What's really wrong with bar graphs of mean values: variable and inaccurate communication of evidence on three key dimensions

Jeremy B. Wilmer and Sarah H. Kerns

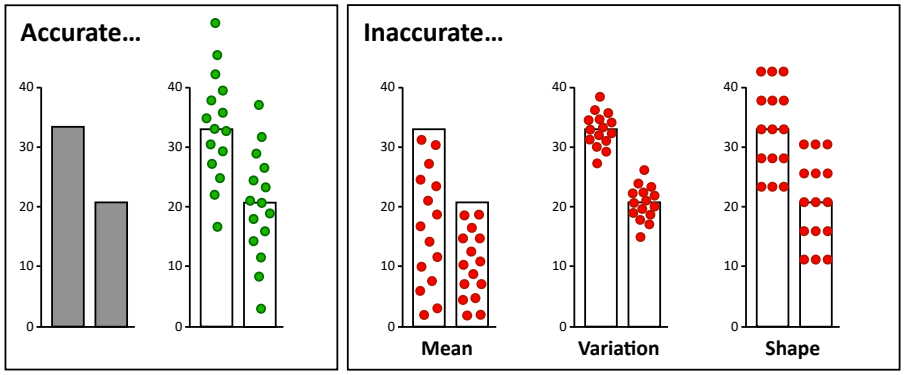

Fig. 1: **Three dimensions of inaccuracy.** Left: A mean bar graph (gray bars) next to a version of the same graph that shows, via green datapoints, hypothetical values that could have been averaged to produce the shown mean values. Right: Illustrative cartoons that show highly inaccurate or implausible (for human behavioral research) reproduction of means (left), variations (center), and distribution shapes (right). The latter were the three dimensions on which we observed both high inter-individual variability and systematic inaccuracy.

Abstract— Human behavioral data are frequently communicated via bar graphs of mean values. Such "mean bar graphs" are presumed to communicate empirical results effectively to non-experts. Yet direct evidence for or against this presumption remains sparse. Here, we ask how a set of widely-consumed scientific mean bar graphs are interpreted by a demographically diverse sample of 133 participants. We use four mean bar graphs of research results, taken from major introductory psychology textbooks, which vary in content (developmental, clinical, social, cognitive), form (unidirectional bars, bidirectional bars), visual aesthetics (four different textbooks' look and feel), data type (objective performance, survey ratings), and study design (experimental, non-experimental). Participants created a detailed sketch of each graph, adding datapoints for their best guess of individual values that were averaged to produce the mean values. Drawn data values were then analyzed as if they were real data. Results were examined for deviations from the ground truth of the published data and for variability between participants. On three separate dimensions—location of the mean, variability around the mean, and normality of distribution shape—we found large, systematic deviations from ground truth and high inter-participant variability. Together, the combination of systematic deviations and inter-participant variability yielded common, extreme misunderstandings, or fallacies, on all three dimensions. We call these fallacies: (1) the Bar-Tip Limit Error: most or all data plotted inside the bar, as if the bar's tip represented the outside limit of the data rather than its balanced center point; (2) the Dichotomization Fallacy: little to no overlap between distributions that should show substantial overlap; (3) the Uniformity Fallacy: data distributed uniformly over its entire range, absent the tails that were present in the real data. These results replicated across the four varying stimulus graphs, suggesting that they are not limited to specific graph form, content, visual aesthetic, data type, or study design. We conclude that the choice to communicate human behavioral data via a mean bar graph carries with it at least two major risks. First, different viewers may walk away from the same graph with widely divergent interpretations of the presented evidence. Second, interpretations may deviate systematically, and, for many viewers, to an extreme degree, from ground truth.

Index Terms— bar graph of means, data visualization, communication of evidence, graph interpretation.

1 INTRODUCTION

Arguably, the main reason to collect and communicate data is to establish an accurate, agreed-upon set of available facts that support reasoned debate and evidence-based decisions. The domain may vary widely: education, health, science, agriculture, technology, law. And the actors too: individuals, groups, institutions. Yet the aim of accurate communication and application of evidence is highly general.

With this aim in mind, ambiguity in visual communication of data could have high costs. Errors of interpretation, even if rare, among key decision-makers could produce poor outcomes. When decisions are made in groups, ambiguity-fueled disagreements on the nature of evidence could slow or halt decision-making. To the extent that ambiguity produces widespread or systematic distortions in understanding of evidence, group decisions might be fast and efficient yet poor. Individual decision-making in everyday life, too, could be poor if individuals misinterpret evidence.

Mean bar graphs have long been criticized, particularly in the sciences, for their ambiguity: for hiding the individual values that were averaged to produce the displayed mean [8, 14, 35, 45, 51, 54–57]. Journal editors have, for example, asked their authors to "kick the bar chart habit ... [for] data that they cannot represent well" [15] and to "Show

- *Jeremy B. Wilmer is with Wellesley College. E-mail: jwilmer@wellesley.edu.*
- *Sarah H. Kerns is with Wellesley College. E-mail: sarah.kerns@wellesley.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

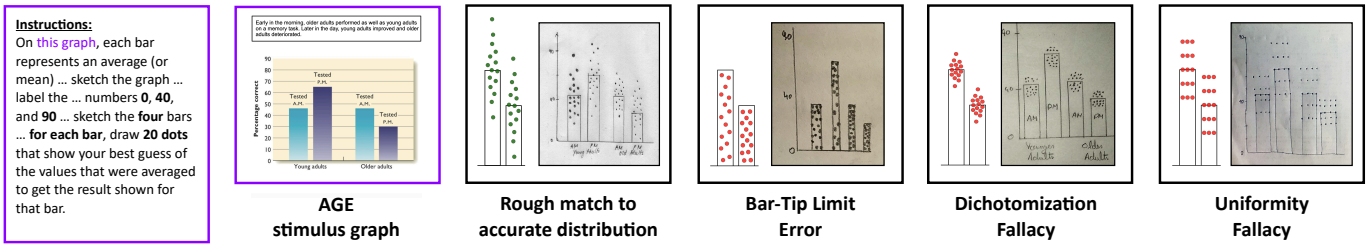

Fig. 2: **Illustrative examples of major deviations of drawings from original research results.** Left (purple outlines): Key excerpts from drawing task instructions and one of our four stimulus graphs. Left-middle (green dots): Submitted drawing whose data correspond roughly to the published results for this stimulus graph. Right graphs (red dots): The leftmost graph shows the Bar-Tip Limit Error (BTLE), where the drawn data points are placed most or all inside the bars. The middle graph shows the Dichotomization Fallacy, where the data points drawn for different bars show little to no overlap. The rightmost graph shows the Uniformity Fallacy, where drawn dots exhibit a flat, or uniform distribution that entirely lacks the tails that are characteristic of most human behavioral data (and many other common forms of data).

the dots in plots ... the data points are the context" [16]. Peer reviewed scientific articles have asked their readers to "show the data" [14] and to "reveal, don't conceal" [57]. Nearly four decades ago, the first bullet-point on page 1 of Tufte's well-known first book *The visual display of quantitative data* stated: "show the data" [53].

In short, the critique that mean bar graphs are ambiguous is not new: it has existed for decades. Yet while such ambiguity is in some sense logically self-evident, what remains largely unknown is whether, and to what degree, ambiguous communication produces variable or inaccurate understanding. That is, with few exceptions [33, 42], direct evidence for how the ambiguity in the mean bar graph impacts understanding of graphs is scarce.

Perhaps partly for this reason, mean bar graphs remain, to this day, widely used [7, 35, 39, 56, 57]. They are particularly ubiquitous in publications, such as introductory psychology textbooks [24,25,31,41], whose intended audience cannot be assumed to have a robust statistical education. Indeed, it is widely believed that the visual simplicity of the mean bar graph yields clarity, especially to non-experts [1,4,62]. Yet this claim too, of simplicity-enhanced understanding, like the claim of ambiguity-hindered understanding, largely lacks direct evidence. Our aim here was therefore to ask, in a relatively simple, direct, explicit manner, how a large, demographically diverse sample of participants interpreted a set of several widely-consumed mean bar graphs.

In contrast to the relative lack of direct empirical data on how mean bar graphs are understood or misunderstood, more is known about how another sort of bar graph is understood and misunderstood. This is the bar graph where each bar represents only a single number: for example, a count (five dogs), amount ($4.28), or percent (67%). When used in this way, bars can reasonably be considered a graphical best-practice because, when aligned at a common baseline, they enable precise and rapid comparisons and height estimates [8, 26].

This is not to say that bar graphs that convey single numbers are immune to all bias or misperception. For example, the choice of y-axis range can impact the reading of a difference between two numbers, either inflating or deflating its apparent severity [12, 44]; and estimated averages of several bars, each of which convey a single number, can produce biases [60, 61]. Indeed, these and other findings related to the reading of single-number bar graphs would likely apply to mean bar graphs as well. Yet mean bar graphs contain an important further ambiguity: an entire underlying distribution of data is not shown. It is this particular ambiguity that is the focus of our work here.

The potential consequences of mean bar graph ambiguity are hinted at by educational researchers like 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 form a part of our statistics teaching schemes" [23]. Or, as Mokros and Russell quoted an 8th grader 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" [40]. Here, we aim to document, in a concrete and detailed way, how specifically this "reverse process," this reverse-engineering step, may operate when it is, indeed, "taken for granted" in the case of a mean bar graph [40].

We presented viewers with four mean bar graphs of real results, taken directly from widely-used introductory psychology textbooks. Viewers then created detailed sketches of those mean bar graphs that included their best guess of where the data were that, when averaged, produced the displayed mean values. For all four graphs we showed, widespread inaccuracy existed in the treatment of the mean, variation, and distribution shape. In the extreme, this resulted in three severe misunderstandings, or fallacies, which we label the Bar-Tip Limit Error, the Dichotomization Fallacy, and the Uniformity Fallacy. Figure 2 shows illustrative examples of these fallacies for one of our four stimulus graphs. Though the fallacies represent extreme responses, they were not uncommon. Moreover, many gradations of inaccuracy existed: indeed, both systematic inaccuracy and high variability characterized all three dimensions.

2 RELATED WORK: 3 DIMENSIONS OF INACCURACY

Here, we flesh out the three dimensions of inaccuracy mentioned above—treatment of mean, variation, and distribution shape—in the context of key prior literature.

2.1 Inaccuracy of means: an embedded replication of our prior work

With regard to the treatment of mean values, the present work constitutes a replication of our own prior study [33]. In that prior study, we showed that about 1 in 5 persons—regardless of age, gender, nationality, educational attainment, and prior coursework—made the error shown in Figures 1 and 2, whereby the bar was treated as a limit rather than as a balanced center point. We called this the Bar-Tip Limit Error and we identified its apparent mechanism as a conflation of mean bar graphs with (visually identical) bar graphs of single values [33].

By embedding a focused replication study within the broader investigation undertaken here, we accomplish two things. First, and most directly, we will demonstrate that the prior result replicates with high precision in an independent sample, a critical, though all too rare, step in the scientific literature. The importance of this step is underscored by major replication projects in cancer biology and psychology that failed to replicate 90% and 60% of prior published results, respectively; also by a major survey of 1,576 natural science researchers that showed "more than half have failed to reproduce their own experiments" [3]. Second, by probing for the Bar-Tip Limit Error here, we are able to filter out those drawings that exhibit it when examining other forms of inaccuracy.

Our prior study includes a comprehensive review of literature relevant to the Bar-Tip Limit Error. Briefly, whereas prior studies had shown that the average viewer considered inside-of-bar locations slightly more likely than outside-of-bar locations [12, 42, 43, 46], the data from our drawing task made clear that the average viewer, essentially, did not exist. What did exist was two categorically different interpretations of the mean bar graph: one with the data relatively

evenly spread across the mean (like the green dots in Figures 1 and 2), and the other with most or all of the data under the mean, as if all individual scores or values were below average (like the leftmost set of red dots in Figures 1 and 2). Notably, the mean is the only aspect of the data that is shown in a mean bar graph, so common misinterpretation of the mean suggests a serious and rather surprising failure of communication.

2.2 Inaccuracy of variation, overlap, and effect size

New to this study is our examination of the accuracy with which viewers of mean bar graphs reproduce , from the original published study, the extent of variation in the data that produced a single mean, or, equivalently, the overlap between distributions. While the Bar-Tip Limit Error implies misunderstanding of the shown mean, the variation of the underlying data is explicitly hidden in a mean bar graph, so one would not necessarily expect perfect reproduction. But this is all the more reason to measure the deviation of the imagined, drawn data from the ground truth of the actual published data. To the extent that a viewer's imagined variation differs from the ground truth, the ambiguity of the mean bar graph has misled that viewer to an incorrect interpretation of the evidence. In other words, absent explicit information, a viewer will insert their implicit assumptions, which could potentially deviate both between individuals and from the ground truth for a given data set. At the same time, the ambiguity of the mean bar graph may provide an opportunity to reveal the nature of implicit assumptions about variation in unseen data.

What are the potential implications of an incorrect interpretation of variation? To preview our result, we find that most viewers underestimate the variation and thus underestimate the overlap between distributions. In other words, these viewers overestimate the (standardized) effect size (Cohen's d, in standard deviation units) of the difference between means. Look back, for example, at the second-to-right drawing in Figure 2, which shows zero overlap in memory performance between older and younger adults when tested in the afternoon. This drawing shows a categorical difference in memory whereby the worst memory among younger adults is (substantially) better than the best memory among older adults. Indeed, one could perfectly predict the age group based upon the memory score alone, and one could likewise perfectly predict (the range of) the memory score from the age group. In fact, this is far from the truth. Many older adults have very strong memory capacity, and many younger adults have weaker memory capacity.

In human behavioral research, an overlap of 69% is, by convention, considered a "large" effect [9, 52], and it is uncommon to find reasonably powered studies whose effects are larger than that [6, 9, 10, 52]. A 69% overlap corresponds to a 0.80 standard deviation difference. In the second-to-right drawing of Figure 2, we see zero overlap, and a massive 7.25 standard deviation difference in memory performance, between older and younger adults in the afternoon. The drawn effect here is several fold larger than the (unusually large) ground truth effect size of the original published result [38].

Plausibly, the underestimation of overlap (and overestimation of difference) to the point of dichotomy may reflect a well-known penchant of the human mind to prefer the relative cognitive simplicity of "black and white" categorical distinctions over more subtle and graded, quantitative and continuum-based "shades of gray." Indeed, in non-visualization domains, the so-called "binary bias" has been documented as a tendency to take continua and boil them down to categories [20]. In visualization specifically, uncertainty visualization is famously avoided in communication of data to non-expert consumers [28]; when it is included, it is often ignored by viewers [28]; and when attended to, it can easily be misinterpreted [12, 27, 32], even by expert consumers [5].

Recent work on data visualization has characterized conceptions of effect size in terms of both subjective ratings of "severity" [11, 44] and so-called Probability of Superiority judgments [27, 32]. While this prior work has outlined important mediators of effect size judgments, to our knowledge, the existing literature in this area has yet to examine effect size judgments for mean bar graphs specifically, nor for real data that provides a ground truth to compare against. More subtly, but perhaps equally important, a large number of the drawn effect sizes that we will see in the present investigation exceed the useful range of common effect size measures [37], including the Probability of Superiority judgment that was used in some prior studies [27, 32]. An important contribution of the drawing-based approach that we use here [33] is that the graph viewer's response produces an entire, concrete, information-rich hypothesized data set that, just like real data, can be examined to determine what analytic approach will do it justice.

2.3 Inaccuracy of distribution shape

Another novel piece of this study is the examination of the accuracy with which viewers of mean bar graphs reproduce the canonical finding in human behavioral research (and in many other domains) that real data tends to be normally distributed, or, at the very least, there are discernible tails in the distribution such that extreme values are less common than values near the middle of the distribution. Again, as with the extent of variation around the mean, and unlike the mean value, the "tailedness" of the distribution is not explicitly given by a mean bar graph. Therefore, the ambiguity of the mean bar graph may provide an opportunity to reveal the nature of implicit assumptions about variation in unseen data. And, to the extent that a viewer's imagined distribution shape differs from the ground truth, the ambiguity of the mean bar graph has misled that viewer to an incorrect interpretation of the evidence. Below, we will use a transformation of the kurtosis ("tailedness") statistic to assess the degree to which drawn data distributions include tails, or "tailing off" of values. As we shall see, a large number of drawings contain no discernible tails whatsoever, showing no more tailedness than a perfectly flat, or uniform, distribution.

The study of distribution shape is widely considered both core to an understanding of data and yet surprisingly difficult for non-experts to grasp [21, 58]. What intuition would cause many viewers to assume a flat, rather than normal (Gaussian, bell-shaped, "tailed") distribution? Several key pieces of intuition on this go back as early as Inhelder and Piaget's mid 20th century work on small children [48]. A key finding of this work is that children may assume that future values are more likely to occur where past values have not occurred (the principle of compensation). For example, when a coin is flipped, the child who sees three straight heads assumes that the next flip will probably produce tails; or, when a dice is rolled, the child who has not seen a 1 in a while assumes that a 1 is increasingly likely until it occurs. Indeed, adults, too, often make these same mistakes. Fundamentally, compensation occurs when one sees independent events (a series of coin flips or dice rolls) as both dependent and related in the particular manner of tending to even out. Plausibly, a thought process similar to compensation could lead many graph viewers to assume that human behavioral data—or other forms of typically bell-shaped data—will be uniformly distributed. Another possibility, also documented by Inhelder and Piaget, is that the statistical novice simply expects human data to be distributed like single dice rolls, whereby in the long run, all plausible values are equally likely [48]. Inhelder and Piaget also proposed the example of rain, which, over the course of a storm, tends to fall relatively equally across different specific spots in a limited (e.g., two meters square) stretch of ground. In this scenario, what the child (or statistical novice) fails to grasp is that the more relevant rain analogy involves consideration only of rain originating from a single spot in the sky, in which case increasing lateral distances become increasingly rare, just as with the normal distribution.

3 METHODS

3.1 Recruitment

Data collection was completed remotely using Qualtrics online survey software. Participants were recruited via Testable Minds, unselected for location or education, and paid $7. Completion time percentiles were: 39m (25th); 54m (50th); 70m (75th).

3.2 Drawing Task and Procedure

Photographs of the procedure, as it appeared on Qualtrics, including exact wording of all questions, are provided as supplemental information and are also posted to Open Science Framework (osf.io/7cxkb).

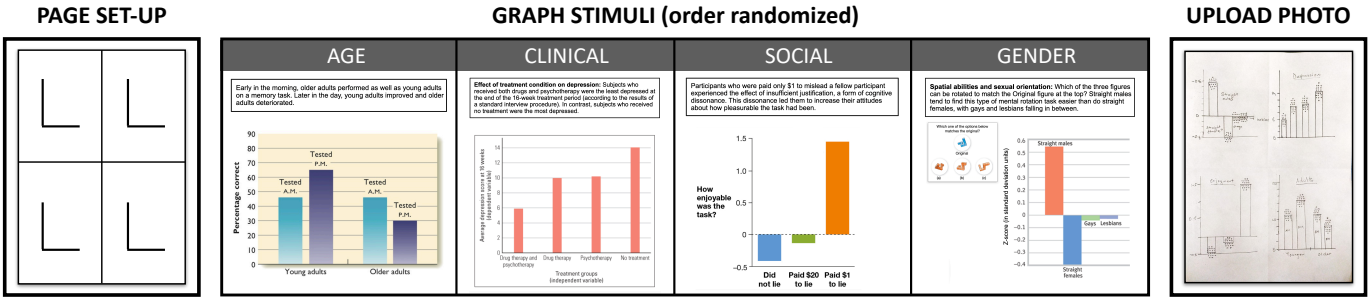

Fig. 3: **Drawing task.** Left (page set-up): The participant is first asked, with the help of a picture of a model example, to separate a piece of standard letter paper into four quadrants and draw an axis in each quadrant. Center (graph stimuli): Graph stimuli are presented one at a time in a random order, and for each—with the graph remaining available the entire time to refer to—the participant is instructed to sketch a version of the graph that includes their "best guess" of 20 dots per bar that could have been averaged to produce the displayed mean values (see Figure 2 for key excerpts from these instructions; see supplemental information or osf.io/7cxkb for detailed photographs of the full procedure from the participant's view). Right (upload photo): The participant is instructed to take a picture of their page of four graph drawings and upload it to Qualtrics.

The development of, and rationale for, the drawing task, and a review of relevant literature, are covered in our prior paper [33]. Briefly, we chose to have participants draw their best guess of data values because it was a relatively concrete, accessible, and information-rich way to access these best guesses. Prior major reviews of methodology in data visualization [29] and in vision science [17] suggest that drawing-based methods, though well-known [22, 33, 34, 59], and though, when appropriately structured, a powerful method for revealing the contents of thought [2, 33], are relatively underutilized compared to other methods such as likert rating scales, reasoning tasks, and psychophysical measurements.

The first page of the drawing task showed participants a picture of a letter-sized piece of paper, separated into four quadrants, with a blank set of axes in each quadrant, and asked them to set up a piece of paper that looked like that one (Figure 3, left). On each of the next four pages, participants were shown, in random order, one of the four stimulus graphs (Figure 3, middle), and were asked to sketch that graph along with dots for their best guess of 20 data points per bar that were averaged to get the shown mean value (see Figure 2, instructions, for key phrases used in these instructions). Upon finishing the four drawings, participants photographed their piece of paper and uploaded it to Qualtrics. All drawn axes and datapoint locations for the two target bars per stimulus graph—the bars representing the highest and lowest mean values—were recorded via the computer program WebPlotDigitizer [50].

Following the drawing task, participants completed 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 that were not the focus of the present investigation.

3.3 Stimuli

The four mean bar graph stimuli pictured in Figure 3 were chosen from Introductory Psychology textbooks due to such texts' strong representation in undergraduate education [47]. We refer to them by their independent variables: AGE, CLINICAL, SOCIAL, and GENDER. These four graph stimuli were selected, respectively, from four separate, widely-used Introductory Psychology textbooks (by Kalat [31], Gray [24], Grison & Gazzaniga [25], and Myers [41]). Each graph stimulus reports an actual scientific result (respectively, from references [13, 19, 38, 49]), which provides a ground truth against which drawn data can be compared.

These stimuli, represent, respectively, varied content (cognitive and developmental, clinical, social, cognitive), form (unidirectional bars, unidirectional bars, bidirectional bars, bidirectional bars), visual aesthetics (four different textbooks' look and feel), data type (objective performance, survey ratings, survey ratings, objective performance), and study design (non-experimental, experimental, experimental, non-experimental). Additionally, the independent variables were chosen to represent potentially self-relevant topics where strong personal opinions might exist. All of these efforts to select widely varying graphs, within the broad area of human behavior, aimed to probe whether such differences might cause variations in results from one graph to another. As we will see below, despite these differences, results varied surprisingly little from one graph to another.

3.4 Collected Data

The 170 participants who completed the study were located in 20 countries and 5 continents. 106 reported male gender, and 64 female. Of these participants, six had technical difficulties (1 upload failure, 5 image quality failures) that made their data unusable, and 30 others were excluded based on the following predetermined criteria (which were identical to our prior study [33]). A participant was excluded if three or four of their individual drawings were unusable due to a basic failure to follow instructions: the drawn datapoints exhibited complete lack of covariation with bar height or direction (1 participant excluded), the drawings failed to include datapoints (11), or had substantially too many datapoints (1), or too few (16). The range of acceptable datapoints was pre-defined as greater than 25% difference from the requested 20 datapoints on the two target bars representing the highest and lowest mean values. 134 participants followed the instructions sufficiently to enable use of their drawings, a yield of 79%, which is typical of online studies of this length [36]. Each of the included drawing pages contained 4 drawn graphs, for 536 graphs total. Individual drawings were excluded for one or more of the same predetermined reasons mentioned above: unreadable dots (2), complete lack of covariation with bars (4), no dots (3), too many dots (4), too few dots (3). The remaining 520 drawn graphs, from 134 participants, broken down by count per-stimulus, were: AGE: 133, CLINICAL: 131, SOCIAL: 130, GENDER: 126.

3.5 Three measures of drawn data accuracy

We developed three measures of drawn data accuracy, one each to probe the accuracy of the implied mean, the variation of data around the mean, and the shape of the distribution.

Bar-Tip Limit Index: This index quantifies the relation of the drawn data to the indicated mean value. It is computed as the proportion of a bar's drawn data points that are outside the bar. If data is symmetrical, as was assumed by the analyses reported by the four original studies from which our stimulus graphs came [13, 19, 38, 49], then the expected value for this index is 50, representing 10 the 20 drawn data points on each side of the mean value, represented by the tip of the bar. Figure 4 (left top) shows a visual representation of Bar-Tip Limit Index scores of 0 (no data points outside the bar), 50 (half of data points outside the bar), and 100 (all data points outside the bar). For this index, a value differing substantially from 50 in either direction would represent

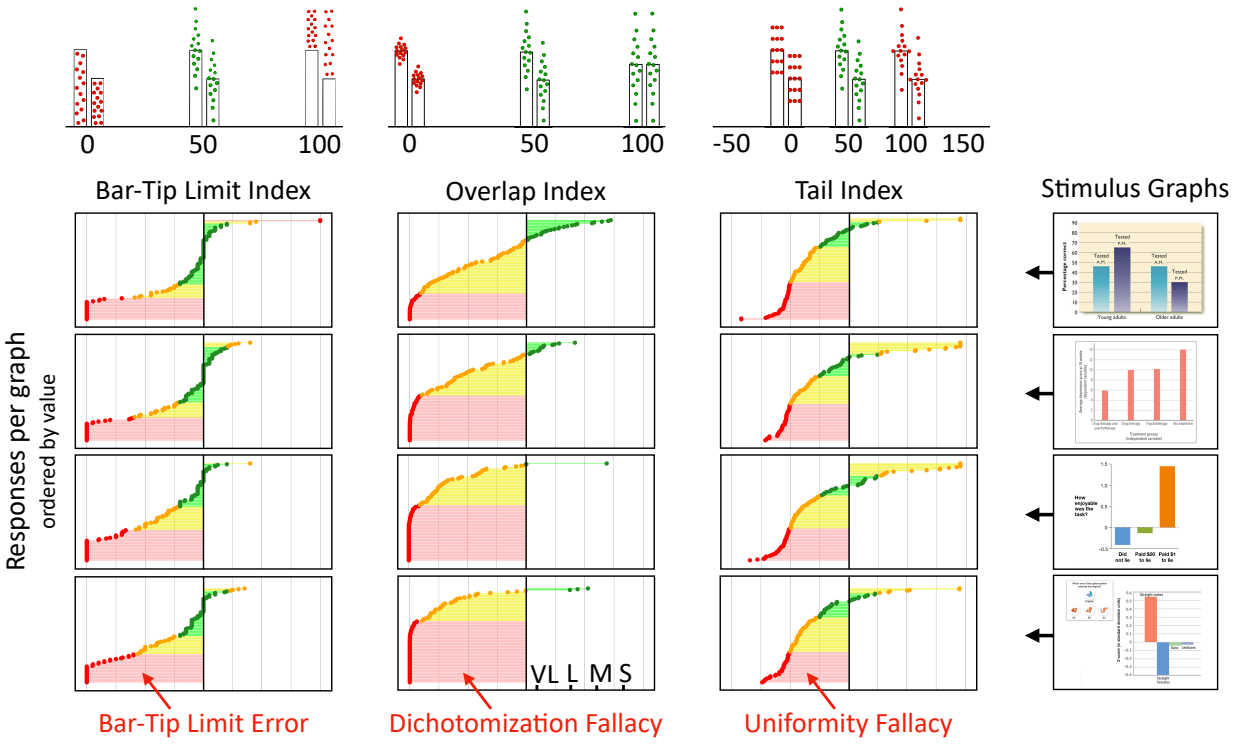

Fig. 4: **Attempts to reproduce published means, overlaps, and distribution shapes from mean bar graphs alone show high variability and systematic inaccuracy.** Pictures at top illustrate key reference values of 0, 50, and 100 for each of three Indices (green = accurate, red = inaccurate); the graphs in each column are aligned such that their x-axis values correspond to the pictures at top. Each row shows results for the single stimulus graph shown at right. Each graph contains datapoints for all participants (N = 133), ordered by their value, for one Index and one stimulus graph. Shaded areas represent the deviation of individual participant drawings from a value 50 on a given Index. The Bar-Tip Limit Index is the percentage of datapoints drawn outside the bar; for a symmetrical distribution, the expected value is 50. The Overlap Index is the percentage overlap between distributions, one of several common measures of effect size; a value of 50 corresponds to a 1.35 SD difference between means, which, in the context of human behavioral research, is uncommonly large to a point of likely implausibility [6, 9, 10, 52]. The Tail Index is computed, based on the kurtosis ("tailed-ness") statistic, such that a flat, uniform distribution obtains a value of 0 and a bell-shaped (Gaussian, normal) distribution obtains a value of 50. Colors are selected to represent reasonable plausibility or accuracy (green), questionable plausibility or accuracy (orange), and implausibly high inaccuracy, or fallacy (red). See text for a discussion of how these ranges were selected. Letters in bottom middle graph mark effect sizes that are conventionally considered to be small (S, 92% overlap, 0.20 SDs), moderate (M, 80% overlap, 0.50 SDs), large (L, 67% overlap, 0.80 SDs), and very large (VL, 55% overlap, 1.20 SDs) [9, 37, 52].

misconception. As we shall see, however, low values (data mostly within the bar) were far more common.

Overlap Index: This index quantifies the variation in the drawn data for each mean value. It is computed as the proportion of the distributions from the two target bars that overlap. Overlap is one of several measures of standardized effect size, and overlap can be transformed into other standardized effect size measures. For example, using the RPsychologist website [37], we can see that an Overlap Index value of 50, shown in Figure 4 (middle top), corresponds to a Cohen's d (difference in standard deviation units) of 1.35, a Cohen's U3 (proportion of higher distribution above the mean of the lower) of 91%, a Probability of Superiority (aka Common Language Effect Size, the proportion the time that a randomly chosen value from the higher distribution would exceed that of a randomly chosen value from the lower distribution) of 83%, and a Number Needed to Treat (the number of persons who would have to switch from the lower to the higher distribution before one person would be expected to have a meaningfully better outcome) of 2.0. Unlike our other indices, the Overlap Index does not have a particular expected value because effect sizes differ across studies. However, overlaps of less than 50 are rare in human behavioral data [6, 9, 10, 52]. Note that for non-infinite sample sizes, Cohen's U3, Probability of Superiority, and Overlap all operate over only a limited range of Cohen's d. For samples of 20 persons per condition, as our participants drew here, these three measures cease to operate above a Cohen's d of about 2, 3, and 4, respectively. As we shall see, Cohen's d values above 2, 3, and even 4, though rare to nonexistent in all but the smallest and most imprecise human behavioral research studies, were commonly drawn.

Tail Index: This index quantifies the shape of the drawn data distributions. It is designed to detect the presence and heaviness of tails in the data. In other words, it is designed to detect whether the participant understands that in real human behavioral data, the more extreme the value, the rarer it generally is. An analogy with human height is perhaps useful here. Height is influenced by many separate genetic and environmental factors, and it is rare that one gets either the "tall" or the "short" version of all of those factors, thus both 4-foot-tall (122 cm tall) or 7-foot-tall (213 cm tall) persons are relatively rare. In human behavior, too, the more extreme the value, the rarer it generally is, and for the same reason: human behavior is sufficiently complex and multifactorial that it is uncommon for all relevant influences to push in the same direction. The canonical version of a distribution with tails is the bell-shaped (normal, Gaussian) distribution. Such a distribution is a basic assumption of most statistical analyses, including all of the analyses reported by four of the original studies from which our stimulus graphs came [13, 19, 38, 49]. The Tail Index is computed as a linear transformation of the kurtosis ("tailedness") statistic. Kurtosis is considered the fourth "moment" of a distribution, with the first three moments being the mean, standard deviation, and skew. In order to

have the Tail Index work similarly to our other two indices, where results become less accurate or plausible as one moves from 50 to 0, we transformed the kurtosis statistic such that a score of 50 corresponded to a perfectly normal distribution and a score of 0 corresponded to a perfectly uniform distribution (that is, a flat distribution where all observed values are equally likely and therefore there are no tails at all). As we shall see, Tail Index scores at, or even below, 0—the score expected of a perfectly flat distribution—were quite common.

3.6 Thresholds of plausibility or accuracy on each index

On continuous measures such as the indices above, thresholds or cutoffs are rarely definitively non-arbitrary, yet they can still be a useful mechanism for identification of more and less accurate values. Next, we will set out ranges on our three indices that we will label, in Figure 4 below, as reasonably plausible or accurate (green), questionable in plausibility or accuracy (orange), and as highly inaccurate or implausible, or fallacy (red).

Bar-Tip Limit Index: We showed in our prior work that scores on the Bar-Tip Limit Index are bimodal, forming two distinct modes, categories, or clusters [33], one near the score of 50 that treats the mean as the balanced center point of the data and one near the score of 0 that treats the bar as a limit. In a case like this, the data itself suggests a natural threshold or cutoff, and we used standard clustering methods to derive a cutoff score of 20 (less than or equal to 20% of dots outside the bar) to identify the Bar-Tip Limit Error (red, in Figure 4). Here, we additionally label (rare) values at the other end of the index (greater than or equal to 80% of dots outside the bar) as fallacy (red), we label values 40-60 (within 10% of the expected 50%) as reasonably plausible or accurate (green), and we label the remaining values (20-40 and 60-80, not inclusive) as questionable in plausibility or accuracy (orange).

Overlap Index: As mentioned when we introduced the Overlap Index just above, scores below 50 (overlaps below 50%) indicate uncommonly large standardized differences (differences above 1.35 in standard deviation units). Given that scores of 100 (100% overlap, or no difference between two mean values) are both common and expected, we labeled all Overlap Index values from 50 to 100 as reasonably plausible or accurate (green), even though values as low as 50 are quite rare in reasonably well-powered (large sample size) human behavioral research [6, 9, 10, 52]. Index values below 50 were labeled as questionable in plausibility or accuracy until a value of 5. A value of 5 (5% overlap) or less represents a difference of more than 4.00 standard deviation units, which is five times Cohen's 0.80 heuristic for a "large." It is, moreover, double the highest proposal by Sawilowsky to extend Cohen's labeled range to 2.00 ("huge") to account for exceptionally rare, but marginally plausible human behavioral results. We label index values of 5 or below on Figure 4 as highly inaccurate or implausible (red), and we label this phenomenon of little to no overlap as the Dichotomization fallacy.

Tail Index: As mentioned when we introduced the Tail Index just above, it was computed such that a value of 50 was the expected (for human behavioral research) and, at least roughly, accurate case (for our stimulus graphs) of a normal distribution. A value of 0 or below, on the other hand, is what would be obtained for a completely flat, uniform distribution. Uniquely among our three indices, since the Tail Index is a linear transformation of a statistic that can range widely, the Tail Index could yield values below 0 or above 100. A value below 0 would be obtained if the tails became so "heavy" as to contain denser data than the center of the distribution, as would be the case for a bimodal distribution. A value above 100 would be obtained in the case of a very small number of extreme outliers that extended the tail to an exceptional degree. Figure 4 shows visual representations of 0, 50, and 100, and one can imagine what further shifts in distribution shape beyond this range might look like. For the Tail Index, we conservatively labeled scores at or below 0 as highly inaccurate or implausible to the point of fallacy (red), and we named this the Uniformity Fallacy. For sake of symmetry, we also labeled (rare) scores that deviated to the same degree (50 points) from 50 (that is, scores above 100) as highly inaccurate or implausible (red). Given the lack of a particularly principled threshold between accurate/plausible (green) and questionably accurate/plausible (orange), we simply placed this threshold at 25, midway between the perfectly normal distribution (50) and the perfectly uniform distribution (0), and, for sake of symmetry, at 75.

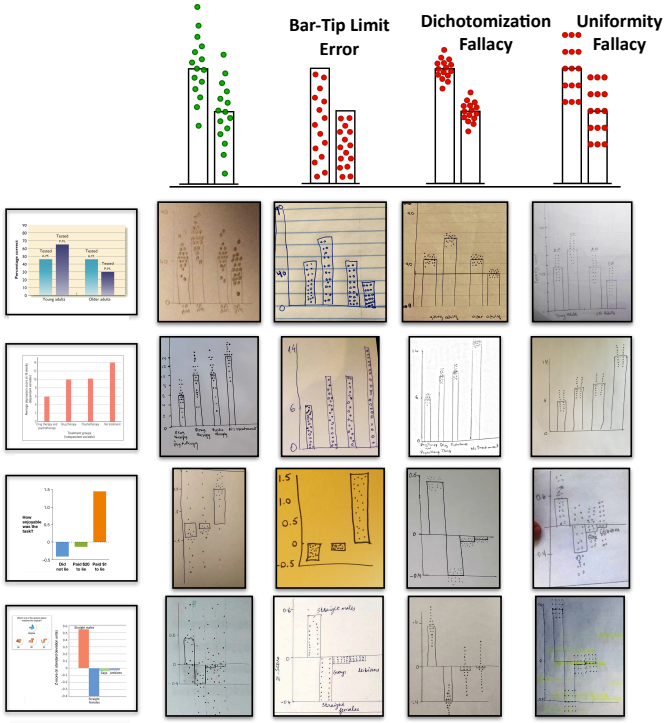

Fig. 5: **Examples of submitted graph drawings that illustrate the three fallacies and correct responses by stimulus graph.** Left column: stimulus graphs (top to bottom) AGE, CLINICAL, SOCIAL, GENDER. Top row: cartoon responses for correct (green, far left) and incorrect: BTLE (red, middle-left), Dichotomization Fallacy (red, middle-right), and Uniformity Fallacy (red, far right). Drawings: one example drawing per type of response (top) and stimulus graph (left).

Table 1: Number (and %) of participants who exhibited each of the three fallacies, separately by stimulus graph.

	AGE	CLINICAL	SOCIAL	GENDER
Total number	133	131	130	126
Bar-Tip Limit Error	29 (22%)	31 (24%)	41 (32%)	38 (30%)
Dichotomization Fallacy	31 (23%)	55 (42%)	65 (50%)	66 (52%)
Uniformity Fallacy	34 (26%)	33 (25%)	29 (22%)	29 (23%)
Any of the above 3	77 (58%)	96 (73%)	112 (86%)	110 (87%)

4 RESULTS

4.1 Inaccuracy of the Mean

The plots of the Bar-Tip Limit Index shown in the left column of Figure 4 closely replicate the results we reported in our previous work [33], showing a bimodal distribution with two modes (regions with vertical slopes of dots), one at a value of 0 (all data points drawn inside the bar) and another at a value of 50 (data points drawn half inside and half outside the bar). A close look at the graphs reveals a somewhat higher rate of the Bar-Tip Limit Error (red dots, red shading) in the graphs whose bars were bidirectional (SOCIAL and GENDER). As Table 1 shows, rates of the Bar-Tip Limit Error ranged from 22% to 32%. Over and above the mode at an index value of 0, a high degree of variation can be observed (the regions of gentle slopes of dots).

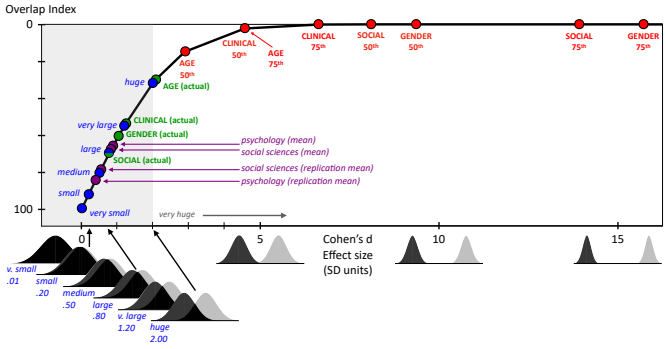

Fig. 6: **Key values plotted for two different effect size measures.** This graph plots two different effect size measures—Overlap Index (y-axis, reversed) and Cohen's d (difference in standard deviation units)—directly against each other (see curved black line). It can be clearly seen in this graph that the Overlap Index operates is capable of measurement over the Cohen's d range of about 0 to 4 standard deviations, but it is unable to capture higher ranges of Cohen's d. The red dots show key percentile values (50th and 75th percentiles) for the four stimulus graphs, with percentiles reversed for the Overlap Index (calculated such that higher percentiles correspond to lower Overlap Index scores) so that higher percentiles correspond to larger effects for both Overlap Index and Cohen's d. It can be seen from the location of the red dots that large portions of the drawn graphs exceed the useful range of the Overlap Index. It can also be seen by comparing the red dots to other color dots that drawn Cohen's d effect sizes for the majority of graphs substantially exceeded, sometimes by an order of magnitude or more, both the ground truth effect sizes for the stimulus graphs (green) and typical effect sizes in psychology and social sciences (purple, [6, 10]). The shaded region shows effect sizes up to what is considered "huge" in human behavior [9, 52]. The pairs of bell curves below the x-axis illustrate the amount of overlap shown by effect sizes of a variety of magnitudes from "very small" (0.01 standard deviation) through "huge" and beyond [9, 52]. These effect sizes are also labeled on the curve itself in blue. Representative effect size distributions (below the x-axis) are shown in dark and light gray.

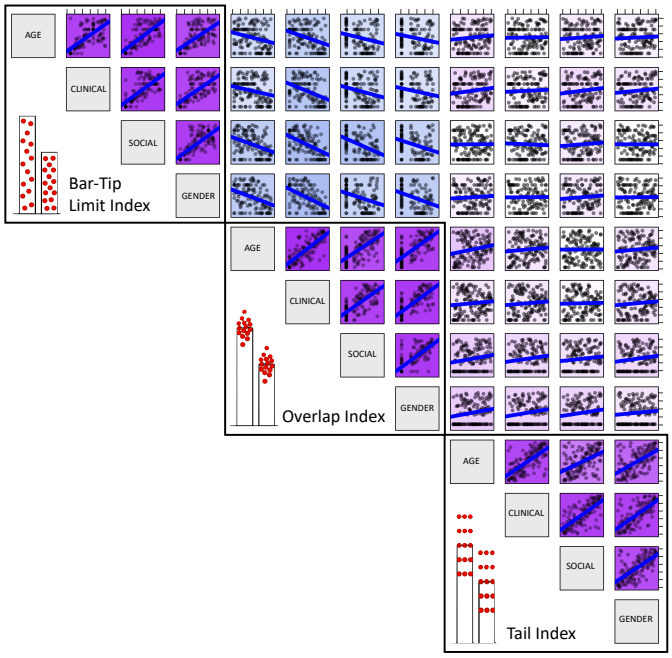

Fig. 7: **The indices used to capture dimensions of accuracy show high consistency across stimuli and show substantial independence relative to each other.** Shown is a color-tinted scatterplot matrix showing associations among the three indices—the Bar-Tip Limit Index, the Overlap Index, and the Tail Index—for each of the four stimulus graphs (AGE, CLINICAL, SOCIAL, and CLINICAL). The exact strength of the (Spearman rank-order) correlations (i.e., the Pearson correlation computed on the shown percentile rank-ordered data) are indicated by the slope of the respective least-squares line (blue); correlation values are also conveyed via color tinting, with darkness conveying strength and color conveying direction of correlation (purple = positive, blue = negative). The non-parametric, rank-order Spearman approach is used here to visually demonstrate the robustness of these correlations despite some minor non-normality of data distributions. The purple triangles outlined with black boxes show the high correlations within each index across stimulus graphs, and the regions outside those boxes show the substantial degree of independence between the three indices. Graphs with red data points provide pictorial representations of the type of inaccuracy captured by each index.

4.2 Inaccuracy of variation, overlap, and effect size

The plots of the Overlap Index shown in the center column of Figure 4 again show a very high degree of variability. Additionally, they show a large mode at or near an index value of 0 (0% overlap between distributions). This we label the Dichotomization Fallacy (red dots, red shading). This fallacy, like the Bar-Tip Limit Error, is more common among the stimulus graphs whose bars were bidirectional (SOCIAL and GENDER). This is perhaps logical, given that the mean values in the bidirectional graphs are physically further from each other. As Table 1 shows, rates of the Dichotomization Fallacy vary substantially from 23% to 52%.

Over and above these rates, it can clearly be seen that, especially for the bidirectional plots, very few participants drew effects that would be considered any less than "very large" by conventional standards [9, 52], and, generally speaking, the great majority of the drawn effects were in a range considered rare for adequately powered human behavioral research [6, 10].

In Figure 6, we plot Overlap Index scores against their corresponding Cohen's d scores. On this graph is plotted, in green, the ground truth, original published effect sizes for each of our four stimulus graphs [13, 19, 38, 49]. All four of these findings would be considered at least "large" by conventional standards, and the AGE stimulus graph would be considered "huge," with a Cohen's d effect size of 2.08 standard deviation units (Overlap of 30% [37]). Nevertheless, even for the AGE graph, the median drawn effect size was larger than the true effect size. In all other cases, the median drawn effect size was far larger than ground truth, and in the case of the GENDER graph, the 75% percentile drawn effect size was, in standard deviation units, nearly fifteen times larger than the actual result.

An important observation here is that the range of effect sizes drawn by participants (red dots mark 50th and 75th percentiles) routinely exceed the useful range of the Overlap Index (asymptote near top). One key advantage of the drawing-based approach used here is that it is possible to observe that a measure's useful range is exceeded and then do a second analysis of the same drawn data using an alternate measure.

4.3 Inaccuracy of distribution shape

The plots of the Tail Index shown in the right column of Figure 4 yet again show a very high degree of variability. Additionally, they each show a discernible mode at or near an index value of 0 (the value corresponding to a perfectly flat, uniform distribution. Values 0 or below we label as the Uniformity Fallacy (red dots, red shading). This fallacy, unlike the other two, is nearly identical in rate among the four graphs, with rates varying only from 22% to 26%. Notably, exceptionally few drawings are in the immediate region of a value of 50, which would correspond to a perfectly normal distribution; and the great majority of drawn graphs are on the uniformity side of normality.

4.4 Different graphs, same fallacies

Figure 5 is designed to provide the reader with some additional visual intuition about the types of drawings that qualify as relatively accurate

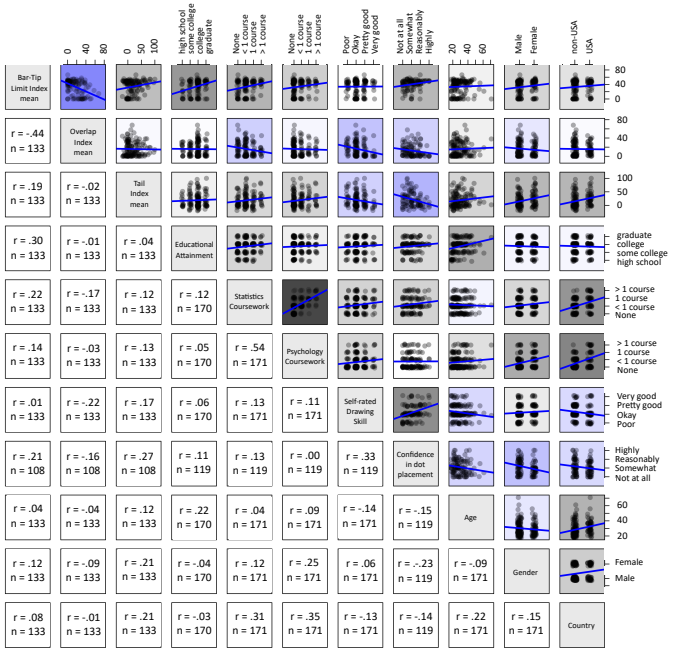

Fig. 8: **Scatterplot matrix shows associations among mean scores for each index, demographic variables, and self-report measures.** Scatterplots are shown between each participant's mean score (averaged across the four stimulus graphs) for each index (Bar-Tip Limit Index, Overlap Index, and Tail Index) and Educational attainment, Statistics coursework, Psychology coursework, Self-rated drawing skill, Confidence in dot placement, Age, Gender, and Country. The physical slope of the least squares lines (blue) equal the Pearson's correlation coefficients (also shown below the diagonal with sample size n). Correlation direction (gray = positive; blue = negative) and strength (darkness) are also indicated by the tint of each individual scatterplot. Pearson r values and sample sizes (n) are shown below the diagonal. For purposes of this analysis, nominal (categorical) variables or ordinal variables were coded via sequential numbers (e.g. non-USA = 0, USA = 1; Poor = 1, Okay = 2, Pretty Good = 3, Very Good = 4), which allows for the computation of meaningful, though somewhat limited in utility, correlation values [9, 52].

(green) or highly implausible or inaccurate on each of the three dimensions that we have examined. An illustrative example is provided for each of the four stimulus graphs. We also include all drawn graphs as supplemental information and have uploaded them as well to the Open Science Framework at osf.io/7cxkb.

4.5 Consistency across graph stimuli, Independence of dimensions

Figure 7 uses a scatterplot matrix of percentile rank values to examine the nonparametric (Spearman) correlations, both within and between the three indices, by stimulus graph. Spearman correlations were used here primarily due to the non-normal (bimodal) Bar-Tip Limit Index distribution, but results are similar for Pearson correlations. The correlations in Figure 7 show that the three indices of inaccuracy are both (1) consistent across graphs within the same individual (purple triangles in black outlined boxes, mean correlation across all of these boxes is rho=0.60) and (2) relatively independent of each other (outside black boxes). Interestingly, there was a moderate negative correlation between Bar-Tip Limit Index and Overlap Index (average rho = -0.33, light blue square region of scatterplots), owing primarily to the tendency for Bar-Tip Limit Error drawings to overlap more than other drawings. Given that those who made the Bar-Tip Limit Error did not appear to understand the nature of the mean bar graph, we consider this negative correlation to be largely an artifact.

4.6 Proportion of drawings that exhibited fallacies

Table 1 summarizes the prevalence data by fallacy and by stimulus graph. The last row of Table 1 shows the prevalence of all three fallacies taken together. In each of our stimulus graphs, the majority of drawings, between 58% and 87%, showed at least one major fallacy, or misunderstanding, of the data that the mean bar graph was designed to communicate. Moreover, there were systematic differences between stimulus graphs, with bidirectional graphs (SOCIAL and GENDER) exhibiting more total fallacies, in large part due to their higher rate of Dichotomization Fallacy. The total prevalence of the three fallacies is partly, but not fully, summative from the individual prevalence values, and individual participants can show one, two, or all three fallacies.

4.7 Demographic correlates

Figure 8 presents a scatterplot matrix of correlations between each participant's mean score (taken across the four stimulus graphs) on each index (Bar-Tip Limit Index, Overlap Index, and Tail Index) and various demographic and other self-reported variables. An important design feature of this scatterplot matrix, as with Figure 7, is that the axis ranges and aspect ratios have been selected such that the physical slope of each blue least squares line precisely equals its respective correlation coefficient (numerical correlation values r and sample size n are shown below the diagonal). The scatterplots are also colored, using a standard conditional formatting approach, such that the strength of the correlations are indicated by darker tint, in this case with gray tints indicating positive, and blue tints negative correlations. Perhaps most notable are the rather low correlations of the three indices with a number of demographic variables including educational attainment, prior coursework, age, gender, and nationality. A possible exception is a moderately sized correlation of the Bar-Tip Limit Index with overall educational attainment, though in the context of the many computed correlations here, this particular correlation can be considered exploratory.

5 DISCUSSION

5.1 Key results

In the present investigation, we used a drawing-based approach to examine how a large sample of demographically diverse individuals interpreted a varied set of widely-consumed mean bar graphs of human behavioral data. Our key results are: (A) a high degree of inter-participant variability and (B) strong, systematic deviations from ground truth on three separate dimensions of interpretation: (1) location of the mean value, (2) variation around the mean value (and, equivalently, overlap between distributions, or standardized effect size of differences between mean values), and (3) the shape of the distribution around the mean value (in particular, whether it contained the tails that are characteristic of a normal, bell-shaped, Gaussian distribution).

Our recruitment strategy intentionally cast a broad net, which succeeded in obtaining a demographically diverse sample in terms of age, gender, educational attainment, prior statics and psychology coursework, and even nationality. Importantly, our key results appear to depend little on such demographic variables.

Several intentional aspects of our approach to graph stimulus selection were: (1) selection of graphs of real scientific results, so that we could compare participant drawings to the ground truth of the original, published results, (2) selection of versions of these graphs that were designed specifically to communicate data to statistical novices, in case seemingly minor stylistic aspects of these graphs might be important for accurate communication (we found evidence for a high level of inaccuracy despite the focus of these graphs on communicating to a non-expert population), (3) selection of graphs that varied on a number of dimensions—content (developmental, clinical, social, cognitive), form (unidirectional bars, bidirectional bars), visual aesthetics (four different textbooks' look and feel), data type (objective performance, survey ratings), and study design (experimental, non-experimental)—to see if key results would replicate across these variations (they did), (4) assignment of every participant to draw all four stimulus graphs to enable robust analyses of (A) whether the same participants showed similar accuracy on a given dimension across different graphs (they

did), and (B) whether different dimensions of accuracy showed reasonable independence from each other, even within the same graph (they did).

We took a drawing-based approach in this study, as we did in a past study [33], because this approach produces an entire, concrete, information-rich hypothesized data set that can be analyzed just like real data. This approach allowed us to examine the same responses from three different perspectives. It also allowed us to ask questions about things like distribution shape or variation without needing to teach these concepts to the participants. Such teaching carries at least two risks: first, one might confuse participants, and second, one might, intentionally or not, nudge them toward a particular response. Finally, the expressive freedom enabled by pencil-and-paper drawing both allowed a wide range of intentional responses and, as well, helped us to identify and discriminate between various types of confused or inaccurate responses. For a more comprehensive treatment of what can be learned from a drawing-based approach to the study of graph interpretation, see our prior paper [33], and for a discussion of how drawing-based approaches can be used to study a variety of other cognitive processes, see a recent tutorial by Bainbridge [2].

5.2 Limitations and future work

While this investigation answers the questions that it set out to tackle, it was, inevitably, limited in many ways that raise a variety of new questions. First, while our sample varied on a number of key demographic variables, there are many more that could be examined. For example, we recruited adults. What would children do? The adults we recruited had no necessary expertise in statistics or in the science of human behavior. What would one see in a select sample of college-level statistics or psychology majors? In a sample of persons who have completed at least one course devoted specifically to data visualization? In a sample of published researchers, expert statisticians, or high-level content experts? Prior research suggests that even published researchers do not always understand common data presentation strategies [5]. Would they be better able to draw data that is accurate relative to ground truth in an area that they are deeply familiar with? Slightly less familiar with?

Second, while our results clearly replicated across several stimulus graphs that varied in a number of ways, thereby demonstrating that the results are not unduly restricted to a particular aspect of graph form or content, all of our graphs were taken from Introductory Psychology textbooks, and all involved human behavior. Would similar results be obtained for graphs from other Introductory Science textbooks? For non-behavioral human data? For data that does not come from humans, or even from living organisms? And, even within Introductory Psychology textbooks would similar results be observed for non-bar representations of mean values, for example line graphs of mean values?

Third, while the drawing-based approach that we took here yielded significant insights, what other insights could be revealed by taking what was learned here and probing it via other response methods: for example, structured decision-making tasks, or detailed interviews.

5.3 Potential implications and applications

Limitations aside, what are some conceivable implications or applications of this work. Again, these can usefully be framed via questions. If the ambiguity of mean bar graphs leads to markedly different conceptions of the data by different viewers, as our work here suggests, what are the specific consequences of this for real-world individual or group decision-making? We proposed in the Introduction that an inaccurate reading of evidence might lead to poor individual decision-making and to unnecessary disagreements and inefficiencies in group decision-making contexts. But further work is needed to directly demonstrate that.

Additionally, what should be used instead of bar graphs. We have proposed previously that the data visualization community has already taken an important step by limiting the use of bar graphs to convey mean values [33]. Those who have criticized mean bar graphs in the past have suggested multiple less ambiguous alternatives [8, 14, 35, 45, 51, 54–57]. Some of these involve showing all of the individual data points, for example via dot plots, bee swarm plots, and sinaplots, with techniques like jitter, opacity, and non-filled data markers used to avoid overplotting. Others, such as boxplots or violin plots retain a certain level of abstraction away from the data. Still other options such as quantile dotplots [18] or hypothetical outcome plots [30] might be borrowed from the uncertainty visualization literature. Ultimately, further work will be needed to sort out which are the best options for a given scenario.

We think that one of the most exciting potential applications of the present work is in direct education of students about what real human data looks like. Increasingly, an understanding of data is considered important, and yet data education can be surprisingly difficult [21, 58]. The first author already regularly uses drawing in every course he teaches from Sensation and Perception to Statistics and Data Analysis to a high-level seminar on Nature and Nurture. Students draw the data they think underlies a particular graph, then compare their drawings to those of other students to notice variation between drawings, then analyze their data and compare the results of their analyses to those of the real data in the published paper. Students also draw data to predict results before seeing the results, again, comparing their drawn data to those of other students and then to the actual results.

Time and further careful empirical work will tell what lessons can be learned, what potential implications will be confirmed, and what applications will have the most value. In the meantime, if nothing else, next time you see a mean bar graph of human data, we hope you will pause for a moment to wonder where several different people you know might draw data points, if asked to do so.

Acknowledgments

Funded in part by NSF award #1624891 to JBW, a Brachman Hoffman grant to JBW, and a subaward from NSF grant #1837731 to JBW. The authors thank Ally Kim for assisting with early pilot work.

References

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

[2] W. A. Bainbridge. A tutorial on capturing mental representations through drawing and crowd-sourced scoring. *Behavior Research Methods*, pp. 1–13, 2021.

[3] M. Baker. Reproducibility crisis. *Nature*, 533(26):353–66, 2016.

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

[5] S. Belia, F. Fidler, J. Williams, and G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389, 2005.

[6] C. F. Camerer, A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018.

[7] J. C. Chen, R. J. Cooper, M. E. McMullen, and D. L. Schriger. Graph quality in top medical journals. *Annals of emergency medicine*, 69(4):453–461, 2017.

[8] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[9] J. Cohen. Statistical power analysis for the behavioral sciences. *Hillsdale, NJ*, 1988.

[10] O. S. Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[11] M. Correll, E. Bertini, and S. Franconeri. Truncating the y-axis: Threat or menace? In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020.

[12] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics*, 20(12):2142–2151, 2014.

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

[14] G. B. Drummond and S. L. Vowler. Show the data, don't conceal them. *Advances in physiology education*, 35(2):130–132, 2011.

[15] Editors. Show dots in plots: we encourage our authors to display data points in graphs, and to deposit the data in repositories. *Nature Methods*, 11(2), 2014.

[16] 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.

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

[18] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. 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, 2018.

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

[20] M. Fisher and F. C. Keil. The binary bias: A systematic distortion in the integration of information. *Psychological Science*, 29(11):1846–1858, 2018.

[21] J. Garfield and D. Ben-Zvi. How students learn statistics revisited: A current review of research on teaching and learning statistics. *International statistical review*, 75(3):372–396, 2007.

[22] D. G. Goldstein and D. Rothschild. Lay understanding of probability distributions. *Judgment & Decision Making*, 9(1), 2014.

[23] S. Goodchild. School pupils' understanding of average. *Teaching Statistics*, 10(3):77–81, 1988.

[24] P. O. Gray and D. F. Bjorklund. *Psychology*. Worth, 8 ed., 2017.

[25] S. Grison and M. Gazzaniga. *Psychology*. Norton, 3 ed., 2019.

[26] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 203–212, 2010.

[27] J. M. Hofman, D. G. Goldstein, and J. Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pp. 1–12, 2020.

[28] J. Hullman. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics*, 26(1):130–139, 2019.

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

[30] J. Hullman, P. Resnick, and E. Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one*, 10(11):e0142444, 2015.

[31] J. W. Kalat. *Introduction to psychology*. Wadsworth, 11 ed., 2016.

[32] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE transactions on visualization and computer graphics*, 27(2):272–282, 2020.

[33] S. H. Kerns and J. B. Wilmer. 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, 2021.

[34] Y.-S. Kim, L. A. Walls, P. Krafft, and J. Hullman. A bayesian cognition approach to improve data visualization. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–14, 2019.

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

[36] L. Litman and J. Robinson. *Conducting online research on Amazon Mechanical Turk and beyond*. Sage Publications, 2020.

[37] K. Magnusson. Interpreting cohen's d, 2021.

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

[39] S. A. Mogull and C. T. Stanfield. Current use of visuals in scientific communication. In *2015 IEEE international professional communication conference (IPCC)*, pp. 1–6. IEEE, 2015.

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

[41] D. G. Myers and G. N. DeWall. *Psychology*. Worth, 12 ed., 2017.

[42] G. E. Newman and B. J. Scholl. Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic bulletin & review*, 19(4):601–607, 2012.

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

[44] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini. How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pp. 1469–1478, 2015.

[45] M. Pastore, F. Lionetti, and G. Altoè. When one shape does not fit all: a commentary essay on the use of graphs in psychological research. *Frontiers in psychology*, 8:1666, 2017.

[46] C. S. Pentoney and D. E. Berger. Confidence intervals and the within-the-bar bias. *The American Statistician*, 70(2):215–220, 2016.

[47] J. J. Peterson and A. Sesma Jr. Introductory psychology: What's lab got to do with it? *Teaching of Psychology*, 44(4):313–323, 2017.

[48] J. Piaget and B. Inhelder. *The origin of the idea of chance in children / La genèse de l'idée de hasard chez l'enfant*. Presses Universitaires de France, 1951.

[49] Q. Rahman, G. D. Wilson, and S. Abrahams. Biosocial factors, sexual orientation and neurocognitive functioning. *Psychoneuroendocrinology*, 29(7):867–881, 2004.

[50] A. Rohatgi. Webplotdigitizer, 2015.

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

[52] S. S. Sawilowsky. New effect size rules of thumb. *Journal of modern applied statistical methods*, 8(2):26, 2009.

[53] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.

[54] A. Vail and J. Wilkinson. Bang goes the detonator plot! *Reproduction*, 159(2):E3–E4, 2020.

[55] H. Wainer. How to display data badly. *The American Statistician*, 38(2):137–147, 1984.

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

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

[58] U. Wilensky. What is normal anyway? therapy for epistemological anxiety. *Educational studies in mathematics*, 33(2):171–202, 1997.

[59] J. Wolfe, K. Kluender, D. Levi, L. Bartoshuk, R. Herz, R. Klatzky, and D. Merfeld. Sensation and perception. 2020.

[60] C. Xiong, C. R. Ceja, C. J. Ludwig, and S. Franconeri. Biased average position estimates in line and bar graphs: Underestimation, overestimation, and perceptual pull. *IEEE transactions on visualization and computer graphics*, 26(1):301–310, 2019.

[61] L. Yuan, S. Haroz, and S. Franconeri. Perceptual proxies for extracting averages in data visualizations. *Psychonomic bulletin & review*, 26(2):669–676, 2019.

[62] A. Zubiaga and B. Mac Namee. Graphical perception of value distributions: An evaluation of non-expert viewers' data literacy. *The Journal of Community Informatics*, 12(3), 2016.