The Paradox of Certainty: When Graphed Ensembles Convey Averages Better than Graphed Averages

Yang Wang (yaw001@ucsd.edu)
Department of Psychology, University of California, San Diego
La Jolla, CA 92037 USA

Sarah H. Kerns (Sarah.H.Kerns.GR@dartmouth.edu)
Department of Psychological and Brain Sciences, Dartmouth College
Hanover, NH 03755 USA

Timothy F. Brady (timbrady@ucsd.edu)
Department of Psychology, University of California, San Diego
La Jolla, CA 92037 USA

Jeremy B Wilmer (jwilmer@wellesley.edu)
Department of Psychology, Wellesley College
Wellesley, MA 02481 USA

Abstract

Data visualizations often display averages without raw data to simplify communication and enhance understanding, especially for lay audiences. However, the theory that such simplification improves understanding remains untested. Here, we test this theory's most basic prediction—that at minimum, the average itself is conveyed better by plotted averages than by plotted raw data. Remarkably, we find the opposite: under a wide range of conditions, overall accuracy of average estimation is higher with raw data. This is due to frequent, severe misinterpretations of both bar and line graphs depicting averages. In contrast, raw data yields some variability but few outright errors; notably, the observed variability is comparable to the uncertainty captured by confidence intervals. We conclude that plotted raw data provides valuable context that helps prevent misunderstandings of the average. Our findings challenge the notion that plotted averages alone yield enhanced understanding and emphasize the value of raw data in communicating evidence.

Keywords: Data visualization, ensemble perception, statistical reasoning, graph comprehension, cognitive biases

Introduction

Data visualizations are essential tools for communicating complex information effectively and efficiently. They allow users to gain insights, identify patterns, and make data-driven decisions by leveraging the human visual system's remarkable ability to process and interpret graphical representations (Cleveland & McGill, 1984; Tufte, 1983). However, the effectiveness of data visualizations relies heavily on the chosen design elements and the user's ability to accurately comprehend the presented information (Shah & Hoeffner, 2002).

One common approach in data visualization is the use of averages or means to summarize and represent central tendencies within datasets (Streit & Gehlenborg, 2014). Averages are widely employed in various graphical representations, such as bar plots and line plots, to provide a concise overview of the data and facilitate comparisons between different groups or conditions (Zacks & Tversky, 1999, Ali & Peeble, 2012). The use of averages is particularly prevalent in fields such as business, finance, and scientific research, where decision-making often relies on summarized information (Montier, 2007; Weissgerber, Milic, Winham, & Garovic, 2015).

However, the reliance on averages in data visualizations raises concerns about the accuracy of users' comprehension and interpretation of the underlying data (Correll & Gleicher, 2014, Newman & Scholl, 2012). Averages, while useful for simplification, can obscure important aspects of the data distribution, such as variability, skewness, and the presence of outliers (Anscombe, 1973). This oversimplification can lead to misinterpretations and flawed decision-making, as users may draw conclusions based on the averages without considering the nuances of the raw data.

One notable example of the limitations of relying solely on averages is the "Bar-Tip-Limit error" (Kerns & Wilmer, 2021). This error occurs when individuals interpret the tops of bars in a bar plot as the upper limit of the data, instead of the average. Consequently, users may underestimate the variability within the data and make incorrect inferences about the underlying distribution. The Bar-Tip-Limit error highlights the need for data visualizations that effectively communicate both the central tendencies and the variability of the data.

Ensemble processing, the human ability to perceive summary statistics from groups of objects, has been studied extensively in various domains, such as visual perception (Ariely, 2001; Chong & Treisman, 2003), numerical cognition (Malmi & Samson, 1983), and decision-making (Maule, 1994). Ensemble processing allows individuals to rapidly extract statistical properties, such as the mean or variance, from a set of objects without consciously attending to each individual item (Whitney & Yamanashi Leib, 2018).

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

This ability is thought to be an evolutionary adaptation that enables efficient processing of complex visual scenes and supports rapid decision-making in dynamic environments (Alvarez, 2011).

Despite the extensive research on ensemble processing in other domains, its role in graph interpretation and data visualization remains underexplored (Cui & Liu, 2021). While some studies have investigated the perception of summary statistics in graphical representations (e.g., Szafir, Haroz, Gleicher, & Franconeri, 2016), there is a lack of comprehensive understanding of how ensemble processing influences the interpretation of averages in different types of data visualizations.

This study aims to address this gap by examining how individuals interpret averages using different graphical representations and the extent to which ensemble processing contributes to accurate comprehension. To investigate this, we conducted a within-subjects experiment in which participants were presented with different data visualizations and asked to judge averages. We compare data visualizations that explicitly show averages, such as bar plots and line plots, with representations that explicitly display raw data points, namely cloud plots and sinaplots (Sidiropoulos, et al., 2018).

Our findings provide valuable insights into the role of ensemble processing in graph interpretation and the importance of displaying raw data in data visualizations. We demonstrate that presenting raw data points can lead to more accurate comprehension of the underlying data distribution, reducing the Bar-Tip-Limit error and enabling users to make more informed judgments. These results have significant implications for the design of effective data visualizations and underscore the need for a paradigm shift towards representations that prioritize the communication of both central tendencies and variability.

Method

Subjects

Participants were 273 undergraduates recruited from the UCSD subject pool. We excluded 11 subjects due to technical difficulties, uncertainty about task requirements, or careless responding. We excluded 41 subjects who failed basic mean calculations, resulting in a final sample (N=221) with demonstrated understanding of statistical averages.

Design

We employed a within-subjects design comparing five visualization types: line plots and bar plots (showing explicit averages highlighted in the instructions), and cloud plots and sinaplots (displaying raw data distributions) from two programs A and B, and "BarCloud" plots that combine the bar plots and cloud plots. Each cloud or sinaplots represented individual fitness test scores, with one program's data normally distributed and the other skewed (±0.75). We systematically varied distribution parameters: sample sizes alternated between 50 and 150 data points, standard deviations between 15 and 25 units, and relative heights between large and small differences. We counterbalanced relative heights (ensuring average score for Program A exceeded B as often as B exceeded A) and skewness direction.

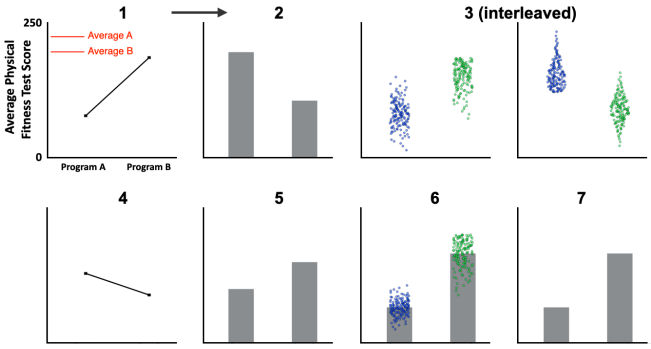

Figure 1. Graph sequence for the procedure. (1) Pre-Line plots (pre-cloud/sinaplots), 2 trials. (2) Pre-Bar plots (pre-cloud/sinaplots), 2 trials. (3) Cloud plot & Sinaplots (interleaved), 32 trials for each. (4) Post-Line plot (post-cloud/sinaplots), 2 trials. (5) Post-Bar plot (post-cloud/sinaplots), 2 trials. (6) BarCloud plot, 2 trials. (7) Bar plot, 1 trial.

Procedure

The experiment presented sequence of trials as illustrated in Figure 1. This design incorporated both explicit (bar plots, line plots and "BarCloud" plots) and implicit (cloud plots and sinaplots) average representation phases to mitigate potential demand characteristics. Subjects were asked to indicate the average fitness test score for each program by using draggable bars. Participants initially viewed two trials each of line plots and bar plots (pre-trials), where the average values were explicitly marked by dots and bar tops, respectively. After the second trial of each plot type, subjects were asked to reflect on their decision-making process. We wanted to rule out the possibility, however small, that participants might feel compelled to place their judgments away from the explicitly marked averages if they only encountered these obvious placements, as they might assume the study was testing their ability to "see beyond" the explicit markers. To address this concern, we interspersed 64 trials of either cloud plots or sinaplots, which displayed raw data distributions without explicit averages and required genuine statistical estimation. Subsequently, participants repeated two trials each of line and bar plots (post-trials), followed by two trials of a hybrid visualization combining bars with overlaid raw data points, and a final single bar plot trial, totaling 75 trials. Similar reflection questions were administered for the post-line plots and the final bar plot. This approach reduced any perceived expectation to deviate from obvious answers in the explicit representation tasks.

Results

Our study examined how individuals make judgments about averages across different graphical representations, comparing explicit average displays (bar plots, line plots, and "bar + cloud" plots) with implicit representations (cloud plots

and sinaplots). One might expect near-perfect average judgment for plots with explicitly marked averages, such as bar and line plots. Yet despite explicit instructions identifying the line plot's dot and the bar plot's top edge of the bar as average indicators, we observed that a substantial proportion of participants made judgments at other locations (Figure 2).

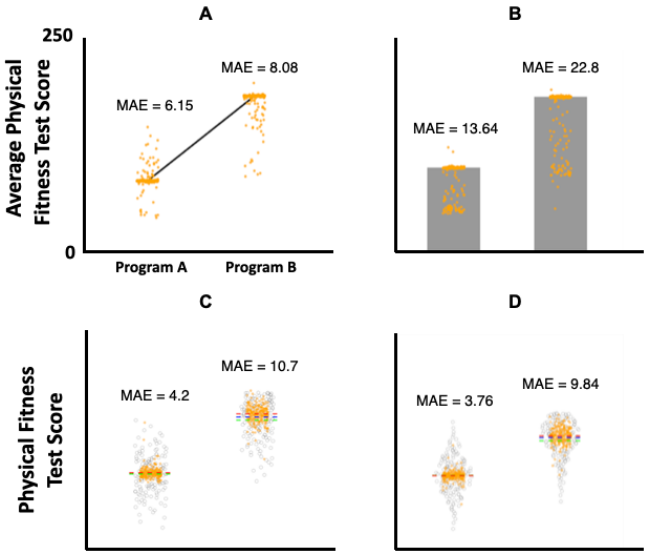

Figure 2: Representative examples of trial responses. (A) Line plot (B) Bar plot (C) Cloud plot (D) Sinaplot. MAE is the mean absolute error of the raw responses. In (C) and (D), the red dashed line is the mean estimated average score for the program, the green dashed lines are the correct mean score for the program, and the blue dashed line is the correct median score for the programs.

To quantify judgment accuracy, we calculated the absolute error between participants' estimates and true means, normalizing these values to reflect the percent deviation from the correct value. We then derived the mean absolute error (MAE) for each participant across all trial conditions (Figure 3). Analysis of the data aggregated across pre-post trials, variance levels, and sample sizes revealed a significant main effect of graph type on MAE via repeated measures ANOVA, $F(5, 1314) = 15.3$, $p < .001$. Post-hoc pairwise paired t-tests with Bonferroni correction demonstrated that bar plots yielded significantly higher normalized MAE compared to other visualization types (all $p < 0.001$). Most notably, bar plots produced significantly larger MAE compared to both normally-distributed raw data plots ($t(220) = 8.49$, $p < .001$, Cohen's $d = 0.57$) and skewed raw data plots ($t(220) = 5.42$, $p < 0.001$, Cohen's $d = 0.37$). Line plots also produced significantly larger MAE compared to normally-distributed raw data plots ($t(220)=4.32$, $p < 0.001$, Cohen's $d = 0.29$). These findings suggest that average estimation from distributed data points can outperform explicit average representations in conventional bar plots and line plots. The accuracy of average estimation for cloud plots and sinaplots, however, varied systematically with the underlying distribution properties. For skewed distributions, participants demonstrated a consistent bias toward the median rather than the mean (Figure 2, C & D). Paired t-tests comparing mean-based error versus median-based error confirmed this bias by showing estimates were significantly closer to the median than to the mean for both cloud plots ($t(220) = 7.29$, $p < .001$, Cohen's $d = 0.25$, 95% CI [0.19, 0.32]) and sinaplots ($t(220) = 7.84$, $p < .001$, Cohen's $d = 0.26$, 95% CI [0.17, 0.34]).

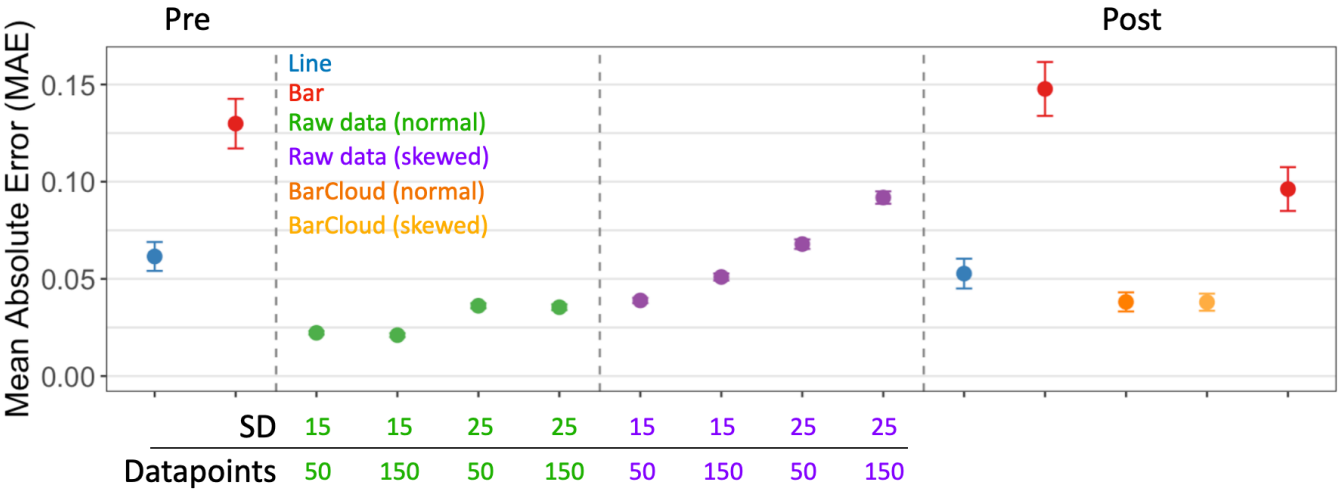

Figure 3. Mean absolute error (MAE) for participants' average judgment across different visualization types, with lower values indicating greater accuracy. Bar plots (red) consistently show the highest error, while line plots (blue) perform better but still exhibit substantial error. Raw data with normal distributions (green) yield the lowest errors, outperforming all other visualization types. For skewed distributions (purple), performance decreases as variability (SD) and sample size increase. BarCloud displays with normal raw data (orange) perform worse than raw data normal displays alone, suggesting that adding bars to raw data can interfere with accurate estimation by introducing competing visual anchors

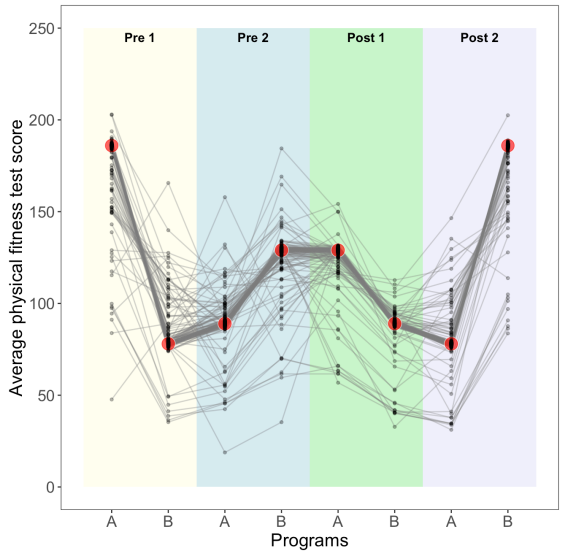

Figure 4. Individual participant response traces across line plot trials. Red dots indicate the correct average fitness test scores for Programs A and B, with gray lines showing how each participant connected their responses across the four phases (Pre1, Pre2, Post1, Post2).

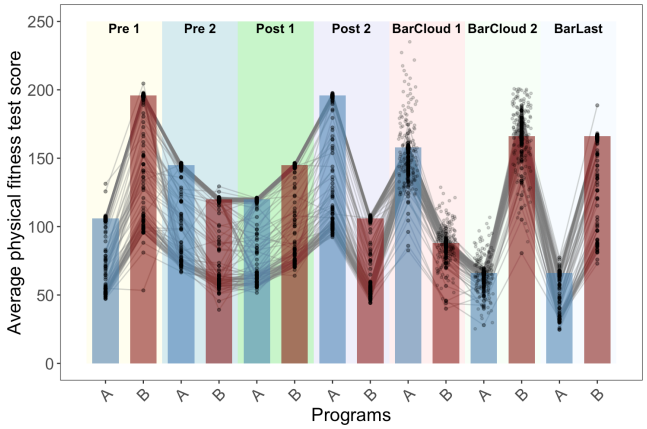

Figure 5. Individual response traces across bar plot trials. The visualization shows participant judgments for fitness test scores across seven trial phases, including Pre-Bar, Post-Bar, BarCloud (combined bars with raw data points), and final bar trials. Gray lines connect individual participant responses.

Despite the bias toward the median in skewed distributions, mean judgments based on raw data visualizations still outperformed most trials with line and especially with bar graphs—where averages were explicitly marked. Figure 4 and Figure 5 illustrate the error distributions for line and bar plots. We observed that the large mean MAE can be attributed to systematic errors patterns which persist across the pre-post trials. To identify systematic patterns in participants' judgments (Figure 6), we conducted an error pattern analysis that integrated quantitative responses with participants' qualitative reflections on their decision-making processes. We developed a response coding scheme to systematically categorize judgment positions. We defined correct judgments as response positions within the true average score +/- 4 (i.e., the score units for the width of the square dot in the line plot). For line graphs, we identified four primary error patterns: On-the-Line (errors falling along the line between the two means), Outside-the-Line (errors falling outside of the line segments appearing to enclose the line segments), Below-the-Line (consistent underestimation), and Above-the-Line (Consistent overestimation). For bar graphs, we categorized four patterns: Middle-Third-Bar (errors in the central third of bars), Top-Third-Bar (errors in the top third of bars), Mixed-Thirds-Bar (errors are not consistently appear in a third) and Above-the-Bar (Consistent overestimation). Notably, the Middle-Third-Bar, Above-the-Bar and Mixed-Third-Bar errors are consistent with the Bar-Tip Limit error (Kerns & Wilmer, 2021).

Overall, we found that 24% of participants for the Pre-Line trials and 23% of participants for the Post-Line trials made an error; and 33% of participants for the Pre-Bar trials and 36% of participants for the Post-Bar trials made an error. Additionally, 28% of participants for the Last-Bar trial made errors, highlighting the persistence of misinterpretations of graphical representations of the average. The classification framework we developed allowed us to quantify the proportion of participants exhibiting each error pattern.

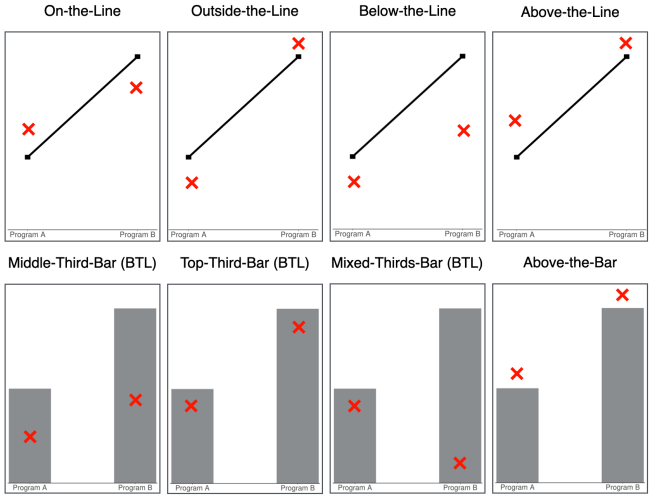

Figure 6: Error patterns.

For line plots (Figure 7), On-the-Line errors were most common, particularly for graph stimuli with large slopes (14.9% in pre-trials and 13.1% in post-trials) compared to small slopes (7.2% in pre-trials and 5.9%% in post-trials). This systematic variation suggests the angle of inclination may be a contributing factor in how participants perceive and interpret average positions in line plots. Participant reflections provided valuable insights into their reasoning processes. For instance, one participant explained, "I have some knowledge of interpreting graphs, and I felt that the average was along those lines because the correlation is ascending," directly illustrating how the line itself guided their judgment. Another participant described their

methodology as, "I just used the first point and halfway point of the line and found the middle of that to place the line for each program," revealing a systematic interpolation strategy that demonstrates how the visual structure of the line plot can lead to systematic misinterpretations. Below-the-Line errors were the second most frequent, with a slight increase in post-trials (6.8-8.6%) compared to pre-trials (3.2-5.9%). This pattern suggests participants sometimes anchored their judgments to lower reference points in the graph. As one participant explained, "I just estimated [the] middle distance between the dots and the x-axis," highlighting a tendency to divide the vertical space rather than recognizing the dots themselves as the marked averages. Notably, this latter behavior is consistent with the notion that the Bar-Tip Limit Error that can, at least occasionally, happen in the absence of a bar (Kerns & Wilmer, 2021).

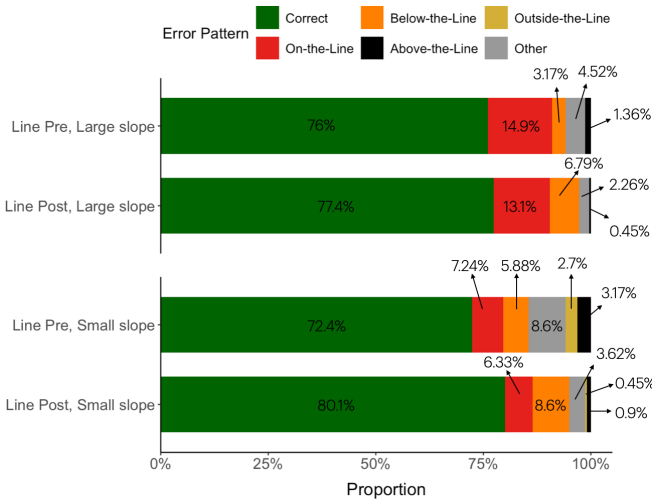

Figure 7. The proportions for error patterns in line plots.

For bar plots (Figure 8), Middle-Third errors were predominant, occurring in 18.3% of pre-trials, 24.4% of post-trials, and 12.7% of the final trial. One participant directly articulated this misconception, stating, "I placed each average in the middle of each bar because the middle is what I presumed to be the average." This finding aligns directly with the Bar-Tip Limit error, indicating that participants frequently perceive the average as located within the bar rather than at its top edge. Top-Third errors (5.9-8.1%) and Mixed-Thirds errors (3.2-5.9%) were less common but still notable. These errors represent different manifestations of the Bar-Tip Limit effect, as evidenced by participant reflections. For Top-Third errors, one participant reasoned, "The top of the bar is the highest and without knowing the real lowest scores, I assume most scores are concentrated close to the top of the bars," revealing a tendency to infer distributional properties not depicted in the visualization. Mixed-Thirds errors were established as a classification category to capture the remaining instances of Bar-Tip Limit errors that did not consistently fall into either the Middle-Third or Top-Third patterns. All three categories ultimately represent the same fundamental misinterpretation: perceiving the bar as

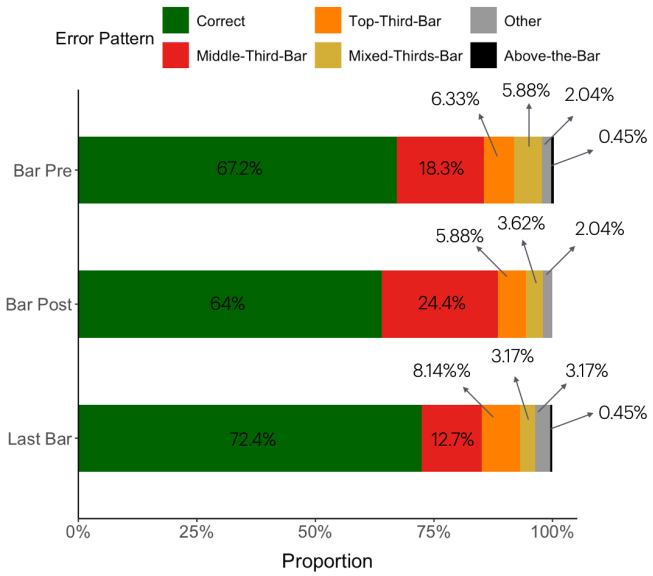

Figure 8. The proportion of error patterns in bar plots.

representing a distribution rather than a single average value at its endpoint.

Discussion

Our study provides valuable insights into the perception and interpretation of averages in data visualizations, highlighting the importance of displaying raw data. The findings have significant implications for the design of effective data visualizations and the communication of statistical information.

One of the most surprising and compelling findings of our study is that the overall accuracy for bar plots and line plots, which explicitly mark the average value, was worse than that of cloud and sinaplots with normally distributed data or even the skewed data in some cases. This result challenges the common assumption that explicitly displaying averages improves comprehension and interpretation.

This finding also highlights the importance of considering the distribution of individual responses and the potential for systematic errors, rather than relying solely on aggregate measures of accuracy. If we had focused only on the average performance, we might have missed the presence of systematic errors in the interpretation of bar and line graphs. By revealing these systematic errors, we gain insights into underlying mechanisms.

These results underscore the crucial role of displaying raw data points alongside summary statistics. Without the context provided by the raw data, individuals may be more susceptible to biases and misinterpretations, leading to flawed decision-making (Tufte, 1983; Weissgerber, Milic, Winham, & Garovic, 2015). Our findings challenge the common assumption that displaying averages alone is sufficient for effective data communication.

By intentionally contrasting our own aggregate results with the deeper insights gained from error pattern analysis, we aim to emphasize the importance of thorough and critical analysis in data visualization research. It serves as a valuable reminder

to look beyond surface-level findings and consider the distribution of individual responses to uncover the true patterns and biases that shape the interpretation of data.

The identification of different error types, such as the Bar-Tip Limit error and its connection to the "Middle-Third Error," sheds light on the cognitive processes involved in interpreting data visualizations. In the absence of explicit raw data, a substantial proportion of participants appear to have a strong prior expectation that the data points are clustered near the middle of the bar. This tendency can lead to the perception of only half the actual difference, on average, when comparing two groups, potentially distorting the interpretation of the data. The observation that participants often anchor and shift from the middle point of the bar suggests a strong prior belief about higher data density towards the top edge of the bar. This finding has important implications for the design of data visualizations, as it indicates that the mere presence of a bar can influence people's perceptions and lead to systematic biases. Even a single line connecting to the two dots that represent the averages can mislead people to interpolating the averages along the lines. While future work will be necessary to isolate the precise cognitive mechanism for this interpolation, we speculate that analogous to the Bar-Tip Limit error, where the bar contains the data, it could be that viewers assume that the line contains some or all of the data. If that were the case, then the average would be pulled away from the average values, at the end of the line, and toward the middle of the line.

Our study suggests that a notable portion of participants in the "BarCloud" trials appeared to overlook the bar representation (showing the actual mean) and instead seemed to rely on their own visual estimates from the cloud distribution. This tendency was observed more frequently when the cloud depicted skewed data (Figure 4). This behavior may indicate that humans tend to compute statistical averages from distributions they observe, even when explicit summary statistics are provided. This observation potentially sheds light on how our perceptual system processes statistical information, sometimes favoring the raw distribution over presented summary values, particularly with asymmetrical distributions. These findings could have implications for visualization design approaches, especially when presenting statistical information for non-normal distributions where summary statistics alone might not tell the complete story.

While our study shows that displaying raw data helps people tap into their natural ability to process multiple items at once, this visual approach isn't always accurate for determining exact averages. Previous research has shown that mean perception for skewed data with large variance and large sample sizes as well as aggregated mean judgments with multimodal data or outliers can result in substantial estimation errors (Wang & Brady, 2021). However, in the case of skewed data, the bias towards the median can be considered beneficial, as the median is often a more appropriate measure of central tendency. Notably, the observed variability in these raw data estimates is comparable to the uncertainty captured by confidence intervals, suggesting that viewers' natural variability in judgment reflects meaningful statistical properties of the data. The key insight here is that the accuracy of mean estimation from raw data is not the sole consideration; displaying raw data can reduce errors like the bar-tip-limit error and facilitate a more comprehensive understanding of the data distribution.

The current study raises important pedagogical considerations, as our data reveal various misunderstandings of averages in data visualization and the concept of average itself. These findings underscore the need for improved statistical education and the development of intuitive and accessible data visualization methods. By addressing these challenges, we can empower individuals to make more informed decisions based on accurate interpretations of data.

Future research should aim to extend our findings to more diverse populations and explore the ecological validity of our results. The current sample consisted of undergraduate students from a single institution, and it would be valuable to collect data from a more general population. Additionally, future studies could employ one-shot judgment tasks with a variety of contextual information, such as independent-measure and repeated-measure samples, to assess the robustness of our findings in different settings.

In conclusion, our study highlights the importance of displaying raw data in data visualizations to facilitate accurate perception and interpretation of averages. The findings challenge the reliance on summary statistics alone and emphasize the need for comprehensive representations that leverage ensemble processing mechanisms. By understanding the cognitive processes and errors involved in interpreting data visualizations, we can design more effective visual representations and improve statistical education. Ultimately, our research contributes to the development of a new data presentation paradigm that prioritizes transparency, accuracy, and informed decision-making.

Acknowledgement

We thank anonymous reviewers for insightful comments.

References

Ali, N., & Peebles, D. (2012). The Effect of Gestalt Laws of Perceptual Organization on the Comprehension of Three-Variable Bar and Line Graphs. *Human Factors*, *55*(1), 183-203.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. Trends in Cognitive Sciences, 15(3), 122-131.

Anscombe, F. J. (1973). Graphs in statistical analysis. The American Statistician, 27(1), 17-21.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. Psychological Science, 12(2), 157-162.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. Vision Research, 43(4), 393-404.

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

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. IEEE Transactions on Visualization and Computer Graphics, 20(12), 2142-2151.

Cui, L., Liu, Z. Synergy between research on ensemble perception, data visualization, and statistics education: A tutorial review. *Atten Percept Psychophys* **83**, 1290–1311 (2021).

Kerns, S. H., & Wilmer, J. B. (2021). Two graphs walk into a bar: Readout-based measurement reveals the Bar-Tip Limit error, a common, categorical misinterpretation of mean bar graphs. *Journal of Vision, 21*(12), Article 17.

Malmi, R. A., & Samson, D. J. (1983). Intuitive averaging of categorized numerical stimuli. Journal of Verbal Learning and Verbal Behavior, 22(5), 547-559.

Maule, A. J. (1994). A componential investigation of the relation between structural modelling and cognitive accounts of human judgement. Acta Psychologica, 87(2-3), 199-216.

Montier, J. (2007). Behavioural Finance: Insights into Irrational Minds and Markets. John Wiley & Sons.

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. Psychonomic Bulletin & Review, 19(4), 601-607.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. Educational Psychology Review, 14(1), 47-69.

Sidiropoulos, N., Sohi, S. H., Pedersen, T. L., Porse, B. T., Winther, O., Rapin, N., & Bagger, F. O. (2018). SinaPlot: An Enhanced Chart for Simple and Truthful Representation of Single Observations Over Multiple Classes. Journal of Computational and Graphical Statistics, 27(3), 673–676.

Streit, M., & Gehlenborg, N. (2014). Bar charts and box plots. Nature Methods, 11(2), 117.

Szafir, D. A., Haroz, S., Gleicher, M., & Franconeri, S. (2016). Four types of ensemble coding in data visualizations. Journal of Vision, 16(5), 11.

Tufte, E. R. (1983). The visual display of quantitative information. Graphics Press.

Wang, Y., & Brady, T. F. (2023, May 9). Intuitive Global Mean Estimation in Scatterplots with Spatial Clusters. https://doi.org/10.31219/osf.io/w4xq3

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. PLoS Biology, 13(4), e1002128.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. Annual Review of Psychology, 69, 105-129.

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. Memory & Cognition, 27(6), 1073-1079.