codes to determine the level of agreement between the raters (shown in the bottom row of Table 1). The average inter-rater reliability for the six questions was .83 and ranged from .70 to .89. This range of inter-rater reliability scores indicates a substantial level of agreement between the two raters [20]. The codes were not mutually exclusive, and many times participants indicated that they used several strategies, in which case they received multiple codes. The proportion of strategies reported in Table 1 is for the codes the two raters agreed on.

Mental averaging. The most commonly reported strategy was mentally computing the average using visual perception. For example, a participant wrote, "*I marked the point on the graphs where it seemed like the generalized average would fall if the points on the graph were boiled down into numbers and you wanted to find the average of those numbers.*" Another participant described, "*I tried to get a visual sense of where the average would fall. I looked for a good mid point of the overall graph.*" Table 1 shows that this was the most commonly reported strategy for each mark type.

Focusing on extrema. The second most common strategy was to focus on the max and min points and select a location between those extrema. For example, a participant wrote, "*I looked at the highest and lowest point and went with the middle.*" Another example includes, "*I looked at the lowest mark and the highest mark and then the middle of that, but looked to see if there were upper or lower trends and adjusted accordingly to that...*" Focusing on the extrema is not the most effective strategy. It is surprising to see that, on average, roughly 17% of the participants indicated that they incorporated the high and low points into their average estimations.

Incorporating variability. Roughly 15% of participants reported incorporating variability into their judgments. However, they incorporated the variability in different ways. For example, one type of strategy included incorporating the areas with both high and low variability. For example, a participant wrote, "*I tried to find out the relatively stable parts of the graph, these were useful when they extended over a long*

Table 1: The six most reported strategies for each experiment and the proportion of participants who correctly guessed the purpose of the study. The last column shows the inter-rater reliability score (IRR), which indicates the level of agreement between raters. IRR scores over .61 indicate substantial agreement between raters [20].

	Reported Strategy						Guessed purpose of study
Exp	mental **averaging**	focusing on **extrema**	incorporated **variability**	equal number of **points** or **line** below and above	equal **area** below and above	**beginning** and **end** points	variability-overweighting
Exp 1 Line	56.74%	21.28%	21.99%	5.67%	3.55%	4.96%	2.13%
Exp 2 Line	52.42%	23.39%	17.74%	8.87%	5.65%	2.42%	4.03%
Exp 2 Points	38.13%	9.35%	10.07%	24.46%	1.44%	2.16%	.72%
Exp 2 Point Arc	32.85%	18.25%	12.41%	7.30%	5.11%	1.46%	1.46%
Average	**44.92%**	**17.93%**	**15.53%**	**11.65%**	**3.88%**	**2.77%**	**2.03%**
IRR	.83	.88	.70	.87	.89	.83	

period of time. I also considered the effect of the crests and troughs and the depth of these extreme occurrences. Using these as a metric, I tried to estimate the average."

In contrast, another group of participants focused more on areas with low variability. For example, "*Most of the time the stock price comes to certain point, and jumps again or fall back, I consider the price where it is often stable for more time.*" or "*It was easier to make the average when the stock prices were not changing much and the graph was more even. When the prices were more "jumping", I tried to find the phase where these trends stayed the longest and put my average around it.*" Participants who reported this type of strategy seemed averse to variability or uncertainty, which is a well-known bias in psychology [7, 16].

Equal number of points or line below and above. A strategy we did not anticipate was ensuring an equal number of points or line lengths above and below the judgment indicator. To indicate their responses, participants drug a line on the graph. By allowing participants to place a line directly on the graphs, participants could then easily count the number of points above and below the line. One participant simply wrote, "*I tried to have the number of points above and below the line be approximately equal.*" Unsurprisingly, this strategy was most common for participants who viewed graphs with points equally spaced along the x-axis (24%). Although less common, some participants who viewed the line encoding also used this strategy (5-8%). A participant explains, "*I tried to get half of the trend line above and half of the trend line below the average line and where I placed it.*"

Beginning and endpoints. Another suboptimal strategy was to focus on the beginning and ending values of the time series. While a small proportion of participants used this strategy (roughly 3%), it is noteworthy because it reflects a misconception. For example, one participant wrote, "*I mainly looked at the stock at the beginning and end of the year. Afterwards, I tried to make an educated guess on what the average stock price would be.*" Another person describes also being confused about the impact of data at the end of the time series. They wrote, "*Depending on how the end of the chart looks, I draw a different strategy. If the chart is rallying, I believe the average price is at the low before this rally. If the chart is going down, I place it at the lowest low there was throughout the chart.*"

Equal area (correct strategy). The correct strategy was to select a location with equal *area* above and below the estimated average. Only a small number of people reported using this strategy (roughly 4%). An example is, "*I just tried to make the volume of the areas above and below the line approximately equal. That was my only strategy. Think I learnt it in a maths or stats course.*"

As the equal-area strategy is the correct approach, we wanted to determine if participants who used it showed less bias in their judgments. To compare performance between those who used the equal-area strategy to those who did not, we computed the bias for the two groups for Experiment 2. Figure 9 shows the nine participants with the correct strategy in the Point Arc and Line groups and the two in the Point group compared to a distribution representing all the other strategies. Note that there is one distribution for all people with the incorrect strategies compared to individual distributions for those with the correct ones. We did this to clarify that a small number of people had the correct strategy and meaningful variation exists between them.

Taking the individual distributions from those with the correct strategy as a whole compared to the distributions for those with the incorrect strategy, we found that people with the correct strategy showed 12.7% less bias (.028 normalized error) than those with the incorrect strategy (.032 normalized error). The disparity between those with the correct strategy (.021 normalized error) and without (.041 normalized error) was most pronounced for the Line encoding with .4 variability (change of .019 or 46% reduction). We opted not to do a statistical analysis on these groups as they were highly unbalanced (20 participants vs. 398) and were not equally distributed across the groups. However, visual analysis reveals a general tendency where using the correct strategy leads to less bias.

6.3.2 Knowledge about the experiment purpose

Several people in each experimental condition made guesses somewhat close to the actual experiment goals in response to the question, "What do you think the experiment was about?" For example, one person wrote, "*How people picture averages differently when there are smooth transitions versus spikes in the graph.*" and another person wrote "*I think it was about how accurate people can estimate the average of a line and if different line conditions affect the accuracy, such as jagged line vs. smooth line...*"

In Experiment 1, three people guessed the purpose of the study, and ten correctly guessed in Experiment 2. We conducted a sensitivity analysis to determine if those participants biased the findings. In this analysis, we removed the participants that relatively accurately guessed the manipulations of the study and reran the preregistered analysis for Experiment 2. Across all the findings, there was no meaningful impact of removing the participants who guessed correctly. To illustrate these effects, in Figure 10, we show the original data from Experiment 2 with density plots. Overlayed on the density plots are quantile dot plots that show the data after removing the participants who guess the manipulations in the study. As seen in Figure 10, where all the distributions for each condition overlap, removing the participants did not meaningfully impact the results.

6.4 Predictive Model

Chart authors should consider different designs if a particular line chart is prone to bias. To help chart authors know whether a chart may be

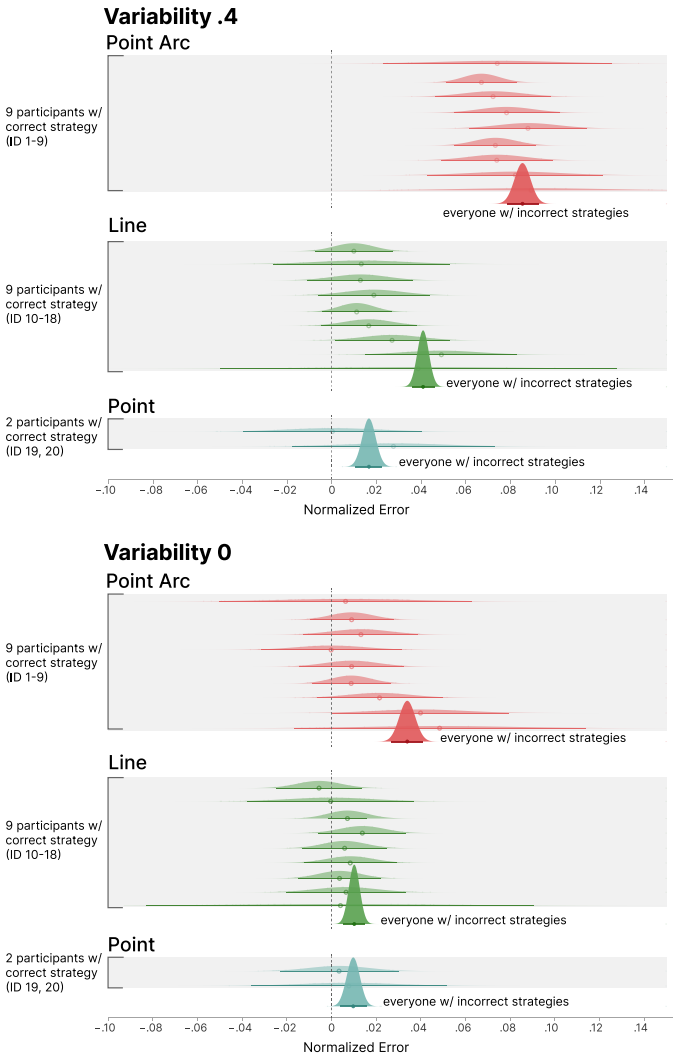

Fig. 9: Strategy analysis for Experiment 2, where individual participants with the correct strategy (gray background) are shown compared to the other participants (foreground density plots). These data are broken down by mark type and variability. The horizontal bars within the density plots show 95% CIs with a mean dot. The dashed line denotes zero normalized error.

biased without running their own perceptual experiments, we sought to build a model that predicts participants' responses. We aim to predict the bias and average estimate only based on properties of the data we can observe in a given line chart rather than based on the parameters of the data generation method since the latter is typically not known. We also chose to only use a simple model with few features (rather than, e.g., the whole time series as input) since we are interested in the model's generalizability.

We hypothesize that such a model is possible. If the salience of longer line segments drives the estimates of averages, we may be able to predict the estimates of averages using the true average and the average of the values along the arc—*arc average* for short.

To understand whether the arc average is a meaningful predictor, we computed the Pearson correlation between the average error of the average estimate for each stimulus to the error of the arc average estimate. For Experiment 1, the correlation is 0.85 ($p < .001$), and for Experiment 2, the correlation is 0.64 ($p < .001$). This suggests that the arc average is meaningful to predict the variability overweighting bias.

To predict the estimated average we created linear regression models that first used the average of the data points to predict participants' responses and then a second model that included the arc average. The

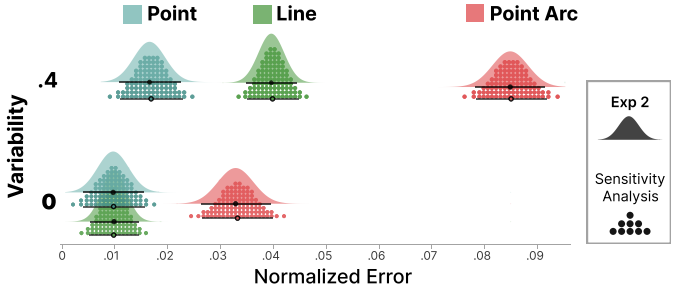

Fig. 10: Sensitivity analysis for Experiment 2, in which the original data from Experiment 2 is displayed with density plots, and the data that excludes people who guessed the purpose of the experiment is shown with quantile dotplots. The black bars within the density plots show 95% CIs with a mean dot.

goal of the second model was to evaluate if the arc average accounted for meaningfully more variability in participants' responses than the average of the data set alone. We then statistically compared the two models to determine if the arc average model was a significantly better fit, using the data from Experiments 1 and 2.

For Experiment 1, we fit a linear regression model using the average of the data point to predict participants' estimates. The average of the data points meaningfully predicted participants' responses ($b = .53$, $t(6814) = 51.80$, $p < .001$) with a model adjusted r-squared of .28. For the second model that included arc average, both the average of the data points ($b = .45$, $t(6813) = 39.42$, $p < .001$) and arc average ($b = .25$, $t(6813) = 15.70$, $p < .001$) meaningfully accounted for variance in participants judgments. The second model had an adjusted r-squared of .31. This result suggests that after accounting for the meaningful impact of the average of the data points, for every one unit change in arc average, participants' judgments were biased by .25. We then compared the two models using an ANOVA. This comparison revealed that the second model, which included arc average, had a significantly better fit than the first model ($F(2, 6813) = 246.58$, $p < .001$).

We also completed the same sequence of model comparisons for the data in Experiment 2 using only the data for the line stimuli. For the first model, the impact of the true average was $b = .69$, $t(6718) = 76.07$, $p < .001$, with an adjusted r-squared of .46. For the second model, the effect of the true average ($b = .39$, $t(6717) = 11.62$, $p < .001$) was larger than the impact of the arc average ($b = .34$, $t(6717) = 9.18$, $p < .001$), with an adjusted r-squared of .47. This result suggests that after accounting for the meaningful impact of the true average, for every one unit change in arc average, participants' judgments were biased by .34 (compared to .25 from Experiment 1). When comparing the two models, we found that the model that included the arc average had a meaningfully better fit than the one that did not ($F(2, 6717) = 84.30$, $p < .001$).

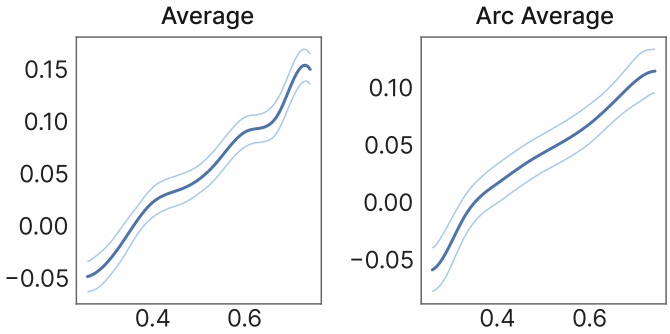

Fig. 11: The two response functions of a generalized additive model (GAM) trained on the line data from Experiments 1 and 2. The response functions resemble linear functions, making linear models appropriate for these data.

We then created a linear model from the data for both experiments with three parameters: intercept ($b = .14$, $t(13533) = 32.58$ $p < .001$), average ($b = .40$, $t(13533) = 34.36$, $p < .001$), and arc average ($b = .31$, $t(13533) = 21.87$ $p < .001$), and an adjusted r-squared of .39. We selected a linear model because we found that a more sophisticated GAM [10] used nearly linear feature functions (Figure 11).

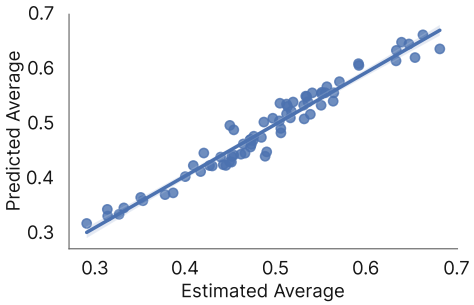

Fig. 12: Predicted and estimated averages for all stimuli.

For most stimuli (90%), the model predicted the correct direction of the bias. The predicted estimated average fits well with the estimated averages (Figure 12). The model's mean absolute error is .014, which means that with values in the range of [0,1], the prediction is only off by 1.4%. The RMSE is .019, also indicating a good fit.

7 Discussion

The results of Experiment 1 (Figure 6, left) support our hypothesis that estimation error is significantly biased toward the direction of larger variability in data. This bias could be caused by more variability leading to steeper line segments that use more ink and are more visually salient. Prior work has also found that areas of higher salience in scatter plots can bias average estimation [13]. To test this theory, we conducted Experiment 2, using a dot plot instead of a line chart to encode the series data. In the dot plot, the amount of salience is proportional to the amount of data in each x-interval and independent of the steepness of the line segment.

Experiment 2 (Figure 6, right) replicates the findings from Experiment 1. The experiment additionally supports our hypothesis that we can reduce the bias by encoding the series data as a dot plot instead of a line chart. To simulate the higher salience of steep line segments, we also tested a design that spaces points along the arc of a line. We found that the line bias was significantly higher than the bias of the dot plot but lower than the average estimation of the points along the arc. These results support our theory that the bias is toward more visually salient areas of the chart but cannot yet explain the full extent of the bias.

We generated the stimuli for our experiments using different levels of variability. Since these parameters are typically unknown, it would be impossible to model the bias in real-world applications. However, if we assume that the bias is caused by the salience of steep line segments, we can compute the direction of the bias directly from the average of the points along the arc of the line. We can also estimate the magnitude of the bias as a function of the average of the series and the average of the points along the arc using a simple regression model (Section 6.4). Future work could refine this model using more features.

Our experiments show that average estimates are biased toward higher variability. We believe that we could similarly bias trend estimation in line graphs. For example, a line graph could have more variability for smaller values in the first half and higher variability for larger values in the second half. Since we found that the estimates of averages for the first half are lower than the true average and that estimates for the second half are higher, we can expect that a person also perceives a more extreme increase (stronger trend) than there is. Our experiment only investigated average estimation in isolated charts. As such, future work must confirm that this bias exists in combined charts. Suppose trend estimation could be biased by variability. In that case, malicious people who can affect the variability of series could influence decisions other people make based on trends in data, such as in stock trading.

The results of our experiments have implications for the design of charts in applications where people estimate the average or trend of a series. Designers should consider whether they can replace a line chart with a dot plot to reduce the bias. However, the dot plot design makes the data order and the delta between consecutive points less clear. There may also be other ways to reduce the bias, such as using a different type of line chart that de-emphasizes steep line segments using thinner line segments or lines with lower opacity.

We asked participants about their strategies for estimating the average and found that people used a variety of strategies, the majority of which were incorrect. The high proportion of misconceptions observed in participants' strategies is concerning. We also found that those who used the correct strategy of aiming for an equal area between the average line and the data line seemed to be less biased (Section 6.3.1). While we have too few people to draw firm conclusions, this insight suggests that people may be able to learn to reduce the bias by using the correct strategy. When we initially developed this work, we hypothesized that the biases would be driven by visual salience, a bottom-up attentional process. While such unconscious processes are certainly part of the cause, these data provide some indication that strategies may play a role. One limitation of this work is that we cannot disambiguate the effects of visual salience and strategies. The interconnection between the two is consistent with theories in visual attention that suggest strategies and bottom-up processes are intrinsically interconnected, forming a feedback loop [22, 30]. It is also possible that both the mark type and response method bias participant strategies, which could have impacted attention and responses. Despite these limitations, our findings point to the possibility that visual literacy training might benefit from teaching people to use the correct strategy.

We carefully designed the stimuli of the experiment such that simulated random responses did not show the bias we expected in real responses (Section 4.2). We only found this subtle issue after some initial data generation and pilots and would therefore encourage everyone who runs experiments that test human biases to test their experiments with random data.

8 Conclusion and Future Work

This paper shows that average estimates are biased toward areas of higher salience, caused by increased variability in the visualized data. Since this bias can affect the conclusions drawn from data, visualization designers who create line graphs must be aware of it. This bias is not only significant but also practically relevant. The amount of variability in a line graph may be due to irrelevant (to the conclusions) factors, such as inconsistencies in the data collection, such as varying sensor noise. In the worst case, a malicious actor could introduce small amounts of noise to mask larger changes or nudge analysts to see larger changes.

By quantifying the bias, we can consider showing viewers warnings when we expect the bias to affect the conclusions drawn from a graph or consider alternative visual encodings that do not have the bias shown in this paper. For example, we showed how points instead of lines reduce the bias. Another idea could be to reduce the salience of steep lines by varying the opacity or line width. Alternatively, designers could consider annotating graphs with averages or other visual encodings when average estimates are needed. However, designers need to consider potential biases that additional visual encodings could introduce.

We discussed that biased average estimates could also lead to biased trend estimates. Future work should investigate how manipulations of time series data visualized as line graphs affect trend estimates. Participants in our study represent a general population of people with some but not expert-level visualization literacy. If trend estimates can be affected, we should also investigate whether experts such as scientists, doctors who look at vitals, and stock traders are as affected as the general population. An avenue for investigating this effect could be to analyze historical data such as stocks and see whether increased variability affected traders' investments.

Acknowledgments

This work was supported in part by grants from the NSF (#2238175 and #1901485).

References

[1] M. Adnan, M. Just, and L. Baillie. Investigating time series visualisations to improve the user experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5444–5455, 2016. doi: 10.1145/2858036.2858300 2

[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. doi: 10.1145/2556288.2557200 2

[3] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. pp. 111–117. doi: 10.1109/INFVIS.2005.1532136 2

[4] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01 5

[5] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. 79(387):531–554. doi: 10.2307/2288400 1, 2

[6] 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*, 2012. doi: 10.1145/2207676.2208556 2

[7] C. R. Fox and A. Tversky. Ambiguity aversion and comparative ignorance. *The quarterly journal of economics*, 110(3):585–603, 1995. doi: 10.2307/2946693 7

[8] M. Gamer, J. Lemon, I. Fellows, and P. Singh. *irr: Various Coefficients of Interrater Reliability and Agreement*, 2019. R package version 0.84.1. 6

[9] A. Gogolou, T. Tsandilas, T. Palpanas, and A. Bezerianos. Comparing similarity perception in time series visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):523–533, 2018. doi: 10.1109/TVCG.2018.2865077 2

[10] T. J. Hastie. Generalized additive models. In *Statistical models in S*, pp. 249–307. Routledge, 2017. 9

[11] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1303–1312, 2009. doi: 10.1145/1518701.1518897 2

[12] J. Heinrich and D. Weiskopf. Continuous parallel coordinates. 15(6):1531–1538. doi: 10.1109/TVCG.2009.131 2

[13] M.-H. Hong, J. K. Witt, and D. A. Szafir. The weighted average illusion: Biases in perceived mean position in scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):987–997, 2021. doi: 10.1109/TVCG.2021.3114783 1, 9

[14] D. Huff. *How to lie with statistics*. Penguin UK, 2023. 1

[15] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *IEEE transactions on visualization and computer graphics*, 16(6):927–934, 2010. doi: 10.1109/TVCG.2010.162 2

[16] M. S. Kimball. Standard risk aversion. *Econometrica: Journal of the Econometric Society*, pp. 589–611, 1993. doi: 10.2307/2951719 7

[17] Y. Liu and J. Heer. Somewhere over the rainbow: An empirical assessment of quantitative colormaps. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–12, 2018. doi: 10.1145/3173574.3174172 1

[18] J. Mackinlay. Automating the design of graphical presentations of relational information. 5(2):110–141. doi: 10.1145/22949.22950 2

[19] C. M. McColeman, F. Yang, T. F. Brady, and S. Franconeri. Rethinking the ranks of visual channels. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):707–717, 2021. doi: 10.1109/TVCG.2021.3114684 2

[20] M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012. doi: 10.11613/BM.2012.031 6, 7

[21] D. Moritz and D. Fisher. Visualizing a million time series with the density line chart. doi: 10.48550/arXiv.1808.06019 2

[22] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive research: principles and implications*, 3(1):1–25, 2018. doi: 10.1186/s41235-018-0120-9 9

[23] W. Playfair. *The commercial and political atlas: representing, by means of stained copper-plate charts, the progress of the commerce, revenues, expenditure and debts of england during the whole of the eighteenth century*. T. Burton, 1801. 1

[24] G. J. Quadri and P. Rosen. A survey of perception-based visualization studies by task. *IEEE transactions on visualization and computer graphics*, 2021. doi: 10.1109/TVCG.2021.3098240 2

[25] P. S. Quinan, L. Padilla, S. H. Creem-Regehr, and M. Meyer. Examining implicit discretization in spectral schemes. In *Computer Graphics Forum*, vol. 38, pp. 363–374. Wiley Online Library, 2019. doi: 10.1111/cgf.13695 1

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0. 5

[27] S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*. Sage, 2002. doi: 10.2307/2075823 5

[28] 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. doi: 10.1145/3231772 1

[29] E. R. Tufte. The visual display of quantitative information. *The Journal for Healthcare Quality (JHQ)*, 7(3):15, 1985. 1

[30] R. VanRullen. Visual saliency and spike timing in the ventral visual pathway. *Journal of Physiology-Paris*, 97(2-3):365–377, 2003. doi: 10.1016/j.jphysparis.2003.09.010 1, 9

[31] M. Waldner, A. Diehl, D. Gračanin, R. Splechtna, C. Delrieux, and K. Matković. A comparison of radial and linear charts for visualizing daily patterns. *IEEE transactions on visualization and computer graphics*, 26(1):1033–1042, 2019. doi: 10.1109/TVCG.2019.2934784 2

[32] Wikipedia contributors. Geometric brownian motion — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Geometric_Brownian_motion&oldid=1140966364`, 2023. [Online; accessed 2-March-2023]. 3

[33] C. Xiong, C. R. Ceja, C. J. H. Ludwig, and S. Franconeri. Biased average position estimates in line and bar graphs: Underestimation, overestimation, and perceptual pull. 26(1):301–310. doi: 10.1109/TVCG.2019.2934400 2

[34] Y. Zhao, Y. Wang, J. Zhang, C.-W. Fu, M. Xu, and D. Moritz. Kd-box: Line-segment-based kd-tree for interactive exploration of large-scale time-series data. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):890–900, 2021. doi: 10.1109/TVCG.2021.3114865 2