The Uniformity Fallacy: A Second Common, Severe Misinterpretation of Bar Graphs of Averages

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

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

Abstract

Past methods for studying graph interpretation have only indirectly assessed people's mental picture of the data that produced the graph. Recently, we developed a more direct, drawing-based measure and used it to reveal a severe misinterpretation of the common bar graph of averages: one in five viewers mistook the average for the data's outer limit. Here, we use the same measure to reveal a second misinterpretation, whereby even more viewers—one in three—incorrectly assume that data frequency remains approximately uniform over its entire range. Missing from their mental picture are the tails of the distribution—the relative rarity of extreme values—which are so characteristic of real data that they are embedded in the core normality assumption of statistics. We label this misinterpretation the "Uniformity Fallacy" and characterize its nature, reproducibility, generalizability, and correlates. We conclude that bar graphs of averages fail to communicate data truthfully in not one, but two fundamental ways.

Keywords: bar graph; averages; statistics; data visualization; graph interpretation; education

Introduction

Arguably, the most important role of a graph is to communicate data truthfully—ensuring that the mental image formed in the viewer's mind closely reflects the underlying data. In this study, we examine that mental image in an unusually direct way, revealing a fundamental aspect of the data that is misrepresented by the common bar graph of average values.

Bar graphs of averages have long been criticized—especially in the sciences—for their ambiguity and for concealing the individual values that were averaged to produce the displayed mean (Cleveland & McGill, 1984; Drummond & Vowler, 2011; Larson-Hall, 2017; Pastore, Lionetti, & Alto, 2017; Rousselet & Pernet, 2017; Vali & Wilkinson, 2020; Wainer, 1984; Weissgerber et al., 2015). Journal editors, for example, have urged authors to "kick the bar chart habit... [for] data that they cannot represent well" (Editors, 2014) and to "show the dots in plots... the data points are the context" (Editors, 2017). Peer-reviewed articles have echoed these calls, urging readers to "show the data" (Drummond & Vowler, 2011) and to "reveal, don't conceal" (Weissgerber et al., 2019). Nearly four decades ago, the first bullet point on page one of Tufte's seminal book The Visual Display of Quantitative Information was simply: "show the data" (Tufte, 1983).

The critique that bar graphs of averages are ambiguous is not new—it has persisted for decades. While this ambiguity is, in some sense, logically self-evident, what remains largely unknown is whether, and to what extent, it leads to variable or inaccurate understanding. That is, with few exceptions (Kerns & Wilmer, 2021; Newman & Scholl, 2012), direct evidence on how ambiguity in bar graphs of averages affects comprehension remains scarce.

Perhaps partly for this reason, bar graphs of averages remain widely used today (Chen et al., 2017; Larson-Hall, 2017; Mogull & Stanfield, 2015; Weissgerber et al., 2015; Weissgerber et al., 2019). They are especially prevalent in sources like introductory psychology textbooks (Gray & Bjorklund, 2017; Grison & Gazzaniga, 2019; Kalat, 2016; Myers, 2017), which are aimed at audiences that often lack a strong background in statistics. It is commonly assumed that the visual simplicity of bar graphs of averages promotes clarity, particularly for non-experts (Angra & Gardner, 2017; Barton & Barton, 1987; Zubiaga & Mac Namee, 2016). Yet this assumption—that simplicity enhances understanding—like the claim that ambiguity impairs comprehension, is largely unsupported by direct evidence.

The potential negative consequences of ambiguity in bar graphs of averages are underscored by educational researchers such as Goodchild, who writes: "In schools, we teach the calculation of [the mean], but we pay little attention to the reverse process of understanding what the [mean] tells us about the population it summarizes... This reverse process should not be taken for granted and should be part of our statistics teaching schemes" (1988). Similarly, Mokros and Russell (1995) quoted an eighth-grade student as saying, "I know how to get an average, but I don't know how to get the numbers to go into an average, from an average."

In this study, we aim to document—concretely and in detail—how the "reverse process," or reverse-engineering step, may function when it is, indeed, taken for granted in the context of bar graphs of averages (Goodchild, 1988). A major limitation of prior research on this process has been its reliance on indirect methods (Kerns & Wilmer, 2021; Hullman et al., 2018). To address this, we recently developed and validated a more direct approach, called *Draw Datapoints on Graphs* (DDoG). In this method, viewers are asked to sketch both the graph and the data they believe produced it (Kerns & Wilmer, 2021).

In that recent work, the DDoG method uncovered a common, severe, and previously undocumented misinterpretation of bar graphs of averages: one in five viewers mistakenly believed that most or all of the data fell within the bar itself, interpreting the bar's tip as the outer limit of the data rather than its average (Kerns & Wilmer,

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

2021). This misinterpretation, illustrated in Figure 1, was termed the *Bar-Tip Limit Error* (BTL Error) (Kerns & Wilmer, 2021). Notably, the average is the one aspect of the data that bar graphs of averages explicitly—and theoretically unambiguously—represent. It is therefore striking that such graphs so often fail to convey even this most basic concept accurately.

Nevertheless, if 4 in 5 viewers were to form an otherwise accurate understanding of the data, that might arguably be sufficient for certain purposes. In this study, we apply the same DDoG method to examine the mental image viewers form of a key aspect of the data that is obscured—and thus remains ambiguous—in a bar graph of averages: the shape of the data distribution.

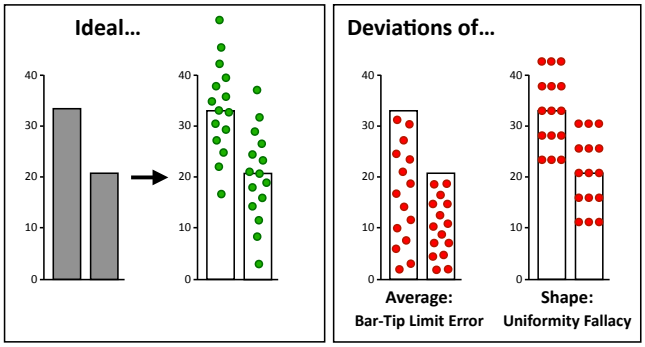

Figure 1: Ideal and Two Deviations. Left panel: Ideally, a bar graph of averages (gray) implies normally distributed data centered on the averages (green dots). Right panel: Two common, extreme deviations (red dots), one from the shown average (Bar-Tip Limit Error; Kerns & Wilmer, 2021), and the other from a normal distribution (Uniformity Fallacy).

In theory, everyday experience should foster a natural inclination to expect a normal (bell-shaped) distribution, where data density is highest near the center and decreases toward the extremes, forming tails. For instance, most people encounter normal distributions in contexts such as human height or temperature. They observe the relative rarity of very tall or short individuals, as well as of extremely hot or cold days. Yet, do people internalize and generalize these experiences sufficiently to apply them to a novel data source presented ambiguously through a bar graph of averages?

We will demonstrate here that the answer is no. Approximately 1 in 3 viewers intuit no tail at all, and about half of those who remain intuit data that is closer to tail-less than normal. We label the former, more extreme misinterpretation, illustrated in **Figure 1**, the Uniformity Fallacy.

Methods

Open data—comprising participant drawings, quantitative drawing-based measures, and survey responses—and open materials—annotated screenshots of the procedure as presented to participants, including the exact wording of the questions—are available at osf.io/ymb7g.

Procedure

Figure 2 shows the basic procedure alongside actual participant responses corresponding to each of the three illustrations in **Figure 1**. Participants were shown four graph stimuli—one at a time, in randomized order—and were instructed to sketch the graph along with their best guess of 20 values that could have produced the displayed averages.

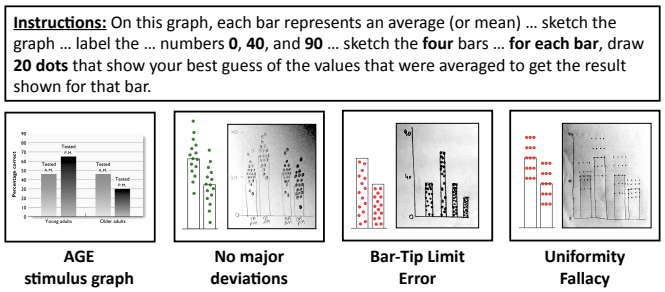

Figure 2: Procedure and Responses. Shown are key excerpts from the drawing task (instructions top, example stimulus left) and three real illustrative responses matching the cartoons shown in **Figure 1**.

Figure 3 displays the four stimulus graphs, each accompanied by an example response that demonstrates the Uniformity Fallacy. The stimulus graphs were selected based on several criteria: (A) they had to present real, published results, allowing interpretations to be evaluated against a ground truth; (B) they needed to depict human data, due to its intuitive appeal and significance; (C) to ensure broad relevance, the graphs had to be widely consumed by an audience with diverse levels of expertise and explicitly designed for that purpose; and (D) to test the generalizability of findings, the stimuli needed to vary in form (unidirectional vs. bidirectional), research design (experimental vs. correlational), and measurement type (subjective vs. objective).

To meet these criteria, the graph stimuli were taken directly from four widely used introductory psychology textbooks (Gray & Bjorklund, 2017; Grison & Gazzaniga, 2019; Kalat, 2016; Myers, 2017), each depicting a published result (DiMascio et al., 2018; Festinger & Carlsmith, 1959; May, Hasher, & Stoltzfus, 1993; Rahman, Wilson, & Abrams, 2004). For clarity, we refer to each stimulus below by its independent variable: AGE, CLINICAL, SOCIAL, and GENDER.

The detailed development, rationale, and validation of our drawing task—along with a review of relevant literature—are discussed elsewhere (Kerns & Wilmer, 2021). Briefly, we selected this task for its well-documented reliability, validity, accessibility, and information richness (Kerns & Wilmer, 2021). Although drawing tasks are recognized as an effective method for revealing the contents of thought (Bainbridge, 2021; Fan et al., 2023), they have been largely overlooked as a means of studying how people interpret data visualizations (Hullman et al., 2018; Elliot et al., 2020; Kerns & Wilmer, 2021; cf. Kim et al., 2017).

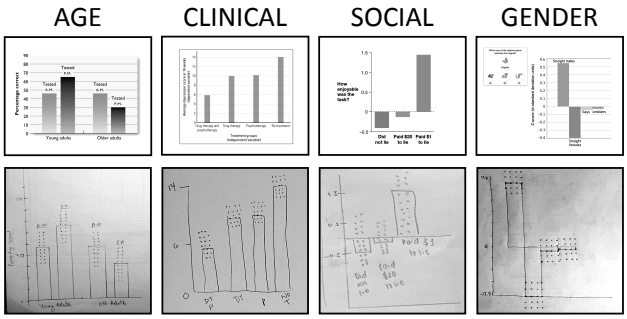

Figure 3: Uniformity Fallacy for each stimulus graph. Shown are the four graph stimuli (top row), and, for each, an example of the Uniformity Fallacy (bottom row). We refer to these graph stimuli below by their independent variables: AGE, CLINICAL, SOCIAL, GENDER.

The computer program WebPlotDigitizer was used to record the axes and data point locations drawn for the two target bars in each stimulus graph—those representing the highest and lowest mean values (Rohatgi, 2015). This method captures the precise numerical values drawn by each participant, allowing the sketches to be analyzed as quantitative data.

After completing the drawing task, participants filled out a demographic survey that included questions about age, handedness, gender, educational attainment, and prior coursework in both psychology and statistics. They also answered a series of graph comprehension questions, which were not the focus of the present investigation.

Participants

Data collection was conducted remotely using Qualtrics online survey software. Participants were recruited through Testable Minds, with no selection criteria based on location or education, and were compensated for their participation. The 170 participants who completed the study represented 20 countries across five continents. Of these, 106 identified as male and 64 as female, and ages ranged from 18-71 (M=30, SD=9.3).

Six participants were excluded due to technical issues—one upload failure and five cases of poor image quality. An additional 30 participants were excluded based on predetermined criteria identical to those used in the previous study (Kerns & Wilmer, 2021). Participants were excluded if three or more of their individual drawings were unusable due to a basic failure to follow instructions. Specifically, exclusions occurred for the following reasons: data points showed no covariation with bar height or direction (1 participant); no data points were included (11 participants); or the number of data points was either too large (>25 per bar, 1 participant) or too small (<15 per bar, 16 participants).

In total, 134 participants met inclusion criteria, resulting in a usable data yield of 79%, which is typical for online studies of this length (Litman & Robinson, 2020). Each included participant provided four drawings, for a total of 536 graphs. Individual drawings were further excluded if they met any of the same predetermined exclusion criteria: unreadable dots (2), no covariation with bars (4), no dots (3), too many dots (4), or too few dots (3). The final dataset comprised 520 usable graphs, distributed across the four stimuli as follows: AGE – 133, CLINICAL – 131, SOCIAL – 130, and GENDER – 126.

Measures

The Tail Index. The normal distribution (also known as Gaussian or bell-shaped) is a fundamental assumption underlying most statistical analyses, including all those reported in the published studies from which our stimulus graphs were derived (DiMascio et al., 2018; Festinger & Carlsmith, 1959; May, Hasher, & Stoltzfus, 1993; Rahman, Wilson, & Abrams, 2004). To assess the extent to which participants' drawn data distributions did—or did not—exhibit the tails characteristic of a normal distribution, we developed a measure called the *Tail Index*.

The Tail Index is calculated as a linear transformation of kurtosis—or "tailedness"—the so-called fourth moment of a distribution (the first three moments being the mean, standard deviation, and skewness). For intuitive interpretation, a perfectly normal distribution corresponds to a Tail Index score of 50, while a perfectly uniform, flat, tail-less distribution—where all observed values are equally likely—corresponds to a Tail Index score of 0.

Notably, Tail Index scores can fall outside the 0 to 50 range. A score above 50 indicates a distribution with tails that extend farther than those of a normal distribution, with values exceeding 100 possible in cases involving extreme outliers. At the lower end, scores below 0 occur when the tails are heavier than the center of the distribution, as seen in bimodal distributions. **Figure 4** presents visual examples of Tail Index scores of 0, 50, and 100.

As we will show, Tail Index scores near 50 are rare, while scores at or below 0—the value expected for a perfectly uniform distribution—are common. We define the Uniformity Fallacy as any Tail Index score of 0 or below. For each stimulus graph, kurtosis was averaged across the two target bars, and this average was then used to calculate the Tail Index. Each participant received up to four Tail Index scores—one for each qualifying stimulus graph drawing.

The Bar-Tip Limit Index. This index quantifies the relationship between the drawn data and the mean value, represented by the tip of the bar. It is calculated as the proportion of a bar's drawn data points that fall outside the bar itself. If the data are symmetrical—as assumed in the analyses of the four original studies from which our stimulus graphs were derived (DiMascio et al., 2018; Festinger & Carlsmith, 1959; May, Hasher, & Stoltzfus, 1993; Rahman, Wilson, & Abrams, 2004)—the expected value for this index is 50, reflecting 10 of the 20 drawn data points on either side of the bar's tip.

Based on prior work (Kerns & Wilmer, 2021), the Bar-Tip Limit Error is defined as a Bar-Tip Limit Index score of 20 or below. In that earlier study, Bar-Tip Limit Index scores were found to be bimodal (Kerns & Wilmer, 2021), with one mode near 50—interpreting the bar tip as the center of the data—and another near 0—treating the bar tip as the boundary or limit of the data. The cutoff score of 20 was determined using standard clustering methods and serves as

a data-driven threshold for identifying the Bar-Tip Limit Error.

Results

Frequency of Severe Misinterpretations

We first assess the frequency of severe misinterpretations. Table 1 shows the frequency of both the Bar-Tip Limit Error and the Uniformity Fallacy, reported separately for each stimulus graph as well as aggregated across all stimulus graphs. The results were consistent across the different stimulus graphs, with no pairwise comparisons reaching statistical significance ($p > 0.05$).

Table 1: Frequency of severe misinterpretations.

Stimulus Graph	Total #	Bar-Tip Limit Error	Uniformity Fallacy	Either or Both
AGE	133	29 (22%)	49 (37%)	63 (47%)
CLINICAL	131	31 (24%)	49 (37%)	64 (49%)
SOCIAL	130	41 (32%)	43 (33%)	70 (54%)
GENDER	126	38 (30%)	41 (33%)	67 (53%)
TOTAL	520	139 (27%)	182 (35%)	264 (51%)

In total, the Bar-Tip Limit Error occurred in 27% of drawings (95% CI [23%, 31%]), or approximately 1 in 4. The Uniformity Fallacy occurred in 35% of drawings (95% CI [31%, 39%]), or about 1 in 3. At least one of these two severe misinterpretations appeared in 51% of drawings (95% CI [47%, 55%]), or just over 1 in 2.

These results replicate—and, if anything, exceed—the previously reported high frequency of approximately 1 in 5 for the Bar-Tip Limit Error (Kerns & Wilmer, 2021). Additionally, they highlight an even more prevalent misinterpretation: the Uniformity Fallacy.

Together, these misinterpretations affect a significantly greater number of drawings than either does alone, providing strong evidence of their substantial independence. Consistent with this, 16% (82) of drawings exhibited only the Bar-Tip Limit Error, 24% (125) showed only the Uniformity Fallacy, and just 11% (57) displayed both.

Rarity of Normal Distributions

To complement the results above, which document the frequency of severe misinterpretations, we now examine the full spectrum of interpretations revealed by the Tail Index. **Figure 4** shows Tail Index scores by stimulus graph, color-coded to indicate the degree of deviation from 50—the score associated with a perfectly normal distribution.

The results are consistent across stimulus graphs. In all cases, the majority of scores fall well below 50, with a mode near 0, which corresponds to perfect uniformity. Overall, 68% of drawings (356) are closer to uniform than to normal (represented by the red and yellow regions on the left), while only 23% (120) fall within 25 units of normal. Although some drawings exhibit excessive tails (Tail Index scores above 75), these cases are rare, accounting for just 8% (44) of responses.

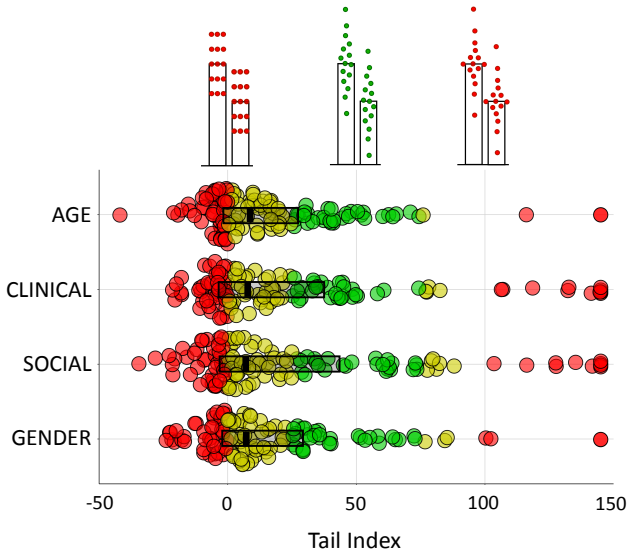

Figure 4: Rarity of Normal Distribution. The Tail Index score (x-axis) for each participant's drawing (data point) is shown for each stimulus graph (y-axis). Illustrative cartoons depict distribution shapes for key Tail Index scores, with 50 representing a normal distribution and 0 representing a uniform distribution. Data points are color-coded according to their distance from the normal distribution: green for values within 25 points, yellow for values between 25 and 50, and red for values above 50. The red data points on the left represent what we define as the Uniformity Fallacy, while the yellow data points on the left are closer to uniform than normal. The black boxes indicate the interquartile ranges, and the black lines represent the medians.

Notably, while the Bar-Tip Limit Error is an inherently categorical phenomenon with a strongly bimodal distribution (Kerns & Wilmer, 2021), the Uniformity Fallacy represents the extreme end of a continuous, albeit highly skewed, distribution of Tail Index scores.

Independence of Error Types despite Similar Behavior Across Stimulus Graphs

Figure 5 highlights two key findings. First, for both the Bar-Tip Limit Index and the Tail Index, higher scores on one stimulus graph are strongly correlated with higher scores on the other stimulus graphs, with average correlation values of $\rho = 0.70$ and 0.63, respectively. These correlations indicate that participant behavior is consistent across stimulus graphs. Second, a higher score on the Bar-Tip Limit Index does not predict a higher score on the Tail Index, as the average correlation is only $\rho = 0.05$. This low correlation reflects a strong independence between the tendency toward the Uniformity Fallacy and the tendency toward the Bar-Tip Limit Error. This independence aligns

with the earlier finding that drawings exhibiting both misinterpretations are relatively rare.

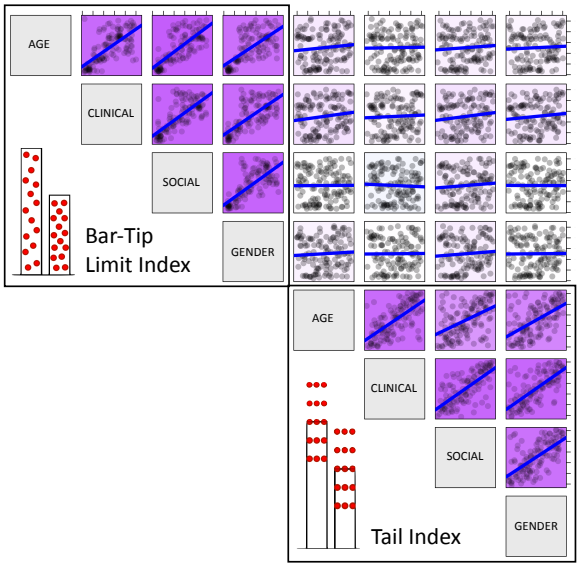

Figure 5: The two errors do not correlate with each other. The figure shows a scatterplot matrix of nonparametric (Spearman's ρ, or rho) correlations computed on rank-ordered data for stimulus graphs and indices. The physical slope of each least squares line (blue) equals ρ, and the purple tint is proportional to ρ. Results are similar for parametric (Pearson) analyses but are presented here in nonparametric form as a demonstration of robustness. The full data set is available at osf.io/ymb7g.

Education and Confidence

Figure 6 presents the correlations between the two indices (averaged across stimulus graphs) and various demographic and self-report measures. In a context of generally similar, low correlations between the Tail Index and other variables, two key observations stand out. First, the Tail Index correlates little to none with educational attainment (r = 0.05), prior statistics coursework (r = 0.11), and prior psychology coursework (r = 0.11) suggesting that standard education does not necessarily produce an expectation of normally distributed data. Second, the moderate correlation between higher confidence in dot placement and less accurate (more uniform) Tail Index scores (r = -0.27) indicates that graph viewers may be unaware of the limitations in their understanding of the data underlying an average value.

Discussion

In this study, we used a recently developed, drawing-based measure to assess people's understanding of the data distributions underlying bar graphs of averages. We discovered a common, severe error in the estimation of distribution shape. Approximately one in three viewers displayed what we call the Uniformity Fallacy—assuming a uniform distribution instead of a normal one. Even among those who did not display this fallacy, another third inferred a distribution closer to uniform than to normal.

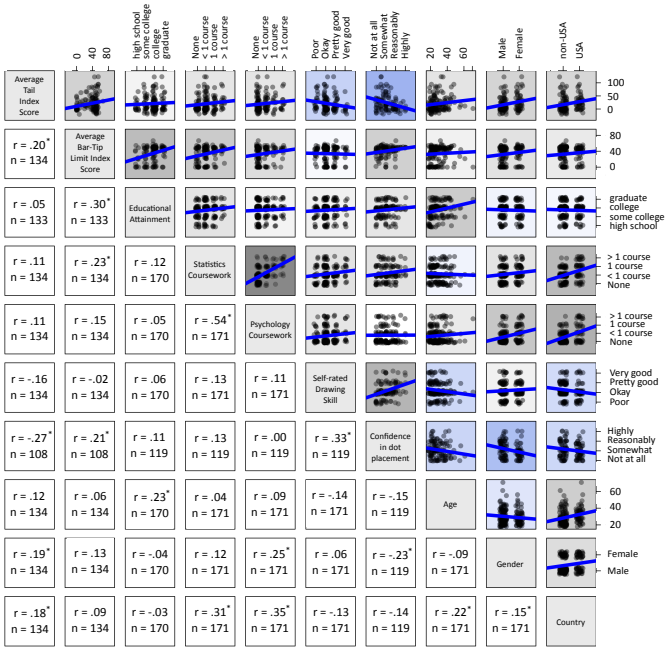

Figure 6: Demographic and Self-report Correlations. Shown is a full matrix of scatterplots (above the diagonal) and correlations (below the diagonal) for our two main indices (Bar-Tip Limit Index and Tail Index, averaged across stimulus graphs) and demographic and self-report measures: educational attainment, statistics coursework, psychology coursework, self-rated drawing skill, confidence in dot placement, age, gender, and country. Statistically significant correlations ($p<0.05$, two-tailed) are marked with an asterisk. The physical slope of each least squares line (blue line) equals Pearson's correlation coefficient r, and the scatterplot tint (gray=positive, blue=negative) is proportional to r. Nominal and ordinal variables are coded as sequential numbers (e.g., non-USA=0, USA=1; poor=1, okay=2, pretty good = 3, very good = 4) to allow for computation of meaningful correlation values. Results are similar for nonparametric (Spearman) correlations but are presented here in parametric form for understandability. The full data set is available at osf.io/ymb7g.

These results exhibited substantial generality. They replicated not only qualitatively, but also quantitatively, across four distinct bar graph stimuli, with substantially varied form and content (**Figure 4**). This imperviousness to content and form raises the question of whether they may extend to an even wider array of graph types, or to non-visual ways of communicating average values. Results also remained roughly constant across varied general and statistical educational backgrounds. This independence from prior training suggests the presence of basic intuitions that may be difficult to override. For comparison, a growing body of research on the Bar-Tip Limit Error indicates that even surprisingly clear and direct instructions are often

insufficient to eliminate it (Cui et al., 2024; Wang et al., 2025). Potentially, the same could be true of the erroneous tendency to assume flat distributions.

Notably, there was little correlation between the Tail Index, which captures the Uniformity Fallacy, and the Bar-Tip Limit Index, which captures the Bar-Tip Limit Error (**Figure 5**). This weak relationship suggests that misinterpretation of the distribution shape is largely independent from misinterpretation of the average. Supporting this, many viewers exhibited one of the two errors without exhibiting the other (**Table 1**). Together, these two severe misinterpretations appeared in more than half of the drawings we collected, indicating a high prevalence. Combined with our prior identification of the Bar-Tip Limit Error (Kerns & Wilmer, 2021), our current identification of the Uniformity Fallacy strengthens the broader argument that bar graphs of averages frequently, severely, and in multiple ways fail to communicate data accurately.

Plausible mechanisms. What intuition might lead so many viewers to imagine an unrealistically flat or uniform distribution? One possibility is that people perceive data as being generated by a process similar to individual dice rolls, where all plausible values (1–6) are equally likely. However, this reasoning overlooks the fact that most measures of complex, multifactorial systems—such as human behavior, biological processes, the weather, and economies—are more akin to the outcome of countless summed dice rolls or coin flips. Just as 10 coin flips are far more likely to produce 5 heads than 0 or 10 heads, the normal distribution—with its frequent central values and rare extremes—routinely emerges from measurements of complex systems.

A related hypothesis involves what Inhelder and Piaget referred to as compensation (Inhelder & Piaget, 1951). Compensation is the assumption that an event which has not occurred recently is more likely to occur soon. For example, if three coin flips yield "heads," one might believe the next flip is likely to be "tails," or if ten dice rolls fail to produce a "1," a "1" is considered more likely on the next roll. At its core, compensation reflects a mistaken belief that independent events (such as coin flips or dice rolls) are dependent, with an inherent tendency to "even out." While Piaget and Inhelder characterized compensation in children (1951), it is plausible that a similar cognitive process leads adult viewers of bar graphs to assume that data will "even out" to produce a uniform distribution.

Strengths of our drawing based approach. It is worth asking why the Uniformity Fallacy, so common and salient in our present results, was not identified by prior studies. We believe there is an important analogy to our earlier discovery of the Bar-Tip Limit Error (Kerns & Wilmer, 2021). In that case, prior research had suggested that viewers may have a tendency to consider within-bar data values more likely than outside-the-bar values (Newman & Scholl, 2012; Correll & Gleicher, 2014). However, the measures used in those studies were too abstract and information-poor to capture the phenomenon in its true form.

In contrast, our drawing-based measure—concrete and information-rich—clearly revealed that 1 in 5 viewers believed nearly all the data was inside the bar, and just as clearly, it showed that the other 4 in 5 viewers did not make this (categorical) error. Since our initial discovery of the Bar-Tip Limit Error (Kerns & Wilmer, 2021), it has been replicated multiple times using different methods (e.g., et al., 2024; Wang et al., 2025). Yet our drawing method provided the eyes to see the error clearly, which enabled that subsequent work.

A key virtue of our drawing-based approach is that it generates a complete, concrete, and information-rich hypothesized dataset that can be analyzed like real data. Here, this allowed us to explore participants' understanding of distribution shape without explicitly teaching the concept, avoiding potential confusion or bias. Additionally, the expressive freedom of pencil-and-paper drawing allowed for a wide range of intentional responses and helped us identify and distinguish various types of confused or inaccurate responses.

Viable alternatives to bar graphs of averages. If bar graphs of averages communicate information ineffectively, what is a viable alternative? Compelling arguments have been made for using graphs that display individual data points (Weissgerber et al., 2015, 2019), and such visualizations have shown considerable promise in supporting statistical inference (Zhang et al., 2023) and decision-making (Fernandez et al., 2018). Our work suggests that even the narrowly defined task of communicating the average itself is often better accomplished by raw, individual-level data than by explicit plotting of the average (Kerns & Wilmer, 2021; Wang et al., 2025). Moreover, technological advancements have eliminated barriers to producing visually appealing graphs that include raw, individual-level data (Sidiropoulos et al., 2018). Our experience teaching statistical novices suggests that visualizations of individual values can effectively combat miscommunications of the sort we describe here. Ongoing research is needed to identify the most effective designs for individual-level graphs and to directly compare them with common graph types that use more abstract representations (e.g., bar, line, box, and violin plots).

Conclusion

We conclude that bar graphs of averages fail to communicate data accurately in not just one, but two fundamental ways. In addition to their previously documented failure to convey the one thing they are meant to depict—the average (Kerns & Wilmer, 2021)—we find here that they also, even more frequently, severely misrepresent the shape of the distribution.

Acknowledgments

We thank anonymous reviewers for helpful comments.

References

Angra, A., & Gardner, S. M. (2017). Reflecting on graphs: Attributes of graph choice and construction practices in biology. CBE—*Life Sciences Education*, 16(3), ar53.

Bainbridge, W. A. (2022). A tutorial on capturing mental representations through drawing and crowd-sourced scoring. *Behavior Research Methods*, 54(2), 663-675.

Barton, B. F., & Barton, M. S. (1987). Simplicity in visual representation: A semiotic approach. *Iowa State Journal of Business and Technical Communication*, 1(1), 9-26.

Chen, J. C., Cooper, R. J., McMullen, M. E., & Schriger, D. L. (2017). Graph quality in top medical journals. *Annals of emergency medicine*, 69(4), 453-461.

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., Wang, C. Y., Wang, Y., Li, P., Kini, M., & Liu, Z. (2024). Bar Tip Limit Error and Characteristics of Drawn Data Distributions on Bar Graphs. *In Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

DiMascio, A., Weissman, M. M., Prusoff, B. A., Neu, C., Zwilling, M., & Klerman, G. L. (1979). Differential symptom reduction by drugs and psychotherapy in acute depression. *Archives of General Psychiatry*, 36(13), 1450-1456.

Drummond, G. B., & Vowler, S. L. (2011). Show the data, don't conceal them. *Advances in physiology education*, 35(2), 130-132.

Editors. Kick the bar chart habit. *Nature Methods*, 11(2), 2014.

Editors. Show dots in plots: we encourage our authors to display data points in graphs, and to deposit the data in repositories. *Nature Biomedical Engineering*, 1(79), 2017.

Elliott, M. A., Nothelfer, C., Xiong, C., & Szafir, D. A. (2020). A design space of vision science methods for visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 1117-1127.

Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018, April). Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-12).

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The journal of abnormal and social psychology*, 58(2), 203.

Goodchild, S. (1988). School pupils' understanding of average. *Teaching Statistics*, 10(3), 77-81.

Gray, P. O., & Bjorklund, D. F. (2017). Psychology. Worth, 8 ed.

Grison, S., & Gazzaniga, M. (2019). Psychology. Norton, 3 ed.

Hullman, J., Qiao, X., Correll, M., Kale, A., & Kay, M. (2018). In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics*, 25(1), 903-913.

Kalat, J. W. (2016). Introduction to psychology. Wadsworth, 11 ed.

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), 17-17.

Kim, Y. S., Reinecke, K., & Hullman, J. (2017, May). Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 1375-1386).

Larson–Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics 1. *The Modern Language Journal*, 101(1), 244-270.

Litman, L., & Robinson, J. (2020). Conducting online research on Amazon Mechanical Turk and beyond. Sage Publications.

May, C. P., Hasher, L., & Stoltzfus, E. R. (1993). Optimal time of day and the magnitude of age differences in memory. *Psychological Science*, 4(5), 326-330.

Mogull, S. A., & Stanfield, C. T. (2015, July). Current use of visuals in scientific communication. In 2015 *IEEE international professional communication conference (IPCC)* (pp. 1-6). IEEE.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for research in Mathematics Education*, 26(1), 20-39.

Myers D. G., DeWall, G. N. (2017). Psychology. Worth, 12 ed.

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

Pastore, M., Lionetti, F., & Altoè, G. (2017). When one shape does not fit all: A commentary essay on the use of graphs in psychological research. *Frontiers in psychology*, 8, 1666.

Inhelder B. & Piaget, J. (1951). The genesis of the idea of chance in children. Presses Universitaires de France.

Rahman, Q., Wilson, G. D., & Abrahams, S. (2004). Biosocial factors, sexual orientation and neurocognitive functioning. *Psychoneuroendocrinology*, 29(7), 867-881.

Rohatgi, A. (2017). WebPlotDigitizer. https://automeris.io/WebPlotDigitizer.

Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2017). Beyond differences in means: robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience*, 46(2), 1738-1748.

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.

Tufte, E. R. (1983). The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT.

Vail, A., & Wilkinson, J. (2020). Bang goes the detonator plot!. *Reproduction*, 159(2), E3-E4.

Wang, Y., Kerns, S. H., Brady, T. F., Wilmer, J. B. (2025). The Paradox of Certainty: When Graphed Ensembles Convey Averages Better than Graphed Averages. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47.

Wainer, H. (1984). How to display data badly. *The American Statistician*, 38(2), 137-147.

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.

Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., ... & Milic, N. M. (2019). Reveal, don't conceal: transforming data visualization to improve transparency. *Circulation*, 140(18), 1506-1518.

Zhang, S., Heck, P. R., Meyer, M. N., Chabris, C. F., Goldstein, D. G., & Hofman, J. M. (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences*, 120(33), e2302491120.