The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

Confidence Intervals and the Within-the-Bar Bias

Christopher S. Pentoney & Dale E. Berger

To cite this article: Christopher S. Pentoney & Dale E. Berger (2016) Confidence Intervals and the Within-the-Bar Bias, The American Statistician, 70:2, 215-220, DOI: 10.1080/00031305.2016.1141706

To link to this article: http://dx.doi.org/10.1080/00031305.2016.1141706

Accepted author version posted online: 18 Feb 2016.
Published online: 09 Jun 2016.

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 215–220
http://dx.doi.org/10.1080/00031305.2016.1141706

Taylor & Francis
Taylor & Francis Group

Confidence Intervals and the Within-the-Bar Bias

Christopher S. Pentoney and Dale E. Berger

ABSTRACT

Bar graphs displaying means have been shown to bias interpretations of the underlying distributions: viewers typically report higher likelihoods for values within a bar than outside of a bar. One explanation is that viewer attention is driven by the whole bar, rather than only the edge that provides information about an average. This study explored several approaches to correcting this bias. Bar graphs with 95% confidence intervals were used with different levels of contrast to manipulate attention directed to the bar. Viewers showed less bias when the salience of the bar itself was reduced. Response latencies were lowest and bias was eliminated when participants were presented with only a confidence interval and no bar.

ARTICLE HISTORY
Received January 2015
Revised November 2015

KEYWORDS
Bar graph; Bias; Confidence interval; Graph

1. Introduction

Bar graphs are one of the most widely used methods for communicating data visually. Their simplicity makes them useful for displaying information to a wide variety of audiences in a straightforward manner—an important task in reporting statistics. However, bar graphs may lead to biased interpretations when displaying averages (Newman and Scholl 2012; Correll and Geicher 2014). Newman and Scholl conducted several experiments in which a "within-the-bar" bias was found for bar graphs depicting means. The general method was to present bar graphs and ask how likely each of several points was to be part of the underlying distribution. Viewers of the graphs judged points that fell within a bar as more likely to be included in the underlying distribution than points equidistant from the mean but outside the bar. This bias persisted consistently across six separate experiments, both between and within subjects.

In several of their experiments, Newman and Scholl (2012) presented viewers with bar graphs that had a mean of zero and where zero was exactly the midpoint of the graph. Subjects were asked to rate on a nine-point scale how likely each of several specific values (e.g., −5 or +5) was to belong to the underlying distribution (see Figure 1, where point B would be judged as more likely than point A, on average). Half of the values fell outside the bar and half fell inside the bar, with pairs of values equidistant from the mean value of zero. Subjects viewed either a bar based at the bottom or the top of the graph. The bias was shown for both rising and falling bars. The explanation for the bias suggested by Newman and Scholl is that a bar displays greater density on one side of the mean, giving the false impression of more data on that side.

In bar graphs, the edge of the bar furthest from the base represents the value that is being displayed. However, according to Newman and Scholl (2012), the bar as a whole attracts the attention of the viewer. The asymmetric representation of data on the two sides of the mean results in a tendency for people to judge points within the bar's area as more likely to be part of the underlying distribution. Two solutions for correcting this misperception offered by Newman and Scholl are (a) to provide only a single point to represent each mean, or (b) to present a plot showing every data point in each sample distribution. If the Newman and Scholl explanation is correct, providing a symmetric visual representation leaves no reason for the bias to occur. This interpretation of the within-the-bar bias rests on the idea that the entire bar acts as a single unit of visual attention, rather than just the end of the bar that represents the mean.

Although there is much evidence for the role of a discrete object as a unit of attention (Elder and Zucker 1993; Scholl, Pylyshyn, and Feldman 2001; Barenholtz and Feldman 2003), different qualities can define what an object actually is. The primary feature that contributes to the perception of an object is the connectedness of pieces, or the boundary closure of the object (Kovacs and Julesz 1993; Scholl 2001; Marino and Scholl 2005). Connected lines form boundaries that give the strong impression of a shape against a background. Within a bar graph, the closed form of the bar provides the distinct image of a rectangle. Differences in texture between the bar and background provide contrast, but texture alone may not distinguish figure from ground as strongly as solid outlines. If the Newman and Scholl (2012) explanation is correct, the within-the-bar bias will be reduced if the salience of the entire bar is reduced relative to the salience of just the mean. If line borders are the feature that defines a bar as a visual object, bias should be greater for bars with solid outlines than bars with no outlines, even when the bar has some texture or shading contrast with the graph background.

Newman and Scholl (2012) suggested that a point could be used to display a mean; however, a single point may not be very salient and provides no useful information about variability. A confidence interval is more informative and may avoid the within-the-bar bias due to being symmetric around the mean. The simplicity and familiarity of bar graphs make them useful for communicating information to a variety of audiences,

CONTACT Christopher S. Pentoney *christopher.pentoney@gmail.com* Department of Psychology, Claremont Graduate University, Claremont, CA 91711.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/TAS*.

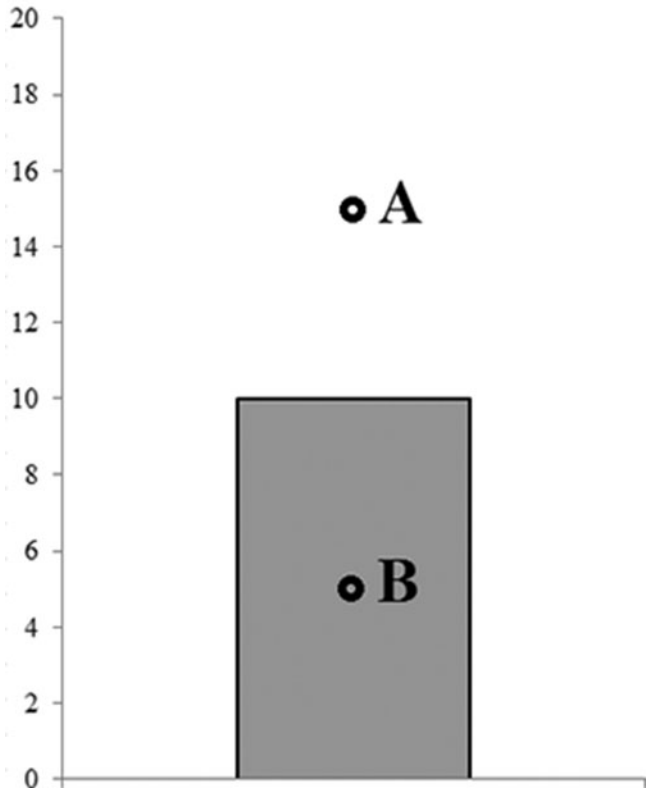

Figure 1. Example bar graph. Points A and B are equidistant from the mean of 10. Adapted from Newman and Scholl (2012) to illustrate the concept of the bias; actual graphs used varied from this example.

and this advantage should not be overlooked. Although a bar graph displayed with a confidence interval provides information about both the sample mean and the variability in the data, a confidence interval alone may not be easy to interpret for novice viewers. Most people do not view confidence intervals very often, so they may be unfamiliar with how to use them. Thus, confidence intervals alone may take longer to process than bar graphs displayed with confidence intervals.

Newman and Scholl (2012) studied the effects of the within-the-bar bias on an applied task. Subjects were told to assume the role of an executive of a tire company that was testing tires for Belt Tensile Strength (BTS, a measure related to tire safety) in which zero was the ideal value. Instructions stated that BTS levels too far above or below zero could result in unsafe tires. Subjects were told that the average BTS was zero for a sample of 30 tires, and subjects were shown a bar graph that was either rising or falling to a value of zero, or no graph at all (control). They were then asked whether they should make changes to the BTS levels. Given that zero was the ideal value, no change was necessary. On average, participants in both bar graph conditions still chose to change BTS levels significantly in comparison to the average control response. For a rising bar with a mean of zero, participants tended to respond that they would like to increase the average BTS such that the bar would extend above the target value of zero, rather than remain centered at zero ($d = 0.28$). For falling bars, participants responded that the BTS levels should be decreased, which would cause the bar to be extended beyond the target of zero ($d = 0.49$). In comparison, the control group (shown no graph) did not choose to make changes significantly different from zero. These results suggest that the within-the-bar bias has implications for real-world decision-making.

The current research tested several approaches to mitigate the within-the-bar bias. Four separate graphs providing differing levels of contrast between a bar and its background were presented. Contrast was manipulated through the presence or absence of solid outlines around the bar and presence or absence of shading within the bar, as shown in Figure 2. Graphs displayed an average of zero so that numeric extremity could not be responsible for biased ratings, analogous to graphs from the last four experiments from Newman and Scholl (2012). Subjects were asked to judge how likely it was that certain values belonged to the distribution represented by each graph. Bias was measured as the difference in likelihood ratings for each pair of corresponding values equidistant on either side of the displayed mean (e.g., the difference between a subject's ratings for the likelihood of −5 and of +5 was used as a measure of bias for that pair).

The first prediction was that if the within-the-bar bias stems from salience of the bar then the bias should be greater for bars with solid outlines than those without. Second, it was predicted that when borders were present, shading presence was not expected to affect viewers' likelihood ratings of points; when borders were absent, absence of shading would remove the bias because the bar was not visible. Specifically, bias would be near zero for graphs with no bar at all (Figure 2(d)), and bias would be significantly less for borderless graphs with shading present

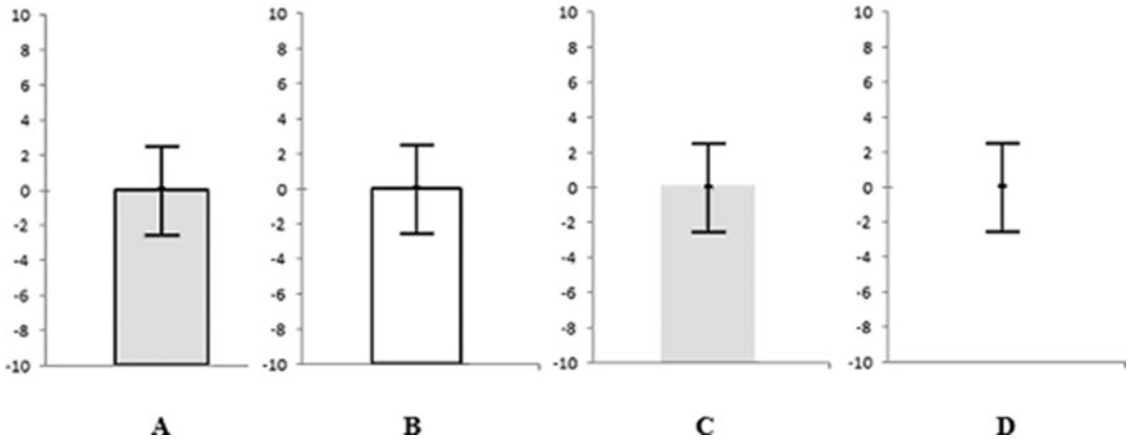

Figure 2. The four graph conditions.

(Figure 2(c)) than for graphs with a border (Figure 2(a) and (b)). Finally, response latencies were predicted to be greater when only a confidence interval was shown compared to when bars were displayed.

2. Method

2.1. Participants and Design

Participants ($N = 192$) were recruited through Amazon Mechanical Turk, and were paid 50 cents ($0.50 USD) for completing the task. Inclusion criteria were that subjects had to be English speakers with an Internet connection from the United States. Gender, experience with statistics, education level, and age were recorded for each participant. Five participants were excluded from analyses due to not completing the task. Two more participants were excluded for having average response times of less than 1 sec per question. This left a total of 185 participants, consisting of 121 males and 64 females. Of these, 42.2% had never taken a statistics course, and 40.5% had taken only one. Most had either a bachelor's degree (45.9%) or some college (32.4%), and 82.3% said that they were currently students. The mean age was 31.9 years (median = 29.0 years; SD = 10.4 years; range = 18 to 68 years). On average, participants spent 6.01 sec answering each question (SD = 3.34).

There were two dependent measures: bias and response latency. All participants viewed all four graphs, with the order of presenting the graphs counterbalanced across participants so that one fourth received each graph first. Participants were assigned in approximately equal proportions to each of the 24 possible orders of presentation for the four graphs. For each graph, participants rated the likelihood of seven values that were presented in random order. Each condition was blocked, and the question order was randomized for each subject within each block.

Performance was expected to improve across trials, so data were analyzed in two ways. Data from the first condition viewed were analyzed in a between-subjects design to assess subjects' naïve ratings of likelihood, and then data from all four conditions for each subject were compared in a within-subjects analysis. A nine-point scale (very unlikely to very likely) was used to assess subjects' estimates of the likelihood of specific values being in a distribution represented by a graph. Likelihood ratings for points outside the bar were subtracted from the ratings for corresponding points inside the bar to produce a measure of the bias.

2.2. Materials

Participants viewed the study online using their own computer. The procedure and stimuli were adapted from Newman and Scholl (2012). An introduction paragraph used a cover story to explain that each graph depicted the average freezing temperature of 30 chemicals tested in a science class. All graphs displayed a mean of zero. Four different graphs with confidence intervals were created by crossing the presence or absence of an outline on the bar with the presence or absence of shading contrast of the bar. Each bar had either a two-point (2pt.) solid black outline or no outline. Bar fill was either solid gray (RGB: 229, 229, 229) with a luminance value of 216 (90% of maximum), or white (luminance of 240% or 100%). When no bar outlines were present and the bar fill was white, only the confidence interval was displayed (Figure 2(d)).

2.3. Procedure

Following a consent form, introduction pages provided the cover story to contextualize the graphs and instructions for the upcoming task. When the participants were ready to begin the study, they clicked a "BEGIN" button that directed their browser to the first graph. Each graph trial began with a graph portraying a mean of zero.

On the same page, participants were asked to use a radio button scale to indicate the likelihood of each specific value by answering the question "What is the likelihood that one of the chemicals used in class had a freezing point of X degrees Fahrenheit?" (adapted from Newman and Scholl 2012). For each question, X was replaced by one of seven values: 0, 3, −3, 5, −5, 7, or −7. Each of these values was presented only once per graph, with the presentation order of values randomized for each respondent. After every response, a "NEXT" button was displayed to allow subjects to move on to the next question. When the button was pressed, the question disappeared, and the next question was displayed in its place. Questions continued with the same graph until all seven values were rated. Every subject viewed all four distinct graphs, each with questions asking about the likelihood of the seven values, for a total of 28 questions. Every trial was presented on the same webpage, and the time from presentation of each question to its submission was recorded as response latency.

3. Analysis and Results

On a five-point scale, subjects on average responded that they were more familiar with standard bar graphs ($M = 3.17$, SD = 1.44) than confidence intervals ($M = 2.60$, SD = 1.31), $t(184) = 4.69$, $p < 0.001$, $d = 0.414$. Responses also showed, however, that people were generally less familiar with confidence intervals combined with bar graphs ($M = 2.05$, SD = 1.10) than confidence intervals alone ($M = 2.60$, SD = 1.31), $t(184) = -6.35$, $p < 0.001$, $d = 0.429$.

Within-the-bar bias was calculated by subtracting subjects' ratings of each point outside the bar from their ratings of corresponding points inside the bar. This measure is an index of asymmetry in a subject's mental representation of data that contributed to the displayed mean. Average bias measures were positive in all bar graph conditions, indicating a bias toward the bar and a replication of the within-the-bar bias effect from Newman and Scholl (2012). To analyze between-subject effects, data from only the first graph a subject viewed were used. A similar number of participants were randomly assigned to view each as their first graph: graphs with borders and shading ($N = 49$), graphs with borders but no shading ($N = 45$), graphs with shading but no border ($N = 43$), and graphs with a confidence interval alone ($N = 48$).

For the first graph viewed by each participant, greatest bias was shown for bars with borders and shading ($M = 1.50$, SE = 0.35, 95% CI [0.80, 2.21]) and bars with borders but no shading ($M = 0.95$, SE = 0.34, 95% CI [0.26, 1.63]). The average bias for shaded bars without borders was somewhat

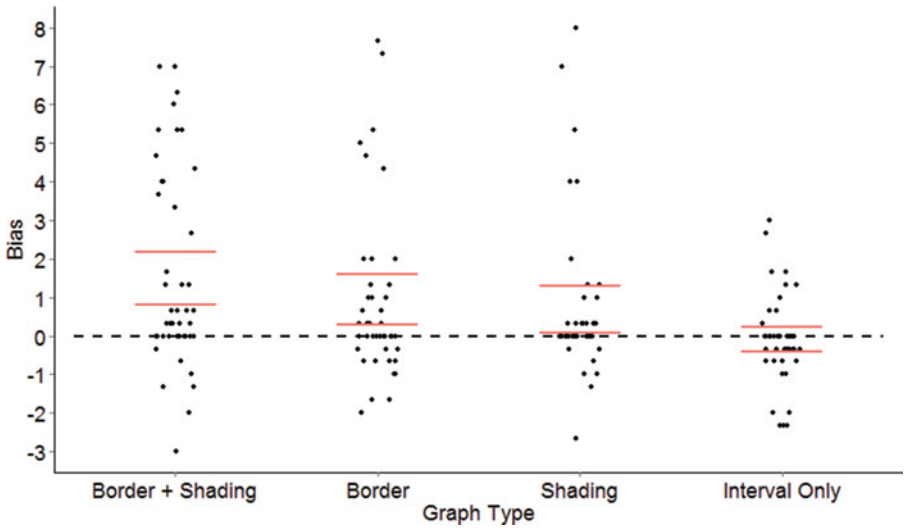

Figure 3. Distributions of bias for between-subjects analysis. A small amount of random noise was introduced across the horizontal axis to better show distribution of points. 95% confidence intervals for the average bias are shown as horizontal bars overlaid on each distribution.

lower ($M = 0.68$, SE = 0.31, 95% CI [0.05, 1.31]), and as predicted, bias was centered near zero when only a confidence interval was displayed ($M = -0.09$, SE = 0.16, 95% CI [−0.41, 0.23]). Pairwise *t*-tests adjusted for unequal variance were used to test the between-subjects differences in bias (see Figure 3). Bias for shaded bars with borders was not significantly greater than for unshaded bars with borders $t(91.99) = 1.13$, $p = 0.258$, but close to significantly greater than for shaded bars with no borders $t(89.77) = 1.75$, $p = 0.08$. Although bias from shaded bars with no borders was descriptively lower than unshaded bars with borders, the difference was not statistically significant $t(85.68) = -0.57$, $p = 0.566$. Furthermore, confidence intervals alone resulted in significantly lower bias ratings than shaded bars with borders $t(66.53) = 4.14$, $p < 0.001$, unshaded bars with borders $t(62.27) = 2.78$, $p = 0.007$, and even shaded bars without borders $t(62.50) = 2.21$, $p = 0.031$.

For the within-subjects analysis, bias was found for bars with borders whether they were shaded ($M = 1.35$, SE = 0.18, 95% CI [1.00, 1.70]) or unshaded ($M = 1.13$, SE = 0.16, 95% CI [0.81, 1.46]), and for shaded bars without borders ($M = 0.85$, SE = 0.16, 95% CI [0.54, 1.16]). However, average bias was near zero when bars were not visible ($M = 0.06$, SE = 0.10, 95% CI [−0.14, 0.26]). Greatest bias was produced by the graph with borders and shading. Although bias for the graph with borders and no shading was nearly as great, $t(180) = 1.81$, $p = 0.07$, bias was significantly lower for shaded graphs without borders, $t(180) = 4.32$, $p < 0.001$. Additionally, borderless graphs with shading had lower bias than bordered graphs without shading $t(180) = 2.61$, $p < 0.001$. Again, the graph with only the confidence interval had significantly less bias than bordered bars with shading $t(184) = -7.31$, $p < 0.001$, bordered bars without shading $t(184) = -6.48$, $p < 0.001$, and borderless bars with shading $t(184) = -5.28$, $p < 0.001$.

The distributions of response latencies for the first graph viewed by each subject were positively skewed due to the presence of several outliers, so a nonparametric Kruskal–Wallis test was used to compare response latencies for the four groups. The difference in response latencies as a function of the graph type attained overall statistical significance $H(3) = 10.18$, $p = 0.017$. Counter to predictions, response latencies for confidence intervals alone were smaller than the other three graphs pooled ($U = 2326.5$, $p = 0.003$). Median response latencies were similar for bordered graphs with shading (median = 8.14), bordered graphs without shading (median = 8.86), and shaded graphs without border (median = 8.29), but were smaller for confidence intervals alone (median = 6.71). Mann–Whitney *U* tests were not significant for any pairwise test comparing conditions with a bar graph. See Figure 4 for the distributions of response latencies.

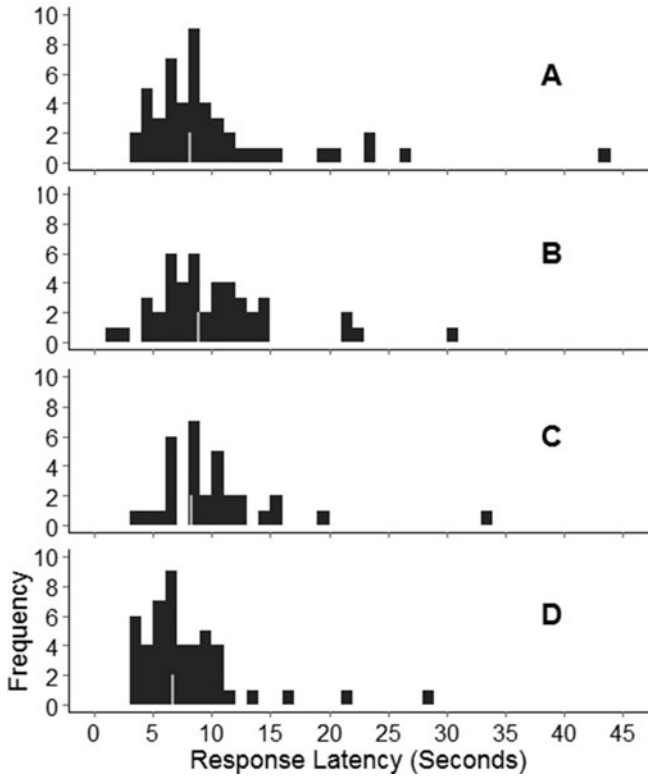

Figure 4. Distributions of response latencies for the first graph viewed. Each histogram represents a distribution of response latencies for the corresponding graph: Histogram A represents bar graphs with borders and shading, B represents graphs with borders and no shading, C represents graphs with shading and no borders, and D represents confidence intervals only. Median response latency is shown as a lightly shaded line.

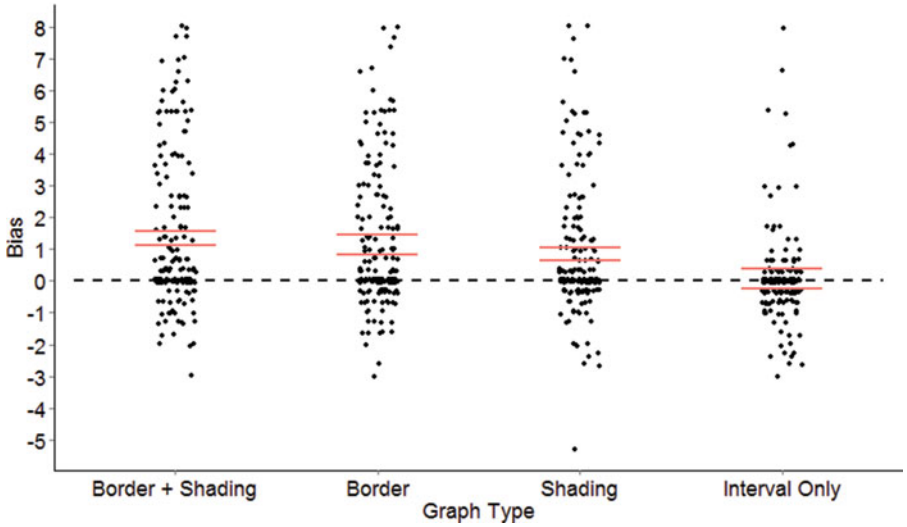

Figure 5. Distributions of bias for within-subjects analysis. A small amount of random noise was introduced across the horizontal axis to better show distribution of points. 95% confidence intervals for the average bias are shown as horizontal bars overlaid on each distribution.

4. Discussion

Two of the three hypotheses were supported. First, the presence of borders resulted in a more biased understanding of the distribution of data used to compute the average. Shading also caused biased ratings, although it did not have as strong of an effect as border presence. Borders and shading are both properties that draw attention of the viewer away from the mean and toward the bar itself. The effect was apparent especially for bars that had borders, which was consistent with the concept that closed shapes are more readily attended to than shapes formed by differences in texture. The pattern of averages for the between-subjects analysis was similar to the pattern found in the within-subjects analysis (compare Figure 3 to Figure 5). These results followed the predicted pattern: bias for the confidence interval alone was centered near zero. One of the most interesting findings was that simply removing the border reduced average bias. In both analyses, the average bias for confidence intervals presented alone was near zero, indicating that the presence of the bar is responsible for the bias.

Within-subjects analyses also showed that the different values being rated were susceptible to differing amounts of bias. Overall, values that were closer to zero had greater bias in likelihood ratings, while extreme values resulted in a smaller bias. This is likely due to a ceiling effect, because many people simply used the extreme end of the scale to rate all extreme points. That is, participants rated 60% of the extreme values (−7 and 7) as a one (very unlikely) on the nine-point scale, limiting the sensitivity of the measure of bias.

The final hypothesis was that responses would be slower for the confidence interval alone than for the three bar graphs. This hypothesis was not supported. In fact, responses were slightly faster when the confidence interval was viewed alone than when any of the three bar graphs were viewed. The reasoning for the original hypothesis was that response latencies would be greater for confidence intervals because familiarity would be lower for confidence intervals compared to the bar graphs. However, confidence intervals alone were rated as more familiar than the bar graphs with confidence intervals. While response latencies were smaller for confidence intervals alone than the other graphs, these time differences might still be due to unfamiliarity because the combinations of confidence intervals and bar graphs were considered to be relatively unfamiliar. Although familiarity was generally low for confidence intervals, responses were less biased and significantly faster with confidence intervals alone than with bar graphs and confidence intervals combined.

These findings provide guidance for best-practices in graphing data. The most important implications are those regarding the use of confidence intervals. An interesting conclusion is that confidence intervals can be informative for nonresearchers, even if only for displaying means. While the exact interpretation of the confidence intervals may not be apparent, they provide a less biased impression of the underlying data compared to bar graphs. Further, confidence intervals presented alone are interpreted more quickly than bar graphs with confidence intervals. Bar graphs with confidence intervals or standard error bars are often displayed by researchers, but the bars add no information beyond the intervals, and only draw attention away from the intended message. Bias is greater for bar graphs displayed with salient borders than for bar graphs displayed without borders.

Two limitations should be noted. First, subjects were recruited online from MTurk, and there was no way to measure how variables such as screen size, screen distance, and real-world distractions may have impacted responses. Nonetheless, the results are useful for an understanding of how different graphs are interpreted on the Internet, at the very least. Another limitation is that all stimuli displayed confidence intervals, so the comparison was really between confidence intervals and bar graphs containing confidence intervals, not standard bar graphs alone. The argument for this comparison is that standard bar graphs give no indication of variability, and would simply contain less information than the stimuli used here. Although bar graphs are used often with confidence intervals, results from this study suggest that it is better to use confidence intervals alone. Confidence intervals avoid the within-the-bar bias, yet still provide rapid assessment of averages.

References

Barenholtz, E., and Feldman, J. (2003), "Perceptual Comparisons Within and Between Object Parts: Evidence for a Single-object Superiority Effect," *Vision Research*, 43, 1655–1666. [215]

Correll, M., and Gleicher, M. (2014), "Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error," *IEEE Transactions on Visualization and Computer Graphics*, 20, 2142–2151. [215]

Elder, J. H., and Zucker, S. W. (1993), "The Effect of Contour Closure on the Rapid Discrimination of Two-Dimensional Shapes," *Vision Research*, 33, 981–991. [215]

Kovacs, I., and Julesz, B. (1993), "A Closed Curve is Much More Than an Incomplete One: Effect of Closure in Figure–Ground Discrimination," *Proceedings of the National Academy of Sciences*, 90, 7495–7497. [215]

Marino, A. C., and Scholl, B. J. (2005), "The Role of Closure in Defining the "Objects" of Object-Based Attention," *Perception and Psychophysics*, 67, 1140–1149. [215]

Newman, G. E., and Scholl, B. J. (2012), "Bar Graphs Depicting Averages are Perceptually Misinterpreted: The Within-the-Bar Bias," *Psychonomic Bulletin and Review*, 19, 601–607. [215,216,217]

Scholl, B. J. (2001), "Objects and Attention: The State of the Art," *Cognition*, 80, 1–46. [215]

Scholl, B. J., Pylyshyn, Z. W., and Feldman, J. (2001), "What is a Visual Object? Evidence From Target Merging in Multiple-Object Tracking," *Cognition*, 80, 159–177. [215]