Bar Tip Limit Error and Characteristics of Drawn Data Distributions on Bar Graphs

Lucy Cui[1] (cuil@rpi.edu)
Ching-Yi Wang[2] (ian0504@ucla.edu)
Yiwei Wang[2] (yiweiwang2001@outlook.com)
Peike Li[2] (peike239@ucla.edu)
Medha Kini[3] (medhakini@ucla.edu)
Zili Liu[2] (zili@psych.ucla.edu)
[1]Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY 12180
[2]Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095
[3]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90095

Abstract

Bar graphs are commonly used graphs, but what do students infer about the data that created the bar graph? Previously, a drawing task revealed that a minority of students conflate mean bar graphs with count bar graphs and draw all data points within the bar of a mean bar graph (*bar tip limit error*, BTLE). The present study extends this literature by manipulating the instructional text for the drawing task, interviewing the participants on their drawings, and recording their drawings and drawing session for further analysis. While we did not see any differences in the BLTE rates across instructional conditions, we did see significant differences in their drawing explanations and drawn data distributions based on condition, and in their drawing explanations based on whether they expressed confusion and whether they committed the BTLE. We discuss possible explanations and their implications.

Keywords: bar graphs, statistics, education, graph perception, data visualization

Introduction

Bar graphs are commonly used in academia as well as many other real-world contexts, such as education, media and business. Despite its ubiquity, bar graphs can be often misinterpreted. Some of these misinterpretations may be perceptual in nature (see Cui & Liu, 2021 for a review). For example, the values in bar graphs are often underestimated. One explanation of this result is that people use spatial location instead of length of bars to determine the value of bar graphs (Yuan, Haroz, & Franceroni, 2019). Other misinterpretations may be due to misconceptions about statistics or oversimplification of relationships between graphical elements and their statistical meaning. For example, people believe data within error bars, which typically depict standard error, are more likely than data outside the error bars (Newman & Scholl, 2012), all values within error bars are equally likely (Ibrekk & Morgan, 1987), and overlapping error bars reveal nonsignificance (Cumming, 2009). Even experts have misconceptions about how error bars relate to statistical significance and the difference between error bars and confidence intervals (Belia, Fidler, Williams, & Cumming, 2005).

One of the commonly seen errors students make with bar graphs regards the relationship between the data points that created the bar graph and the bar graph depiction. Newman and Scholl (2012) coined the term *within-bar bias* to refer to the tendency for people to think data points within the bar in a bar graph are more likely to be part of the data set than data points outside the bar. Newman and Scholl (2012), along with other researchers following them (Correll & Gleicher, 2014, Okan, Garcia-Retamero, Cokely, & Maldonado, 2018, Pentoney & Berger, 2016), used a probability rating scale, which asked participants how likely a given data point is given the bar graph, in order to determine this *within-bar bias*. Another approach to assess for *within-bar bias* is the balls-and-bins approach, which asked participants to produce a "histogram-like" data distribution based on a bar graph (Goldstein & Rothschild, 2014, Kim, Walls, Kraft & Hullman, 2019, Andre, 2016, Hullman et al. 2018). These approaches categorized the bar graph misinterpretation as a bias - that everyone is biased in the direction of data points within a bar being more likely data points.

However, when a different approach was used, that is one that allows for participants to more freely respond, a different interpretation of previous results arose. Kerns and Wilmer (2021) created the Draw Datapoints on Graphs (DDoG) measure, where participants draw data points on the graphs to show their understanding of data and the graph. These drawings revealed that what was once thought of as a bias (everyone exhibits) is actually an error that only a small minority of people commit. Only 20.6-27% of participants produced this error (from the data of Kerns & Wilmer, 2021, Pentoney & Berger, 2016, Newman & Scholl, 2012). Those participants conflate mean bar graphs with count bar graphs and draw all (most) of the dots within the bar, which the authors coined the *bar tip limit error* (see Figure 1). Kerns and Wilmer (2021) defined *bar tip limit error* as having a bar-limit tip index: (# of data points drawn within the bar / # of data points) x 100 be over 80.

Kerns and Wilmer (2021) asked participants to draw data points that "could be averaged to get the value shown by the bar". The resulting drawings from participants showed regular and symmetric dots across the bar edge and few possible y values, suggesting that participants drew for the task and not for what they believed about data distributions.

We extend the previous work by looking at whether instructional changes influence students' drawing behavior and what characteristics they draw into their data distributions. In our experimental conditions, we activate

increasing levels of statistical concepts to see whether considerations of these statistical concepts change the data distributions they draw on the bar graphs.

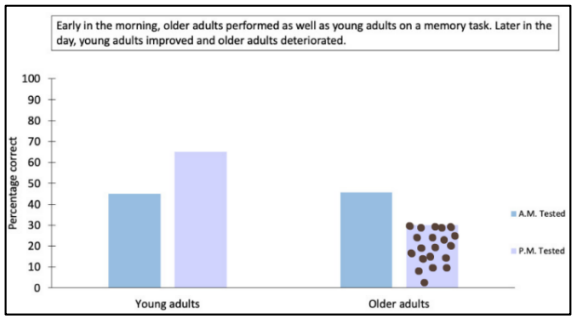

Figure 1: Example of *Bar Tip Limit Error*

We hypothesize that changing the instructions for the task in include varying levels of statistical concepts would change the occurrence of the *bar tip limit error*, drawing behavior, and the types of data distributions drawn.

In the present study, participants were either asked to draw data points onto a bar graph, with consideration to the naturally occurring variance in data, or with additional consideration to the y-axis. We screen-recorded their drawing session and saved their drawings. Research assistants briefly interviewed them on their drawing. We compared the frequency of *bar tip limit error* across conditions, qualitatively coded for their explanation for their drawings and their drawing behavior from the screen-recorded video, and statistically analyzed their drawn data distributions.

Methods

Participants

Participants were 148 undergraduates (97 Female, 29 Male, 22 no response, M_{age}=21.90, SD_{age}=6.28 from the University of California, Los Angeles. They participated for partial course credit. Most students have taken some statistics class(es) (M = 2.03 classes, SD = 1.07). Only two participants have not taken a statistics class before.

Materials

We used an Amazon Fire HD 10 tablet and a compatible stylus for our in-person experiment. We used Nearpod to administer the experiment and collect the drawings and responses. We selected one graph (i.e., the old vs. young adults on memory task) from Kerns & Wilmer (2021) to use due to its simplicity (i.e. 2 x 2 design, or 4 bars) and its values being all positive. This graph was provided as a reference image, on which participants can draw their 20 data points. At the top of the screen, participants were given the instructions for their drawing task (see Figure 2 for an example condition).

There were three conditions, each had different text at the top of the screen (see Table 1). For reference, Kerns & Wilmer (2021) had as their instructions, "draw 20 dots that show possible individual values that could be averaged to get the value shown by the bar". The resulting drawings showed some evidence of demand characteristics, that is drawing to the task by drawing regular and symmetric dots across the bar edge. We chose to reword the instructions from Kerns & Wilmer (2021) to be vaguer and more open to interpretation in order to avoid participants primarily drawing to the task.

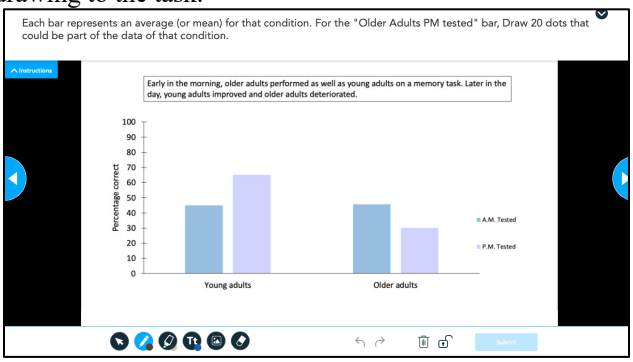

Figure 2: Example of screen for drawing task. Participants got instructions on the top. The graph had its own caption (in the black box) as well as a legend (AM tested, PM tested). At the bottom of the screen, participants can choose their pen or eraser.

Table 1. Instructions for the three conditions.

Condition	Instructions
Vague	"Draw 20 dots that could be part of the data of that condition"
+ Variance	"Draw 20 dots that could be part of the data of that condition, keeping in mind the average shown by the bar's height AND the amount of variation that could exist in data sets."
+ Y-axis	"The y-axis represents all possible values for this task. Draw 20 dots that could be part of the data of that condition, keeping in mind the average shown by the bar's height AND the amount of variation that could exist in data sets."

Procedure

Participants completed the experiment on a digital tablet with a stylus. Participants were randomly assigned into one of three instructions conditions: vague, variance, y-axis. First, they had an opportunity to play around with the drawing interface, such as drawing dots and erasing them. Next, they were presented the bar graph and its accompanying instructions for their condition (see Figure 2). After they were done drawing, the research assistant asked the participant to explain how they drew the dots on the bar graph. After this interview portion, the participants filled out their demographics information, such as age, gender and number of statistics classes taken before.

Data Processing

Image processing was constrained to the bar-of-interest for the task (i.e., the rightmost purple bar in Figure 2). We used the FindContours function in the OpenCV Computer Vision library for object (dot) detection. The original image was inverted into binary coding (black dots on white background). Once the contour (i.e., dot) is detected, we used the center of the contour as the coordinate for the point. Those points were scaled to be meaningful in the following way:

(1): y-coordinates were scaled to match what they would be on the y-axis of the graph, extrapolating if the drawn dot is under 0 or over 100 from the y-axis.

(2) x-coordinates were based on the left-boundary of the bar, that is 0 is a point on the left-edge and values increase as you move right in the bar, with any dots drawn to the left of that boundary being negative. The x-coordinates were scaled based on how many dots fit across the bar, which turned out to be 13 (each dot averaged to be 6 pixels in diameter). Therefore, dots with x-coordinates between 0 and 13 can be interpreted as drawn within the confines of the bar, whereas x-coordinates below 0 and above 13 are drawn outside the confines of the bar.

Once we collected the coordinates, we calculated the bar tip limit index (# of data points drawn within the bar / # of data points) x 100) as well as the following statistics for the dots drawn: mean, median, minimum value, maximum value, range, standard deviation and skewness. For reference, the bar edge has a y-coordinate of 29.54 so all dots drawn below that coordinate were considered within the bar.

Qualitative Coding

Interview Notes After the participant left, the research assistant recorded whether the participant committed *bar tip limit error (BTLE)*, which we defined for coding simplicity as drawing all the dots within the bar. Actual BTLE index calculated and discussed later in the paper.

We qualitatively coded the interviewer's (research assistant) notes from the participants' explanations of their drawings based on the grounded theory approach with no a priori categories. In other words, the primary coder made categories as they read through the responses (see Table 1 for codebook). The secondary coder used the categories description from the primary coder to code a random 30 responses. We used percentage agreement (see Table 3) for interrater reliability because for many of the codes, there were more zeros than ones.

Table 1: Codebook for participants' explanation of their drawings. Binary coded (0 or 1).

Code	Description
Even	creating an equal distribution by placing an equal number of dots above and below the average line of the bar and trying to balance the data around the average
Average	trying to ensure that the overall distribution centers around the average and distributes dots around the average line
Normal	trying to mirror the dots above and below the average line in their distribution of dots. The drawing follows a normal distribution.
Outliers	drawing most of their dots around or below the average line but adding a few outliers that are farther away above the bar.
Random	not reporting a clear strategy for drawing the dots or having randomly distributed the dots across all bars.
Edge	drawing dots directly on the bar's edge or tracing the average line.
Confused	express confusion or lack of understanding to the task
Hesitated	hesitation at the beginning or drawing (uncertainty or delay in starting)
Wrong Bar	did not draw on the target bar, but drew on a different bar or on all bars
Skew	drew noticeably more dots within the bar or outside of the bar skewing the distribution

Videos We qualitatively coded the drawing videos based on a priori categories. Two independent coders used the following codebook in their coding and overlapped in 30 responses. Percentage agreement (see Table 4). was used again for interrater reliability.

Table 2: Codebook for videos of participants' drawing session. Binary coded (0 or 1).

Code	Description
Alternate Mirror	draws one dot within bar and a symmetric dot outside bar and alternate on this behavior
Block Mirror	draws dots either within or outside the bar all at once and then mirrors what is drawn on the other side of the bar
Span Over	draws dots from one direction: top-down or bottom-up
Random	draws dots randomly over the space within and above bar with no perceived order

Results

First, we tested whether our instructional conditions affected the rate of the *bar tip limit error* (BTLE). We conducted a chi-square test to assess the relationship between instructional conditions and BTLE rates and found no reliable relationship, $\chi^2(2) = 2.07$, $p = .35$. In other words, activating different statistical concepts did not reduce the changes a student made the BTLE. The BTLE rates for each of the conditions: 15.69% (vague), 21.57% (+ variance), 27.91% (+y-axis) were also consistent with the BTLE rates found in previous data sets: 20.6-27%.

We continued our data analysis with qualitative data that we have coded and used some of these codes to further explore the BTLE rates.

Drawing Explanations

Next, we tested whether our instructional conditions affected any of our codes for participants' explanation of their drawings (see Table 1 for details) and the codes for their video-captured drawing strategies (see Table 2 for details).

We conducted a chi-square test assessing the relationship between the instructional conditions and each of the codes in Table 3. We only found a significant relationship between instructional conditions and the participant expressing some amount of confusion (Confused code), $\chi^2(2) = 9.81$, $p = .007$. Participants expressed some confusion in the + variance condition the most (41.5%), vague (27.5%), and then + y-axis (12.90%). However, it appears that this confusion did not translate into any difference in BTLE rates from the primary data analysis. We also found a significant relationship between instructional conditions and the focus on recreating an average (Average code), $\chi^2(2) = 6.69$, $p = .035$. Participants focused on the average the most in the + y-axis condition (64.50%), then the + variance (49.00%) condition, and then the vague condition (35.30%). The Confused code and the Average code were amongst the most prevalent aspects of participants' explanations of their drawing (see Table 3). Chi-squares with other codes were nonsignificant.

Overall, the participants seemed to be focused on recreating the average (47.37%) the most and the drawing dots in a random pattern (21.80%) with outliers (14.79%) and an even distribution (aka symmetric) of dots on each side of the bar edge. The topic of skew only came up in a minority of participants' explanations (5.97%). It is also reassuring to see that majority of the participants (92.54%) drew the dots on the Correct Bar compared to the 7.46% that drew on multiple bars.

Table 3: Participants' explanation of their drawings. Percentage (%) is of 1's. Reliability is % agreement between two independent coders.

Code	%	Reliability
Even	12.78	0.97
Average	**47.37**	0.93
Normal	8.27	0.93
Outliers	15.79	0.87
Random	**21.80**	0.87
Edge	5.26	0.90
Confused	**30.83**	0.93
Hesitated	5.97	0.97
Wrong Bar	7.46	0.97
Skew	5.97	0.93

Next, we were interested in whether some of the above codes were related to whether a participant committed a BTLE or not. We conducted a chi-square test assessing the relationship between the Confused code and BTLE rates and found no reliable relationship, $\chi^2(1) = 2.21$, $p = .14$. We also conducted a chi-square test between the Hesitated code and the BTLE rates and found no reliable relationship, $\chi^2(1) = 1.35$, $p = .25$. This suggests that whether a participant was confused or hesitated to start drawing is not a good indicator as to whether they would commit the BTLE. However, they could be predictive of other characteristics of their drawing. Because the Confused code was so prevalent, we will primarily compare the Confused code and the BTLE error as splitting variables in the next few analyses.

We looked at whether confused participants had different explanations than their non-confused counterparts. There was a significant relationship between confusion (Confused code) and the focus on recreating the average (Average code), $\chi^2(1) = 7.79$, $p = .005$. Confused individuals were less likely to focus on recreating the average (29.30%) than non-confused individuals (55.40%). Additionally, we found a marginally significant relationship between confusion (Confused code) and whether the participant drew dots randomly (Random code), $\chi^2(2) = 3.41$, $p = .065$. Confused individuals were more likely to draw dots randomly (31.7%) than those who were less confused (17.4%).

We looked at whether participants who committed the BTLE had different explanations than the participants who did not commit the BTLE. There was a significant relationship between BTLE and focusing on recreating the average (Average code), $\chi^2(2) = 12.40$, $p < .001$. Those who committed the BTLE were less likely to focus on recreating the average (17.90%) than those who did not commit the BTLE (55.3%). There was a marginally significant relationship between BTLE and create a normal distribution with the bar edge as a mean (Normal code), $\chi^2(2) = 2.94$, $p = .086$, such that those who did not commit the BTLE were more likely to draw a normal distribution (9.7%) than those who committed the BTLE (0%). Finally, there was a marginally significant relationship between BTLE and drawing dots on the wrong bar(s) (Wrong Bar code), $\chi^2(2) = 2.91$, $p = .088$, with those who did not commit the BTLE more likely to draw on the wrong bar(s) (9.6%) than those who did commit the BTLE (0%), suggesting that those that committed the BTLE had at least read the instructions carefully enough to draw on the target bar.

Drawing Strategies

We analyzed the screen-recorded videos of participants' drawing sessions using the codes outlined in Table 2. We investigated whether there was a relationship between these codes and instructional conditions, the Confused code, and the BTLE. Unfortunately, we did not find any differences between instructional conditions, between confused vs. non-confused individuals and between those who did and did not commit the BTLE on any of the video codes. In other words, the strategies that participants used to draw their dots did not differ based on their instructions, on whether they were confused, or on whether they committed the BTLE.

Table 4 shows the overall results of the video codes, collapsing across conditions. The most common strategy used was drawing the dots randomly (48.46%) and while our two independent coders may have interpreted the word "random" differently, our reliability score is still pretty good

at a percentage agreement of 80%. The other strategies were very unpopular, suggesting that participants were not being super meticulous about ensuring the data points they drew created the average and that the distribution is perfectly symmetrical. Our codes do not capture all of the different strategies that could have existed but creating *ad hoc* categories of videos is difficult.

Table 4: Drawing behavior from videos of participants' drawing session. Percentage (%) is of 1's and Reliability is the percentage agreement between two independent coders.

Code	%	Reliability
Alternate Mirror	1.54	1.00
Block Mirror	3.08	0.96
Span Over	7.69	0.92
Random	**48.46**	0.80

Drawn Data Distributions

Next, we looked at whether there are any differences in the statistical characteristics of the data distributions drawn. We compared mean, median, minimum value, maximum value, range, standard deviation and skewness of the drawn data distributions based on instructional conditions.

We conducted a one-way ANOVA on each of the statistical characteristics using instructional conditions as the independent variable. There was a significant relationship between instructional condition and the range of values participants drew, *Welch's test* (unequal variances): $F(2, 88.1) = 3.43$, $p = .037$. The range of values drawn increased as more statistical concepts were activated: vague condition ($M = 43.50$, $SD = 16.30$), variance condition ($M = 50.00$, $SD = 23.10$), and y-axis condition ($M = 54.00$, $SD = 23.90$). The standard deviations for each condition show that in each condition, the ranges of values drawn varied greatly. Post-hocs using Games-Howell (unequal variances) correction revealed that there was a significant difference (10.54) in range of values drawn between the vague condition and the y-axis condition, $t(72) = 2.45$, $p = .043$.

There was a marginally significant relationship between instructional conditions and the maximum value of the y drawn (y max), *Welch's test*: $F(2, 88.1) = 2.50$, $p = .088$. The y-axis condition had the highest y-max ($M = 60.7$, $SD = 25.40$), then the variance condition ($M = 52.00$, $SD = 16.80$) and then the vague condition ($M = 52.00$, $SD = 16.80$). This order makes sense as the y-axis condition drew participants attention to the y-axis and all possible values that the y variable could take. However, post-hocs using Games-Howell correction revealed no significant comparisons.

All statistical characteristics were significantly different when independent-samples t-tests were run between participants who committed the BTLE and those who did not, $p < .05$ (descriptive statistics for each feature can be found in Table 5). No other statistical characteristics of the drawn data distributions were different across the conditions.

Since the bar tip is at 29.54, it is no surprise that those who committed the BTLE had a y mean around the midpoint of the bar (~15), a y max below 29.54, and smaller standard deviations for most characteristics (due to constraining dots within bar) compared to the participants who did not commit a BTLE. In addition, we observe that most participants had high BLT indexes (i.e., close to 100, or all of the dots within the bar), whereas those who did not commit the BTLE had roughly half of the dots within the bar (BTLE index of 46.30). While subtle, a very surprising difference based on BTLE error is skewness, as those who did not commit the BTLE error were drawing, on average, a positively skewed data distribution while participants who committed the BTLE were drawing symmetric distributions on average. This suggests that those who did not commit the BTLE may not assume normality in data distributions or believe extreme outliers (e.g., close to perfect scores) are highly probable in data sets. In fact, BLT index was negatively correlated with skewness, $r(143) = -0.13$, $p = .12$, such that the lower the BLT index (proportion of dots drawn within the bar), the more positively skewed the data distribution was. Similarly, BLT index was strongly negatively correlated with deviation (how far from the actual mean the drawn data distribution was), $r(143) = -0.93$, $p < .001$.

Table 5: Descriptive statistics of statistical characteristics of drawn data distributions between those who committed BTLE and those who did not. Means shown in bold and standard deviations in parentheses.

Characteristic	No BTLE	BTLE
BTLE index	**46.30** (17.30)	**98.00** (4.45)
Y mean	**33.60** (8.86)	**15.70** (1.90)
Y median	**32.40** (9.18)	**15.50** (2.73)
deviation from real mean	**4.11** (8.86)	**-13.80** (1.90)
Y min	**8.93** (8.29)	**4.73** (2.56)
Y max	**65.20** (16.90)	**26.70** (3.95)
Range	**56.20** (18.10)	**22.00** (6.02)
Skewness	**0.27** (0.40)	**0.008** (0.378)

Table 6: Proportion of drawings that have means over, similar to and below the actual mean of the bar. Similar is defined as being 2 dot widths within the actual mean.

Drawn Mean is:	No BTLE	BTLE
Over	56.25%	0%
Similar	23.21%	0%
Under	20.54%	100%

To investigate this further, we divided the drawings into three categories, having means that are over the bar mean, similar to the bar mean and under the bar mean (see Table 6 for the breakdown). Means that were 2 dot widths (10 pixels) within the bar edge (29.54) were considered "similar" to the actual bar mean. As you can see from Table 6, majority of the data distributions drawn by those who did not commit the BTLE had averages over the bar mean.

Statistics Courses Taken

We investigated whether the number of statistics courses taken prior to the experience could explain the BTLE or the

drawing differences. An independent-samples t-test revealed no difference in the number of statistics classes taken between those who committed the BTLE error ($M = 1.89$, $SD = 0.99$) and those who did not ($M = 1.97$, $SD = 1.10$). Furthermore, the number of statistics classes taken did not correlate with how close the drawer was at recreating the average (deviation: drawn – actual), $r(96) = -0.05$, $p = .65$. However, there was a marginally significant positive correlation between number of statistics classes taken and the number of unique y values drawn, $r(96) = 0.18$, $p = .08$. In other words, with more statistics education, students understand that there are more possible y values in their data set. The lower unique y values could also reflect tracing the edge of the bar or only drawing a few data points and only closer to the bar.

Discussion

The purpose of this study was to see whether students can spontaneously realize where data points can fall on a bar graph if given enough statistical cues and clues before their drawing process and whether this cuing reduces the *bar tip limit error* (BTLE) rates. For example, drawing attention to the y-axis was meant to have them realize that data points can exist above the bar tip. We manipulated instructional text to have students think about statistical concepts like mean, variance and y-axis, ask participants to explain their drawings, recorded their drawing and drawing session, and analyzed their drawings both quantitatively and qualitatively.

Increasing statistical concept activation (our conditions) did not produce statistically different rates of the BTLE. BTLE seems to be persistent and not something that can be corrected through spontaneous realization by the student. Intervention from an instructor seems to be necessary. What is more surprising is that whether a participant mentioned any confusion during their interview with the research assistant did not predict whether they produced the BTLE, suggesting that those participants may be confident in their responses or oblivious to how data is represented in the bar graph. Furthermore, looking at the Wrong Bar code, we see that all of the participants who committed the BTLE drew on the target bar only (instead of the wrong bars), suggesting that they did read the instructions carefully. Number of statistics courses taken previously also did not predict whether participants committed the BTLE, which is consistent with previous literature showing that expertise does not necessarily reduce graph interpretation errors.

Instructional text did significantly influence participants' drawn data distributions, with the maximum y value drawn and range of y values drawn increasing in order of included statistical concepts in the instructions: vague, variance, y-axis. This makes sense as when one is probed to think about variance, one would draw more variable y-values, with many opportunities to draw upward. When one is probed to think about the y-axis, one would also draw more y-values higher.

Majority (76.79%) of participants did not draw a dot distribution that had a similar mean as the bar mean. This could be due to limitations in people's perceptual averaging abilities and readjusting tendencies during a generative process– some of our human-coded data, not reported here, show an 87.5% agreement between human and computer on determining whether the dot mean was over, similar to, or under the bar mean. On the other hand, perceptual averaging abilities related to viewing graphs tend to be quite good (e.g., with scatterplots: trend judgment, Ciccione, Sablé-Meyer, Boissin, Josserand, Potier-Watkins, Caparos, & Dehaene, 2023; barycenter/mean position, Hong, Witt, & Szafir, 2021). There also was tendency for people to draw a data distribution with a mean higher than the bar mean if the BTLE was not committed.

We further found that drawn data distributions from those who did not commit the BTLE were, on average, positively skewed, whereas those who did commit the BTLE drew, on average, symmetric dot distributions. This suggests that the former group may be drawing some outliers above the bar tip, which is consistent with the Outlier code we produced. The average maximum y value drawn in this group is also over twice as much as the bar tip (65.20 vs. 29.54), which suggest that most of this former group believe that naturally occurring data has outliers and more specifically, some individuals who perform much better than others (see Figure for context). This drawing behavior could also reflect a tendency to not assume normality in data distributions, which could be relate to a tendency to believe data points within a bar in a bar graph are more likely than outside the bar (*within-bar bias*).

While the number of statistics classes taken did not predict committing the BTLE error, those who have taken more statistics classes were more likely to draw more unique y values, suggesting that statistics education helps students understand and imagine more possible data values.

Our results and its interpretation are limited to the quality of the notes that the research assistants took from interviewing the participants as well as the self-awareness and reporting accuracy from the participants. Additionally, one needs to remember and notice patterns in drawing behaviors in order create *ad hoc* categories of those videos, which is harder to do than with text, so we may not be able to draw more information from the screen-recorded videos. Finally, because the BTLE only occurs in 20-30% of participants, our null effect could reflect an underpowered sample or sampling error. The low prevalence also makes it difficult to recruit enough participants to run follow-up intervention studies on this population.

Our study reveals that the BTLE is a persistent problem and not one that can be resolved without direct instruction. Future studies should investigate strategies to reduce or eliminate the BTLE in students and improve students' understanding of the relationship between data and bar graphs. Our study also starts a conversation of what students believe is true about naturally occurring data sets, in terms of its variance and skewness. Future studies could look into whether students think most data sets are positively skewed, whether it is specific to data sets on performance, and whether belief about skewness exists for other graphs that display means or distributions.

References

Andre, Q. (2016). distBuilder, doi:10.5281/zenodo. 166736. Retrieved from https://quentinandre.github.io/DistributionBuilder/.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers Misunderstand Confidence Intervals and Standard Error Bars. *Psychological Methods, 10*(4), 389–396. https://doi.org/10.1037/1082-989X.10.4.389

Ciccione, L., Sablé-Meyer, M., Boissin, E., Josserand, M., Potier-Watkins, C., Caparos, S., & Dehaene, S. (2023). Trend judgment as a perceptual building block of graphicacy and mathematics, across age, education, and culture. *Scientific Reports, 13*(1), 10266. https://doi.org/10.1038/s41598-023-37172-3

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. (2021). Synergy between research on ensemble perception, data visualization, and statistics education: A tutorial review. *Attention, Perception, & Psychophysics, 83*, 1290-1311.

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine, 28*(2), 205–220. https://doi.org/10.1002/sim.3471

Goldstein, D., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making, 9*(1), 1–14.

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.

Hong, M.-H., Witt, J. K., & Szafir, D. A. (2021). The Weighted Average Illusion: Biases in Perceived Mean Position in Scatterplots. ArXiv:2108.03766. http://arxiv.org/abs/2108.03766

Ibrekk, H., & Morgan, M. G. (1987). Graphical communication of uncer- tain quantities to nontechnical people. *Risk Analysis, 7*(4), 519–529. https://doi.org/10.1111/j.1539-6924.1987.tb00488.x

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

Kim, Y., Walls, L., Krafft, P., & Hullman, J. (2019). A Bayesian cognition approach to improve data visualization. *ACM Human Factors in Computing Systems (CHI)*, 682, 1–14.

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. https://doi.org/10.3758/s13423- 012-0247-5

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2018). Biasing and debiasing health decisions with bar graphs: Costs and benefits of graph literacy. *Quarterly Journal of Experimental Psychology, 71*(12), 2506–2519.

Pentoney, C., & Berger, D. (2016). Confidence intervals and the within-the-bar bias. *The American Statistician, 70*(2), 215–220, doi:10.1080/00031305.2016.1141706

Yuan, L., Haroz, S., & Franconeri, S. (2019). Perceptual proxies for extracting averages in data visualizations. *Psychonomic Bulletin & Review, 26*(2), 669–676. https://doi.org/10.3758/s13423-018-1525-7