Two graphs walk into a bar: Readout-based measurement reveals the Bar-Tip Limit error, a common, categorical misinterpretation of mean bar graphs

Sarah H. Kerns
Department of Psychology, Wellesley College, Wellesley, MA, USA

Jeremy B. Wilmer
Department of Psychology, Wellesley College, Wellesley, MA, USA

How do viewers interpret graphs that abstract away from individual-level data to present only summaries of data such as means, intervals, distribution shapes, or effect sizes? Here, focusing on the mean bar graph as a prototypical example of such an abstracted presentation, we contribute three advances to the study of graph interpretation. First, we distill principles for Measurement of Abstract Graph Interpretation (MAGI principles) to guide the collection of valid interpretation data from viewers who may vary in expertise. Second, using these principles, we create the Draw Datapoints on Graphs (DDoG) measure, which collects drawn readouts (concrete, detailed, visuospatial records of thought) as a revealing window into each person's interpretation of a given graph. Third, using this new measure, we discover a common, categorical error in the interpretation of mean bar graphs: the Bar-Tip Limit (BTL) error. The BTL error is an apparent conflation of mean bar graphs with count bar graphs. It occurs when the raw data are assumed to be limited by the bar-tip, as in a count bar graph, rather than distributed across the bar-tip, as in a mean bar graph. In a large, demographically diverse sample, we observe the BTL error in about one in five persons; across educational levels, ages, and genders; and despite thoughtful responding and relevant foundational knowledge. The BTL error provides a case-in-point that simplification via abstraction in graph design can risk severe, high-prevalence misinterpretation. The ease with which our readout-based DDoG measure reveals the nature and likely cognitive mechanisms of the BTL error speaks to the value of both its readout-based approach and the MAGI principles that guided its creation. We conclude that mean bar graphs may be misinterpreted by a large portion of the population, and that enhanced measurement tools and strategies, like those introduced here, can fuel progress in the scientific study of graph interpretation.

Introduction

Background on bar graphs

How can one maximize the chance that visually conveyed quantitative information is accurately and efficiently received? This question is fundamental to data-driven fields, both applied (policy, education, medicine, business, engineering) and basic (physics, psychological science, computer science). Long a crux of debate on this question, "mean bar graphs"—bar graphs depicting mean values—are both widely used, for their familiarity and clean visual impact, and widely criticized, for their abstract nature and paucity of information (Tufte & Graves-Morris, 1983, p. 96; Wainer, 1984; Drummond & Vowler, 2011; Weissgerber et al., 2015; Larson-Hall, 2017; Rousselet, Pernet, & Wilcox, 2017; Pastore, Lionetti, & Altoe, 2017; Weissgerber et al., 2019; Vail & Wilkinson, 2020).

Conversely, bar graphs are considered a best-practice when conveying counts—whether raw or scaled into proportions or percentages. These "count bar graphs" are more concrete and hide less information than mean bar graphs. They are usefully extensible: bars may be stacked on top of each other to convey parts of a whole (a stacked bar graph) or arrayed next to each other to convey a distribution (a histogram). And they take advantage of a core bar graph strength: the alignment of bars at a common baseline supports rapid, precise height estimates and comparisons (Cleveland & McGill, 1984; Heer & Bostock, 2010).

When a bar represents a count, it can be thought of as a stack, and the metaphor of bars-as-stacks is relatively accessible across ages and expertise levels (Zubiaga & MacNamee, 2016). In fact, the introduction of bar graphs in elementary education is often

Citation: Kerns, S. H., & Wilmer, J. B. (2021). Two graphs walk into a bar: Readout-based measurement reveals the Bar-Tip Limit error, a common, categorical misinterpretation of mean bar graphs. *Journal of Vision*, *21*(12):17, 1–36, https://doi.org/10.1167/jov.21.12.17.

https://doi.org/10.1167/jov.21.12.17 Received July 20, 2020; published November 30, 2021 ISSN 1534-7362

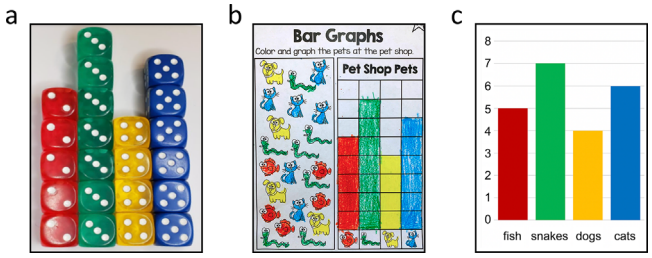

Figure 1. **Elementary progression of count bar graph instruction.** (a) Children are taught first using manipulatives; (b) then they transition to drawn stacks; (c) finally, they are introduced to undivided bars.

accomplished in just this way: with manipulative stacks (Figure 1a), translated to more abstract drawn stacks (Figure 1b), and further abstracted to undivided bars (Figure 1c).

While the mean bar graph potentially borrows some virtues from the count bar graph, the use of a single visual symbol (the bar, Figures 2a, 2c) to represent two profoundly different quantities (means and counts, Figures 2b, 2d) adds inherent ambiguity to the interpretation of that visual symbol.

The information conveyed by these two visually identical bar graph types differs categorically, as Figure 2 shows. Because a count bar graph depicts summation, its bar-tip is the limit of the individual-level data, which are contained entirely within the bar (Figures 2d, 2e). We call this a Bar-Tip Limit (BTL) distribution. A mean bar graph, in contrast, depicts a central tendency; it uses the bar-tip as the balanced center point, or mean, and the individual-level data are distributed across that bar-tip (Figure 2b). We call this a Bar-Tip Mean distribution.

The question of mean bar graph accessibility

While the rationale for using mean bar graphs—or mean graphs more generally—varies by intended audience, a common theme is visual simplification. Potentially, the simplification of substituting a single mean value for many raw data values could ease comparison, aid pattern-seeking, enhance discriminability, reduce clutter, or remove distraction (Barton & Barton, 1987; Franzblau & Chung, 2012). Communications intended for nonexpert consumers, such as introductory textbooks, may favor mean bar graphs because their visual simplicity is assumed to yield accessibility (Angra & Gardner, 2017). In contrast, communications intended for experts, such as scholarly publications, may favor mean bar graphs because their visual simplicity is assumed to yield efficiency (Barton & Barton, 1987). In either case, the aim is to enhance communication. Yet simplification that abstracts away from the concrete underlying data could also mislead if the viewer fails to accurately intuit the nature of that underlying data.

Visual simplification can, in some cases, enhance comprehension. Yet is this the case for mean bar graphs? Basic research in vision science and psychology provides at least four theoretical causes to doubt the accessibility of mean bar graphs. First, in many domains, abstraction, per se, reduces understandability, particularly in nonexperts (Fyfe et al., 2014; Nguyen et al., 2020). Second, less-expert consumers may lack sufficient familiarity with a mean bar graph's dependent variable to accurately intuit the likely range, variability, or shape of the distribution that it represents. In this context, the natural tendency to discount variation (Moore et al., 2015) and exaggerate dichotomy (Fisher & Keil, 2018) could potentially

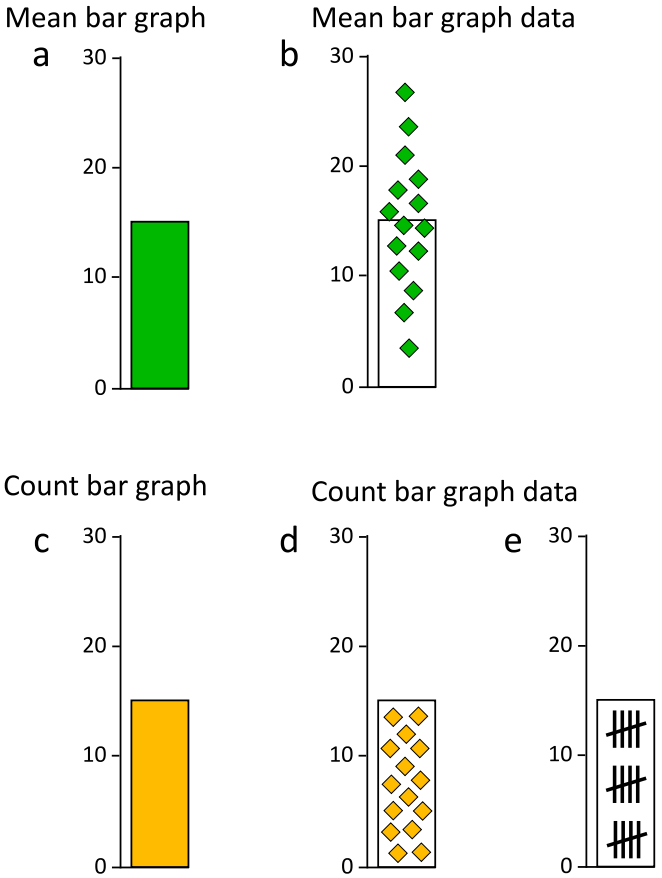

Figure 2. **Data distribution differs categorically between mean and count graphs.** (a) Mean bar graphs and (c) count bar graphs do not differ in basic appearance, but they do depict categorically different data distributions. (b) In a mean bar graph, the bar-tip is the balanced center point, or mean, with the data distributed across it. We call this a Bar-Tip Mean distribution. (d, e) In a count bar graph, the bar-tip acts as a limit, containing the summed data within the bar. We call this a Bar-Tip Limit (BTL) distribution.

distort interpretations. Third, the visual salience of a bar could be a double-edged sword, initially attracting attention, but then spreading it across the bar's full extent, in effect tugging one's focus away from the mean value represented by the bar-tip (Egly et al., 1994). Finally, the formal visual simplicity (i.e., low information content) of the bar does not guarantee that it is processed more effectively by the visual system than a more complex stimulus. Many high-information stimuli like faces (Dobs et al., 2019) and scenes (Thorpe et al., 1996) are processed more rapidly and accurately than low-information stimuli such as colored circles or single letters (Li et al., 2002). Relatedly, a set of dots is accurately and efficiently processed into a mean spatial location (Alvarez & Oliva, 2008), raising questions about what is gained by replacing datapoints on a graph with their mean value. Together, these theory-driven doubts provided a key source of broad motivation for the present work.

Adding to the broad concerns raised by basic vision science and psychology research are numerous specific, direct critiques of mean bar graphs in particular. Such critiques, however, are chiefly theoretical, as opposed to empirical, and focus on expert audiences, as opposed to more general audiences (Tufte & Graves-Morris, 1983, p. 96; Wainer, 1984; Drummond & Vowler, 2011; Weissgerber et al., 2015; Larson-Hall, 2017; Rousselet, Pernet, & Wilcox, 2017; Pastore, Lionetti, & Altoe, 2017; Weissgerber et al., 2019; Vail & Wilkinson, 2020). In contrast, our present work provides a direct, empirical test of the claim that mean bar graphs are accessible to a general audience (see Related works and Results for a detailed discussion of prior work of this sort).

Aim and process for current work

To assess mean bar graph accessibility, we needed an informative measure of abstract-graph interpretation; that is, a measure of how graphs that abstract away from individual-level data to present only summary-level, or aggregate-level, information, are interpreted. Over more than a decade, our lab has developed cognitive measures in areas as diverse as sustained attention, aesthetic preferences, stereoscopic vision, face recognition, visual motion perception, number sense, novel object recognition, visuomotor control, general cognitive ability, emotion identification, and trust perception (Wilmer & Nakayama, 2007; Wilmer, 2008; Wilmer et al., 2012; Degutis et al., 2013; Halberda et al., 2012; Germine et al., 2015; Fortenbaugh et al., 2015; Richler, Wilmer, & Gauthier, 2017; Deveney et al., 2018; Sutherland et al., 2020).

The decision to develop a new measure is never easy; it takes a major investment of time and energy to iteratively refine and properly validate a new measure. Yet the accessibility of abstract-graphs presented a special measurement challenge that we felt warranted such an investment. The challenge: record the viewer's conception of an invisible entity—the individual-level data underlying an abstract-graph—while neither invoking potentially unfamiliar statistical concepts nor explaining those concepts in a way that could distort the response.

Our lab's standard process for developing measures has evolved to include three components: (1) Identify guiding principles (Wilmer, 2008; Wilmer et al., 2012; Degutis et al., 2013). (2) Design and refine a measure (Wilmer & Nakayama, 2007; Wilmer et al., 2012; Halberda et al., 2012; Degutis et al., 2013; Germine et al., 2015; Fortenbaugh et al., 2015; Richler, Wilmer, & Gauthier, 2017; Deveney et al., 2018; Kerns, 2019; Sutherland et al., 2020). (3) Apply the measure to a question of theoretical or practical importance (Wilmer & Nakayama, 2007; Wilmer et al., 2012; Halberda et al., 2012; Degutis et al., 2013; Germine et al., 2015; Fortenbaugh et al., 2015; Richler, Wilmer, & Gauthier, 2017; Deveney et al., 2018; Kerns, 2019; Sutherland et al., 2020). We devoted over a year of intensive piloting to the refinement of these three components for the present project. While the parallel nature of the development process produced a high degree of complementarity between the three, each represents its own separate and unique contribution.

A turning point in the development process was our rediscovery of a group of drawing-based neuropsychological tasks that have been used to probe for pathology of perception, attention, and cognition in brain damaged patients (Landau et al., 2006; Smith, 2009). In one such task, the patient is asked to draw a

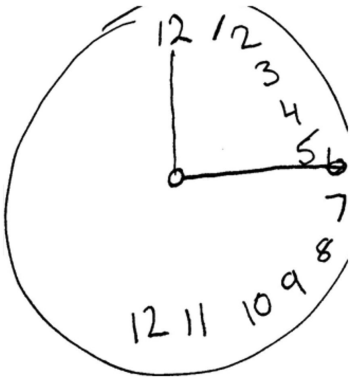

Figure 3. **Readout of a clock drawn by a hemispatial neglect patient.** (https://commons.wikimedia.org/wiki/File:AllochiriaClock.png) The Draw Datapoints on Graphs (DDoG) measure was inspired by readout-based neuropsychological tasks like the one that produced this distorted clock drawing. Such readout-based tasks have long been used with brain damaged patients to probe for pathology of perception, attention, and cognition (Smith, 2009).

	Principle	Execution	Goal
Facilitates *general usage* and *valid measurement*	***Expressive freedom***	Employ a flexible response medium, such as drawing, that allows relatively unconstrained recording of thought processes.	Minimize leading or restricting of responses, allow clusters of responses to stand out, and enable unexpected responses.
	Limited instructions	Keep instructions succinct, concrete, and task-directed; minimize terms, concepts, or task steps that require explanation.	Facilitate a response that is undistorted by information embedded in the instructions themselves.
	Limited mental transformations	Elicit a stimulus-matched response: a response in the same basic format as the stimulus.	Help the response to accurately reflect thinking by avoiding translations; for example, between numbers, words, and spatial representations.
Facilitates *valid measurement*	***Ground-truth linkage***	Select graph stimuli and responses for which logically and/or empirically correct and incorrect answers exist.	Allow objective assessment of comprehension.
	Ecological validity	Excerpt graphs for use as stimuli from real-world sources, relatively unchanged and appropriately contextualized.	Support generalization of insights back to real-world sources, in part by lowering the risk of excluding unexpectedly influential factors.
	Information richness	Elicit high-bandwidth responses, i.e., responses that contain numerous, dense, continuous, and/or multidimensional pieces of information.	Enable higher resolution analyses.

Table 1. **Measurement of Abstract Graph Interpretation (MAGI) principles.**

clock, and pathological inattention to the left side of visual space is detected via the bunching of numbers to the right side of the clock (Figure 3). This clock drawing test exhibits a feature that became central to the present work: readout-based measurement.

A readout is a concrete, detailed, visuospatial record of thought. Measurement via readout harnesses the viewer's capacity to transcribe their own thoughts, with relative fullness and accuracy, when provided with a response format that is sufficiently direct, accessible, rich, and expressive. As the clock drawing task in Figure 3 seeks to read out remembered numeral positions on a clock face, our measure seeks to read out assumed datapoint positions on a graph.

Contribution 1: Identify guiding principles

Distillation of the Measurement of Abstract Graph Interpretation (MAGI) principles

The first of this paper's three contributions is the introduction of six succinct, actionable principles to guide collection of graph interpretation data (Table 1). These principles are aimed specifically at common graph types (e.g., bar/line/box/violin) and graph markings (e.g., confidence/prediction/interquartile interval) that abstract away from individual-level data to present aggregate-level information. These Measurement of Abstract-Graph Interpretation (MAGI) principles center around two core aims, general usage and valid measurement.

General usage—that is, testing of a general population that may vary in statistical and/or content expertise—is most directly facilitated if a measure: (1) avoids constraining the response via limited options (*Expressive freedom*); (2) avoids priming the response via suggestive instructions (*Limited instructions*); and (3) avoids obscuring or hindering thinking via unnecessary interpretive or translational steps (*Limited mental transformations*).

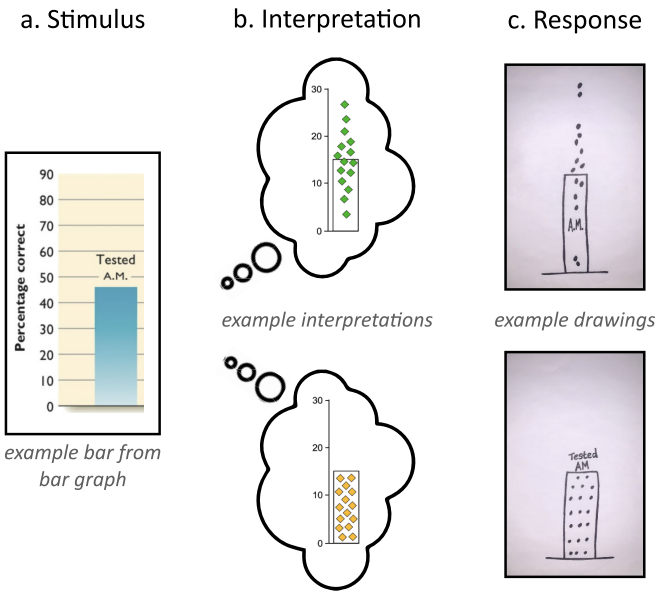

Figure 4. **The Draw Datapoints on Graph (DDoG) measure maintains the graph as a consistent reference frame across its three stages.** (a) Participants are presented with a graph stimulus that (b) produces a mental representation of the data; (c) this interpretation is recorded by sketching a version of the graph along with hypothesized locations of individual data values. Drawings are representative examples of the two common responses seen in pilot data collection: the correct Bar-Tip Mean response (top), and the incorrect Bar-Tip Limit (BTL) response (bottom).

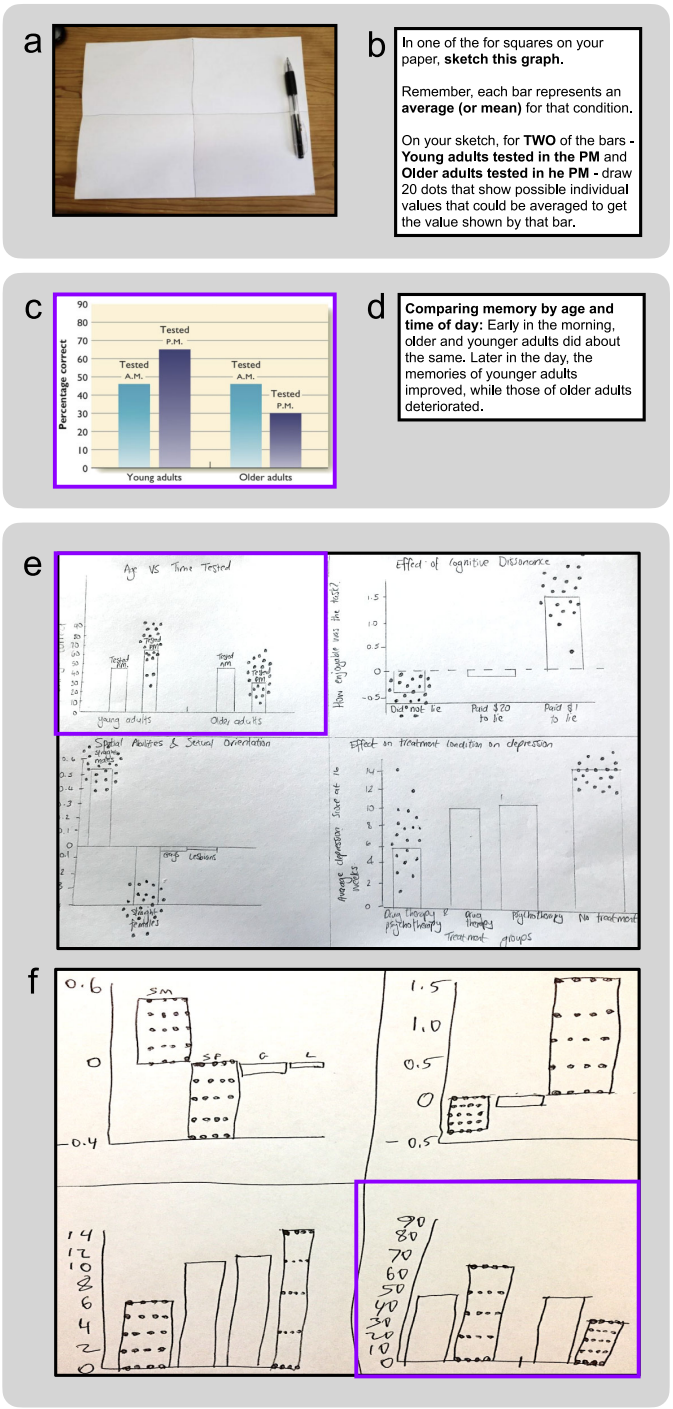

Figure 5. **The DDoG measure implements the MAGI principles. The DDoG measure collects readouts of abstract-graph interpretation.** Shown are the major pieces of the DDoG measure's procedure, with the relevant MAGI principle(s) in parentheses. (a) Drawing page: showed the participant how to set up their paper for drawing (*Expressive freedom, Limited instructions*). (b) Instructions: explained what to draw (*Limited instructions*). (c) Stimulus graph: was the graph to be interpreted (*Ecological validity, Ground-truth linkage, Limited mental transformations*). (d) Figure caption: was presented with each stimulus graph to help clarify its content (*Ecological validity*). (e–f) Representative readouts: four-graph readouts from two separate participants, with readouts of graph

→

Valid measurement is facilitated by all six principles. The three principles just mentioned—*Expressive freedom, Limited instructions,* and *Limited mental transformations*—all help responses to more directly reflect the viewer's actual graph interpretation. This is especially true of *Limited mental transformations*. As we will see below, a set of mental transformations embedded in a popular probability rating scale measure delayed for over a decade the elucidation of a phenomenon that our investigation here reveals.

Valid measurement is additionally facilitated by: (1) objective scoring, via the existence of correct answers (*Ground-truth linkage*); (2) real-world applicability of results, via the sourcing of real graphs (*Ecological validity*); and (3) a clear, high-resolution window into viewers' thinking, via a response that has high detail and bandwidth (*Information richness*).

As shown in the next two sections, the MAGI principles provided a foundational conceptual basis for the creation of the Draw Datapoints on Graphs (DDoG) measure and for understanding the ease with which that DDoG measure identified the common, categorical Bar-Tip Limit (BTL) error. These principles additionally provided a structure for comparing our new measure with existing measures (Related works). In these ways, we demonstrate the utility of the MAGI principles as a metric for the creation, evaluation, and comparison of graph interpretation measures.

Contribution 2: Design and refine a measure

Creation of the Draw Datapoints on Graphs (DDoG) measure

Our second contribution is the use of the MAGI principles to create a readout-based measure of graph comprehension. We call this the Draw Datapoints on Graphs (DDoG) measure. Figure 4 illustrates the DDoG measure's basic approach: it uses a graph as the stimulus (Figure 4a); the graph is interpreted (4b); and this interpretation is then recorded by sketching a version of the graph (4c). The reference-frame of the graph is retained throughout.

Figure 5 shows a more detailed schematic of the DDoG procedure used in the present study, which expresses the six MAGI principles (Table 1) as follows: The instructions (Figure 5b) are succinct, concrete, and task-directed (Limited instructions). The graph stimulus (5c) and its caption (5d) are taken directly from

←

stimulus shown in c outlined in purple (we refer to this stimulus below as AGE) (*Expressive freedom, Information richness, Limited mental transformations*). Representative AGE stimulus graph (c) sketches are outlined in purple. Representative readouts demonstrate (e) the correct Bar-Tip Mean response and (f) the incorrect Bar-Tip Limit (BTL) response.

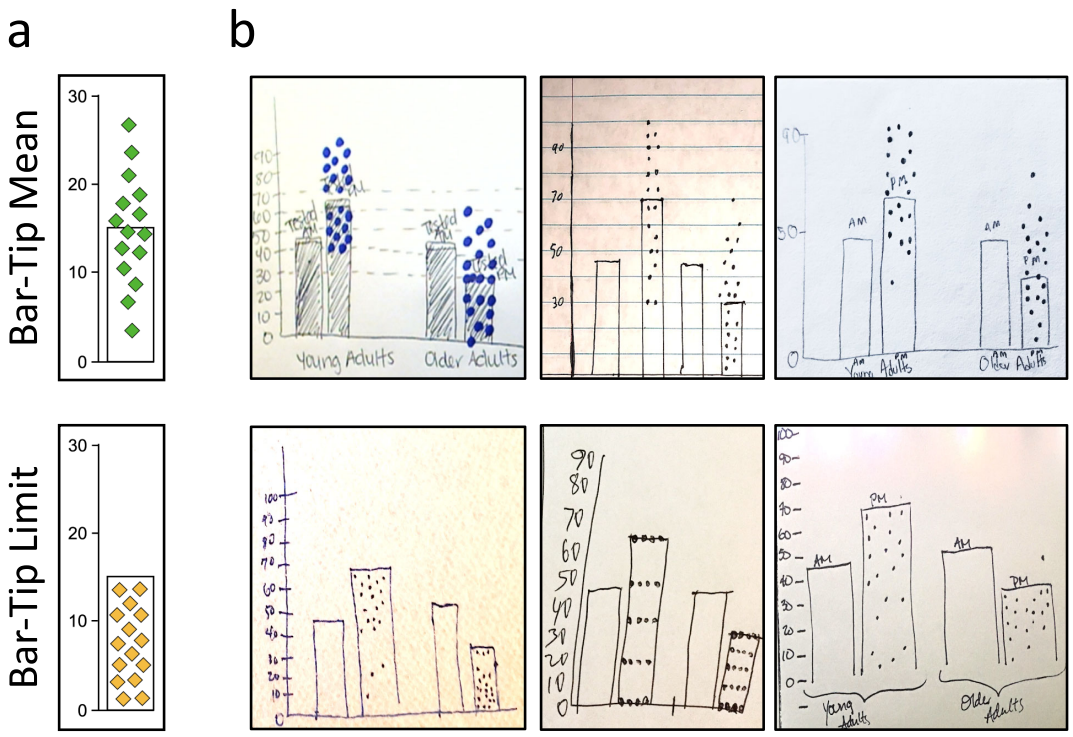

Figure 6. **The difference between Bar-Tip Mean and Bar-Tip Limit thinking is easily observable.** (a) Cartoon and (b) readout examples of the two common DDoG measure responses: the correct Bar-Tip Mean response (top) and the incorrect Bar-Tip Limit (BTL) response (bottom). These readouts were all drawn for the same stimulus graph, which we refer to as AGE (see Figure 11).

a textbook (*Ecological validity*). The drawing-based response medium (5a, e, f) provides the participant with flexibility to record their thoughts in a relatively unconstrained manner (*Expressive freedom*). The matched format between the stimulus (5c) and the response (5e, 5f, purple outline) minimizes the mental transformations required to record an interpretation (4b) (*Limited mental transformations*). The extent of the response—160 drawn datapoints, each on a continuous scale (5e, 5f)—provides a high-resolution window into the thought-process (*Information richness*). Finally, the spatial and numerical concreteness of the readout (5e, 5f), allows easy comparison to a variety of logical and empirical benchmarks of ground-truth (*Ground-truth linkage*). In these ways, the DDoG measure demonstrates each of the MAGI principles.

While we use the DDoG measure here to record interpretations of mean bar graphs, it can easily be applied to any other abstract-graph; that is, any graph that replaces individual-level data with summary-level information.

Contribution 3: Apply the measure

Using the DDoG measure to test the accessibility of mean bar graphs

From our earliest DDoG measure pilots looking at mean bar graph accessibility, a substantial subset of participants drew a very different distribution of data than the rest (Figure 6, Figures 5e & f, and Figure 4c).

When instructed to draw underlying datapoints for mean bar graphs (Figure 5b), most participants drew a (correct) distribution with datapoints balanced across the bar-tip (Figure 6ab, top). A substantial subset, however, drew most, or all, datapoints within the bar (Figure 6ab, bottom), treating the bar-tip as a limit. This minority response would have been correct for a count bar graph (as shown in Figure 2d), but it was severely incorrect for a mean bar graph (Figure 2b). We named this incorrect response pattern the Bar-Tip Limit (BTL) error.

The dataset that we examine in Results contains over three orders of magnitude more drawn datapoints than the two early-pilot drawings shown in Figure 4c: 44,000 datapoints, drawn in 551 sketches, by 149 participants. This far larger dataset yields powerful insights into the BTL error by establishing its categorical nature, high prevalence, stability within individuals, likely developmental influences, persistence despite thoughtful responding, and independence from foundational knowledge and graph content. Yet, in a testament to the DDoG measure's incisiveness, the two early pilot drawings shown in Figure 4c already substantially convey all three of the core contributions of the present paper: a severe error in mean bar graph interpretation (the BTL error), saliently revealed by a new readout-based measure (the DDoG measure) that

collects high-quality graph interpretations from experts and nonexperts alike (using the MAGI principles).

Moreover, it takes only a few readouts to move beyond mere identification of the BTL error, toward elucidation of an apparent mechanism. Considering just the readouts we have seen so far, observe the stereotyped nature of the BTL error response across participants (Figure 6b), graph form, and content (Figures 5e & f). Notice the similarity of this stereotyped response to the Bar-Tip Limit data shown in Figure 2d. These clues align perfectly with a mechanism of conflation, where mean bar graphs are incorrectly interpreted as count bar graphs (Figure 2).

In retrospect, the stage was clearly set for such a conflation. The use of one graph type (bar) to represent two fundamentally different types of data (counts and means) makes it difficult to visually differentiate the two (Figure 2) and sets up the inherent cognitive conflict that is the source of this paper's title ("Two graphs walk into a bar"). Additionally, the physical stacking metaphor for count bar graphs (Figure 1) makes a Bar-Tip Limit interpretation arguably more straightforward and intuitive; and the common use of count bar graphs as a curricular launching pad to early education could easily solidify the Bar-Tip Limit idea as a familiar, well-worn path for interpretation of bar graphs (Figure 1).

Yet, remarkably, none of the many prior theoretical and empirical critiques of mean bar graphs considered that such a conflation might occur (Tufte & Graves-Morris, 1983, p. 96; Wainer, 1984; Drummond & Vowler, 2011; Weissgerber et al., 2015; Larson-Hall, 2017; Rousselet, Pernet, & Wilcox, 2017; Pastore, Lionetti, & Altoe, 2017; Weissgerber et al., 2019; Vail & Wilkinson, 2020; Newman & Scholl, 2012; Correll & Gleicher, 2014; Pentoney & Berger, 2016; Okan et al., 2018). The concrete, granular window into graph interpretation provided by DDoG measure readouts, however, elucidates both phenomenon and apparent mechanism with ease.

Related works

Here we embed our three main contributions—the Measurement of Abstract Graph Interpretation (MAGI) principles, the Draw Datapoints on Graphs (DDoG) measure, and the Bar-Tip Limit (BTL) error—into related psychological, vision science, and data visualization literatures.

Literature related to the DDoG measure

Classification of the DDoG measure

A core design feature of the DDoG measure is its "elicited graph" measurement approach whereby a graph is produced as the response. This approach can, in turn, be placed within two increasingly broad categories: readout-based measurement and graphical elicitation measurement. Like elicited graph measurement, readout-based measurement produces a detailed, visuospatial record of thought; yet its content is broader, encompassing nongraph products such as the clock discussed above (Figure 3). Graphical elicitation, broader still, encompasses any measure with a visuospatial response, regardless of detail or content (Hullman et al., 2018; a similar term, graphic elicitation, refers to visuospatial *stimuli,* not *responses*, Crilly et al., 2006). Having embedded the DDoG measure within these three nested measurement categories—elicited graph, readout-based, and graphical elicitation—we next use these categories to distinguish it from existing measures of graph cognition.

Vision science measures

Vision science has long been an important source of graph cognition measures (Cleveland & McGill, 1984; Heer & Bostock, 2010), and recent years have seen an accelerated adoption of vision science measures in studies of data visualization (hereafter datavis). Of particular interest is a recent tutorial paper by Elliott and colleagues (2020) that cataloged behavioral vision science measures with relevance to datavis. Despite its impressive breadth—laying out nine measures with six response types and 11 direct applications (Elliott et al., 2020)—not a single elicited graph, readout-based, or graphical elicitation approach was included. This fits with our own experience that such methods are rarely used in vision science.

This rarity is even more surprising given that the classic drawing tasks that helped to inspire our current readout-focused approach are well known to vision scientists (Landau et al., 2006; Smith, 2009; Figure 3); indeed, they are commonly featured in Sensation and Perception textbooks (Wolfe et al., 2020). Yet usage of such methods within vision science has been narrow; restricted primarily to studies of extreme deficits in single individuals (Wolfe et al., 2020).

It is unclear why readout-based measurement is uncommon in broader vision science research. Perhaps the relative difficulty of structuring a readout-based task to provide a consistent quantitative measurement across multiple participants has been prohibitive. Or maybe reliable collection of drawn samples from a distance was logistically untenable until the recent advent of smartphones with cameras and accessible image-sharing technologies. We hypothesize that factors such as these may have limited the use of readout-based measurement in vision science, and we believe the proof-of-concept provided by the MAGI principles (Table 1) and the DDoG measure (Figure 4) supports their broader use in the future.

Although the DDoG measure shares its readout-based approach with classic patient drawing tasks (Figure 3), it differs from them in a subtle but important way. Patient drawing tasks are typically stimulus-indifferent, using stimuli (e.g., clocks) as mere tools to reveal the integrity of a broad mental function (e.g., spatial attention). The DDoG measure, in contrast, seeks to probe the interpretation of a specific stimulus (*this* bar graph) or stimulus type (bar graphs in general). The focus on how the stimulus is interpreted, rather than on the integrity of a stimulus-independent mental function, distinguishes the DDoG measure from classic patient drawing tasks.

Graphical elicitation measures

Moving beyond vision science, we next examine the domain of graphical elicitation for DDoG-related measures. A birds-eye perspective is provided by a recent review of evaluation methods in uncertainty visualization (Hullman et al., 2018). Uncertainty visualization's study of distributions, variation, and summary statistics makes it an informative proxy for datavis research related to our work. Notably, that review called the method for eliciting a response "a critical design choice." Of the 331 measures found in 86 qualifying papers, from 11 application domains (from astrophysics to cartography to medicine), only 4% elicited any sort of visuospatial or drawn response, thus qualifying as graphical elicitation (Hullman et al., 2018). None elicited either a distribution or a full graph; thus, none qualified as using an elicited graph approach. Further, though some studies elicited markings on a picture (Hansen et al., 2013) or map (Seipel & Lim, 2017), or movement of an object across a graph (Cumming, Williams, & Fidler, 2004), none sought to elicit a detailed visuospatial record of thought; thus, none qualified as using a readout-based approach. Survey methods such as multiple choice, Likert scale, slider, and text entry were the most common response elicitation methods—used 66% of the time (Hullman et al., 2018).

While readouts are rare in both vision science and datavis, there exists a broader graphical elicitation literature in which readouts are more common and elicited graphs are not unheard of (Jenkinson, 2005; O'Hagan et al., 2006; Choy, O'Leary & Mengersen, 2009). Yet still, these elicited responses tend to differ from our work in two key respects. First, they are typically used exclusively with experts, whereas the DDoG measure prioritizes general usage. Second, their aim is typically to document preexisting knowledge, and they therefore use stimuli as a catalyst or prompt, rather than as an object of study. The DDoG measure, in contrast, holds the stimulus as the object of study: we want to know how the graph was interpreted. The DDoG measure is therefore distinctive even in the broadly defined domain of graphical elicitation.

A small subset of graphical elicitation studies do elicit a visuospatial response, drawn on a graph, from a general population. There remains, however, a core difference in the way the DDoG measure utilizes the elicited response. DDoG uses the elicited response as a measure (dependent variable). The other studies, in contrast, use it as a way to manipulate engagement with, or processing of, a numerical task (as an independent variable; Stern, Aprea, & Ebner, 2003; Natter & Berry, 2005; Kim, Reinecke, & Hullman, 2017a; Kim, Reinecke, & Hullman, 2017b). While it is possible for a manipulation to be adapted into a measure, the creation of a new measure requires time and effort for iterative, evidence-based refinement and validation (e.g., Wilmer, 2008; Wilmer et al., 2012; Degutis et al., 2013). The DDoG measure is therefore distinctive in having been created and validated explicitly as a measure.

Frequency framed measures

Another way in which the DDoG measure may be distinguished from previous methods is in its use of a frequency framed response. Frequency framing is the communication of data in terms of individuals (e.g., "20 of 100 total patients" or "one in five patients") rather than percentages or proportions (e.g., "20% of patients" or "one fifth of patients"). The DDoG measure collects frequency-framed responses in terms of individual values.

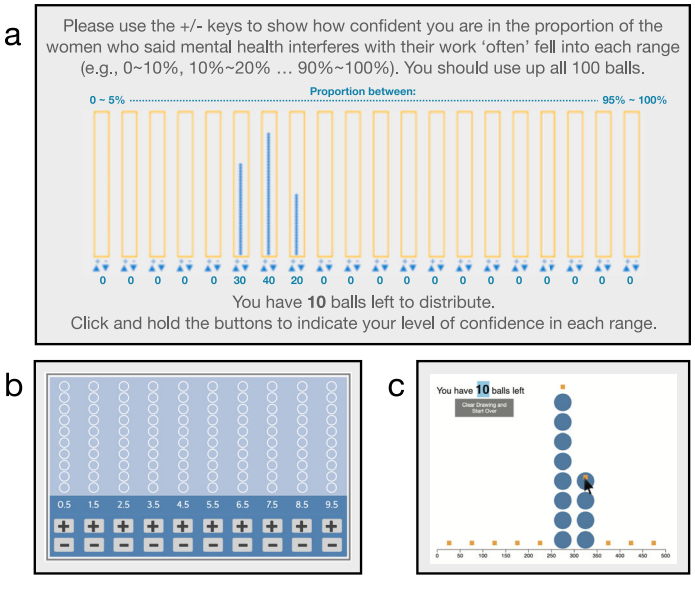

Figure 7. **The balls-and-bins approach.** Three recent adaptations of the balls-and-bins approach to eliciting probability distributions. This approach was originally developed by Goldstein and Rothschild (2014). The sources of these adaptations of balls-and-bins are: (a) Kim, Walls, Kraft & Hullman, 2019; (b) Andre, 2016; (c) Hullman et al. 2018.

Frequency framing has long been considered a best practice for data communication to nonexperts (Cosmides & Tooby, 1996), yet it remains uncommon in the measurement of graph cognition. Illustratively, as of 2016, only a small handful of papers in uncertainty visualization had utilized frequency framing (Hullman, 2016); and in *all cases*, frequency framing was used in the instructions or stimulus rather than in the measured response (e.g., Hullman, Resnick, & Adar, 2015; Kay et al., 2016).

Two other, more recent, datavis studies have elicited frequency framed responses (Hullman et al., 2017; Kim, Walls, Krafft & Hullman, 2019). These studies were geared toward nonexperts and used the so-called balls-and-bins paradigm, originally developed by Goldstein and Rothschild (2014), where balls are placed in virtual bins representing specified ranges (Figure 7). The studies aimed to elicit beliefs about sampling error around summary statistics. This usage contrasts with our aim of eliciting direct interpretation of the raw, individual-level, nonaggregated data that produced a mean.

This difference in usage joins more substantive differences, captured by two MAGI principles: *Limited mental transformations* and *Expressive freedom* (Table 1). While the current implementation of balls-and-bins requires spatial, format, and scale translations between stimulus and response, the DDoG measure achieves *Limited mental transformations* by applying a stimulus-matched response: imagined datapoints are drawn directly into a sketched version of the graph itself. Similarly, while balls-and-bins predetermines key aspects of the response such as bin widths, bin heights, ball sizes, numbers of bins, and the range from lowest to highest bin, the DDoG measure achieves *Expressive freedom* by allowing drawn responses that are constrained only by the limits of the page. Further, the lack of built-in tracks or ranges that constrain or suggest data placement reduces the risk of the "observer effect," whereby constraints or suggestions embedded in the measurement procedure itself alter the result.

Defining thought inaccuracies

Textual definitions of errors, biases, and confusions

To discuss the literature regarding this project's third contribution—identification of the Bar-Tip Limit (BTL) error—it is helpful to distinguish three different types of inaccurate thought processes: errors, biases, and confusions. We will do this first in words, and then numerically.

- Errors are "mistakes, fallacies, misconceptions, misinterpretations" (Oxford University Press, n.d.). They are binary, categorical, qualitative inaccuracies that represent wrong versus right thinking.
- Biases are "leanings, tendencies, inclinations, propensities" (Oxford University Press, n.d.). They are quantitative inaccuracies that exist on a graded scale or spectrum. They exhibit consistent direction but varied magnitude.
- Confusions are "bewilderments, indecisions, perplexities, uncertainties" (Oxford University Press, n.d.). They indicate the absence of systematic thought, resulting, for example, from failures to grasp, remember, or follow instructions.

Numerical/visual definitions of errors, biases, and confusions

Let us now consider how each type of inaccurate thought process would be instantiated in data. Such depictions can act as visual definitions to compare with real datasets. In Figure 8, patterns of data distributions representing correct thoughts (8a), systematic errors (8b), biases (8c), and confusions (8d), are illustrated as frequency curves. They are numerical formalizations of the textual definitions above, providing visual cues for differentiating these four thought processes in a graphing context.

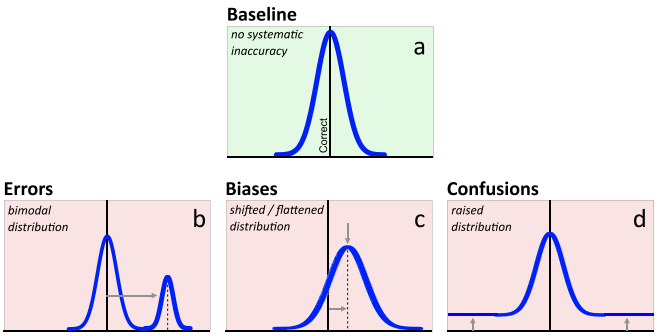

Figure 8. **Three types of inaccurate graph interpretation.** Prototypical response distributions provide visual definitions for three types of inaccuracy. Top row: (a) Baseline response pattern (green) indicates no systematic inaccuracy. Responses cluster symmetrically around the correct response value, with imprecision in task input, processing, and output reflected in the width of the spread around that correct response. Bottom row: Light red backgrounds illustrate the presence of inaccurate responses. Gray arrows indicate the major change from the baseline response pattern for each type of inaccurate response. (b) Systematic error responses form their own distinct mode. (c) Systematic bias responses shift and/or flatten the baseline response distribution. (d) Confused responses—expressed as random, unsystematic responding—are uniformly distributed, thus raising the tails of the baseline distribution to a constant value without altering the mean, median, or mode.

Figure 8a, represents a baseline case: a data distribution (blue curve) that is characteristic of generally correct interpretation, absent systematic inaccuracy. This baseline case prototypically produces a symmetric distribution, with the mean, median, and mode all corresponding to a correct response. Normal imprecision, from expected variation in stimulus input, thought processes, or response output, is reflected by the width of the distribution. (If everyone answered with 100% precision, the "distribution" would be a single stack on the correct answer).

Figure 8b illustrates a subset of systematic errors within an otherwise correct dataset. In distributions, an erroneous, categorically inaccurate subset prototypically presents as a separate mode, additional to the baseline curve. Errors siphon responses from the competing baseline distribution, which remains clustered around the correct answer, retaining the correct modal response. The hallmark of an error is, therefore, a bimodal distribution of responses whose prevalence and severity are demonstrated, respectively, by the height and separation of the additional mode.

Figure 8c illustrates a subset of systematic biases within an otherwise correct dataset. Given their graded nature, a biased subset tends to flatten and shift the baseline distribution of responses. The prototypical bias curve demonstrates a single, shifted mode. To the extent that a bias reflects a pervasive, relatively inescapable aspect of human perception, cognition, or culture, there will be relatively less flattening, and more shifting, of the response distribution.

Finally, Figure 8d illustrates a subset of confusions within an otherwise correct dataset. A confused subset tends to yield random responses, which lift the tails of the response distribution to a constant, nonzero value as responses are siphoned from the competing baseline distribution.

A rough but informative way to distinguish between erroneous, biased, and confused thinking is to simply compare an observed distribution of responses directly to the blue curves in Figure 8. This visual approach may be complemented with an array of analytic approaches that provide quantitative estimates of the degree of evidence for a particular type of inaccurate thinking (e.g., Freeman & Dale, 2013; Pfister et al., 2013; Zhang & Luck, 2008). In Results, we will use both visual and analytic approaches.

The success of any approach rests upon the precision and accuracy of measurement: as measurement quality is reduced, it becomes increasingly difficult to detect clear thought patterns reflected in the data. Our results will show unequivocal bimodality in the collected data, supporting both the presence of an erroneous graph interpretation (the BTL error) and the precision and accuracy of the DDoG measure.

Literature related to the BTL error

Probability rating scale results

In examining prior work related to our discovery of the BTL error, studies by Newman and Scholl (2012), Correll and Gleicher (2014), Pentoney and Berger (2016), and Okan and colleagues (2018) are seminal and contain the most reproducible prior evidence against mean bar graph accessibility. Their core result is an asymmetry recorded via a probability rating scale: the average participant, when shown a mean bar graph, rated an inside-of-bar location for data somewhat more likely than a comparable outside-of-bar location (~1 point on a 9-point probability rating scale, Figure 9a).

The original paper, by Newman and Scholl (2012), reported five replications of this asymmetry, including in-person and online testing, varied samples (ferry commuters, Amazon Mechanical Turk recruits, Yale undergraduates), and conditions that ruled out key numerical and spatial confounds. Five independent research groups have since replicated and extended this finding. The first was Correll and Gleicher (2014),

a. Rating-scale data b. DDoG measure data

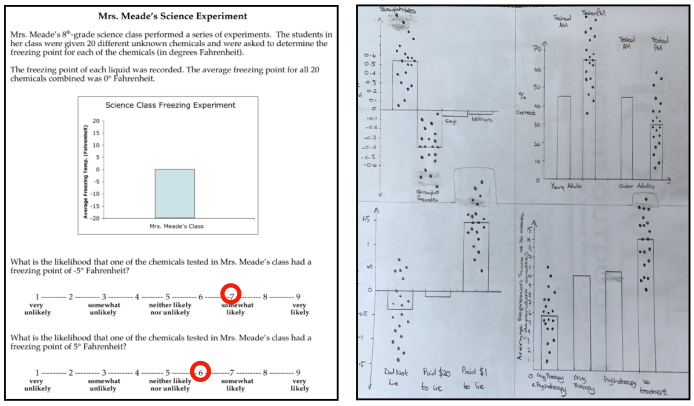

Figure 9. **A side-by-side comparison of a probability rating scale and a DDoG measure response.** DDoG measure's more concrete, detailed, visuospatial (aka readout-based) approach may have contributed to the ease with which it identified the BTL error and its apparent conflation mechanism, which were missed by prior studies using the probability rating scale approach. (a) The probability rating scale response sheet provided to each participant in the original report of asymmetry by Newman and Scholl (2012, Study 5). These Likert-style scales, anchored by colloquial English words (from "very unlikely" to "very likely"), were used to characterize the likelihood that individual values occurred at two specific y-axis values, one within the bar (−5) and one outside the bar (+5). The red circles indicate the resulting mean ratings of 6.9 and 6.1 (from Study 5, Newman & Scholl, 2012). (b) The DDoG measure response sheet from an individual participant given four stimulus graphs and asked to sketch each graph along with 20 hypothesized datapoints for two specified bars on the graph.

who extended the result to viewers' predictions of future data, and who showed that graph types that were symmetrical around the mean (e.g., violin plots) did not produce the effect. Soon after, Pentoney and Berger (2016) found this asymmetry present for a bar graph with confidence intervals, but absent when the bar was removed, leaving only the confidence intervals; this replicated Correll and Gleicher's (2014) finding that the effect requires the presence of a traditional bar graph. Okan and colleagues (2018) also replicated both the core asymmetry (this time using health data) and the finding that nonbar graphs did not produce the effect; they additionally found that the asymmetry, counterintuitively, increased with graph literacy. Finally, Godau, Vogelgesang, and Gaschler (2016) and Kang and colleagues (2021) showed that the asymmetry remains in aggregate—carrying through to judgments of the grand mean of multiple bars.

Newman and Scholl's (2012) hypothesized mechanism was a well-known perceptual bias, whereby in-object locations are processed slightly more effectively than out-of-object locations. The studies that Newman and Scholl (2012) cited for this perceptual bias, for example, had on average 4% faster and 0.6% more accurate responses for in-object stimuli relative to out-of-object stimuli (Egly et al., 1994; Kimchi, Yeshurun, & Cohen-Savransky, 2007; Marino & Scholl, 2005). While subtle, this bias was believed to reflect a fundamental aspect of object processing (Newman & Scholl, 2012), and it was therefore assumed to be pervasive and inescapable. Newman and Scholl's (2012) hypothesis was that this "automatic" and "irresistible" perceptual bias had produced a corresponding "within-the-bar bias" in graph interpretation.

Probability rating scale comparison

At the end of Results, we will compare our detailed findings, obtained via our DDoG measure, to those of the probability rating scale studies reviewed just above. Here, we compare the more general capabilities of the two measurement approaches. Figure 9 shows side-by-side examples of a probability rating scale response from Newman and Scholl's (2012) Study 5 (Figure 9a), and a DDoG measure readout (Figure 9b). Using the MAGI principles (Table 1) as a basis for comparison, we can see key differences in: *Limited mental transformations, Ground-truth linkage, Information richness*, and *Expressive freedom.*

The MAGI principle of *Limited mental transformations* aims to facilitate an accurate readout of graph interpretation. As discussed in Frequency framed measures even a mental transformation as seemingly trivial as a translation from a ratio (one in five) to a percentage (20%) can severely distort a person's thinking (Cosmides & Tooby, 1996). The DDoG measure limits mental transformations via its stimulus-matched response (Figures 4 and 9b). Absent such a matched response, the mental transformations required by a measure may limit its accuracy.

In the example from Newman and Scholl (2012), Study 5 (Figure 9a), several mental transformations are needed for probability rating scale responses; among them, translation from numbers (−5 and 5) to graph locations (in or out of the bar, and to what degree in or out), from graph locations to probabilities (represented in whatever intuitive or numerical way a person's mind may represent them), from probabilities to the rating scale's English adjectives (e.g. "somewhat," "very"), and, separately, from the vertical scale on the graph's y-axis to the horizontal rating scale. While it is possible that some of these mental transformations are accomplished effectively and relatively uniformly for most participants, our reanalysis below of the probability rating scale studies (Results: A reexamination of prior results) will suggest that their presence contributes a certain amount of irreducible "noise" (inaccuracy) to the response.

With regard to *Ground-truth linkage*, it is straightforward to evaluate a DDoG measure readout (Figure 9b) relative to multiple logical and empirical benchmarks of ground-truth. Logical benchmarks include, for example, "Is the mean of the drawn datapoints at the bar-tip?" Empirical benchmarks include "How closely does the drawn data reflect the [insert characteristic] in the actual data?" In contrast, there is no logically or empirically correct (or incorrect) response for the probability rating scale measure (Figure 9a). Even discrepant rated values for in-bar versus out-of-bar locations—though suggestive of inaccurate thinking—could be accurate in a case of a skewed raw data distribution, which could plausibly lead to more raw data in-bar than out-of-bar (or to the opposite). The probability rating scales therefore lack a conclusive *Ground-truth linkage*. DDoG measure responses, in contrast, specify real numerical and spatial values with a concreteness that is easier to compare to a variety of ground-truths.

The DDoG measure's precision is most directly supported by its *Information richness*: in the present study, each participant produces 160 drawn datapoints on a continuous scale. In contrast, the probability rating scale shown in Figure 9a yields only two pieces of information (the two ratings) per participant. Granted, the relative information-value of a single integer rating, versus a single drawn datapoint, is not easily compared. Further, as we will see below, multiple ratings at multiple graph locations can increase precision of the probability rating scale measure (Results: A reexamination of prior results). Nevertheless, it would be difficult to imagine a case where a probability rating scale measure yielded more information richness, and, in turn, higher precision, than the DDoG measure.

	Principle	Execution	DDoG measure	Balls-and-bins measure	Probability rating scale measure
Facilitates *general usage* and *valid measurement*	***Expressive freedom***	Employ a flexible response medium, such as drawing, that allows relatively unconstrained recording of thought processes.	**Capable**	**Incapable** Constrained by answer-bin format.	**Incapable** Constrained by rating scale format.
	Limited instructions	Keep instructions succinct, concrete, and task-directed; minimize terms, concepts, or task steps that require explanation.	**Capable**	**Capable**	**Capable**
	Limited mental transformation	Elicit a stimulus-matched response: a response in the same basic format as the stimulus.	**Capable**	**Capability limited** by multiple mental transformations between graph and balls-and-bins answer format.	**Capability limited** by multiple mental transformations between graph and rating-scale answer format.
Facilitates *valid measuerment*	***Ground-truth linkage***	Select graph stimuli and responses for which logically and/or empirically correct and incorrect answers exist.	**Capable**	**Capable**	**Effectively incapable** due to the technical impossibility of demonstrating incorrectness of single or paired responses.
	Ecological validity	Excerpt graphs for use as stimuli from real-world sources, relatively unchanged and appropriately contextualized.	**Capable**	**Capable**	**Capable**
	Information richness	Elicit high-bandwidth responses, i.e., responses that contain numerous, dense, continuous, and/or multidimensional pieces of information.	**Capable**	**Capability substantial** but limited, by number and range of bins.	**Capability limited** by increment of scale and number of ratings per-participant.

Table 2. **Using MAGI principles to compare graph interpretation measures.**

Finally, comparing *Expressive freedom* between the two measures, the rating scale measure constrained each response to a nine-point integer scale. On this scale, responses are led and constrained to a degree that disallows many types of unexpected responses and limits the capacity of distinct graph interpretations to stand out from each other. These constraints contrast sharply with the flexibility of the drawn readout that the DDoG measure produces.

The probability rating scale and the DDoG measure therefore differ in terms of four separate MAGI principles (Table 1): *Limited mental transformations, Ground-truth linkage, Information richness,* and *Expressive freedom*.

MAGI principles used as a metric

We have used the MAGI principles above to highlight notable differences between the DDoG measure and specific implementations of two other measurement approaches: balls-and-bins (Figure 7) and probability rating scale (Figure 9a). However, the question of whether these differences are integral to the measure-type, or restricted to specific implementations, remains. Table 2 models the use of the MAGI principles as a vehicle to compare, in a more structured way, potentially integral features and limitations of these same three graph interpretation methods.

We believe that all three measures share a capability for *Limited instructions* and *Ecological validity*, though choices specific to each implementation may affect whether the principles are actually expressed (see Hullman et al. (2017) and Kim, Walls, Kraft and Hullman (2019) for implementations of the balls-and-bins method, and see Newman and Scholl (2012), Correll and Gleicher (2014), Pentoney and Berger (2016), and Okan and colleagues (2018) for implementations of probability rating scales). In contrast, the three measures appear to differ more or less intrinsically in their approaches to *Expressive freedom, Limited mental transformations*, and (to a lesser extent) *Information richness*. That said, measure development is an inherently iterative, dynamic process, and what seems intrinsic at one point in time can sometimes shift as development progresses.

Because the DDoG measure and MAGI principles were built around each other and refined in parallel, it makes sense that they are closely aligned in their optimization for assessment of abstract-graph interpretation, with a focus on general usage and valid measurement. Recognizing that measures with different aims and motivations may be best suited for different tasks, this use of MAGI principles, in table form (Table 2), provides a model of its utility for comparison and targeting of measures with a similar set of aims and motivations.

Summary of related works

In the four subsections of Related works above, we first documented and sought to better understand the relative rarity of elicited-graph, readout-based, graphical elicitation, and frequency-framed measurement in studies of graph cognition, while arguing for the value of greater usage of elicited-graph and readout-based measurement in particular. We next distinguished—in both written/verbal and

numerical/visual form—three distinct types of inaccurate graph interpretation: errors, biases, and confusions. Third, we examined key results and methods from prior studies of mean bar graph inaccessibility. And, finally, we provided an illustrative example of how the Measurement of Abstract Graph Interpretation (MAGI) principles can be used to compare and contrast relevant measures.

Methods

The current investigation has two parts: (A) examine the accessibility of a set of ecologically valid mean bar graphs in an educationally diverse sample via our new Draw Datapoints on Graphs (DDoG) measure (sections Participant Data Collection through Define the Average) and (B) assess the relative frequency of mean versus count bar graphs across educational and general sources (section Prevalence of Mean Versus Count Bar Graphs). Included in the Methods sections devoted to "A" are detailed specifications for the current implementation of the DDoG measure, as well as the thinking behind that implementation, as a guide to future DDoG measure usage.

Participant data collection

Piloting

The development of the DDoG measure was highly iterative. Early piloting was performed at guest lectures, lab meetings, and in college classrooms at various levels of the curriculum by author JBW, and in several online data collection pilot studies by both authors. Early wording differed from the more refined wording used in the present investigation. Yet even the earliest pilots produced the core result of the current investigation (Figure 4): a common, severely inaccurate interpretation of mean bar graphs that we call the Bar-Tip Limit (BTL) error. The DDoG measure thus appears—at least in the context of the BTL error—robust to fairly wide variations in its wording. A constant through all DDoG measure iterations, however, was succinct, concrete, task-directed instructions: what came to be expressed as the *Limited instructions* MAGI principle (Table 1).

Recruitment

Data collection was completed remotely using Qualtrics online survey software. There were neither live nor recorded participant-investigator interactions, to minimize priming, coaching, leading, or other forms of experimenter-induced bias (*Limited instructions* principle). Participants were recruited via the Amazon Mechanical Turk (MTurk), Prolific, and TestableMinds platforms. Because no statistically robust differences were observed between platforms, data from all three were combined for the analyses reported below. Participants were paid $5 for an expected 30 minutes of work; the median time taken was 33 minutes.

Procedure

Figure 10 provides a flowchart of the procedure. Participants (a) read an overview of the tasks along with time estimates; (b) read and recorded mean values from stimulus graphs; (c) (grey zone) completed the DDoG measure drawing task for stimulus graphs; (g) provided a definition for the average/mean value; and (h) reported age, gender, educational attainment, and prior coursework in psychology and statistics.

Foundational knowledge tasks

Graph reading task: Find the average/mean

Participants were asked to "warm up" by reading mean values from a set of mean bar graphs that were later used in the DDoG measure drawing task (wording: "What is the average (or mean) value for [condition]?") (Figure 10j). This warm-up served as a control, verifying that the participant was able to locate a mean value on the graph.

To allow for some variance in sight-reading, responses within a tolerance of two tenths of the distance between adjacent y-axis tick-marks were counted as correct (see Figure 11). The results reported in Results: Independence of the BTL error from foundational knowledge remain essentially unchanged for response-tolerances from zero (only exactly correct responses) through arbitrarily high values (all responses). No reported prevalence values dip below 19%.

Graph reading was correct for 93% of graphs. As a control for possible carryover/learning effects, 114 of the 551 total DDoG measure response drawings were completed without prior exposure to the graph (i.e., without a warm-up for that graph). No evidence of carryover/learning effects was observed.

Definition task: Define the average/mean

After completing the DDoG measure drawing task, as a control for comprehension and thoughtful responding, participants were asked to explain the concept of the average/mean. Most participants (112 of 190, or 64%) got the following version of the question: "From your own memory, what is an average (or mean)? Please just give the first definition that comes to mind. If you have no idea, it is fine to say that. Your definition does not need to be correct (so please don't look it up on the internet!)" For other versions of the question, see open data spreadsheet. No systematic

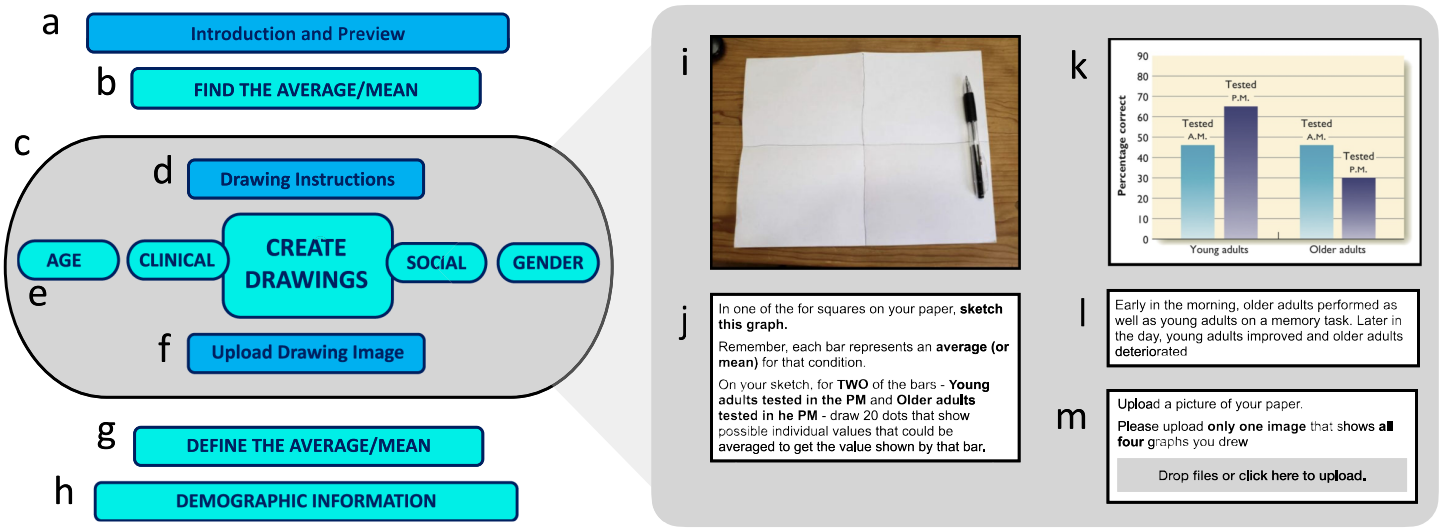

Figure 10. **Flowchart of study procedure.** The present study consisted of five main sections (a, b, c, g, h). Sections and subsections colored teal (b, e, g, h) produced data that were analyzed for this study. Subsections on the right (i, j, k, l, m) are an expansion of section c of the flowchart, showing (i) the drawing page, (j) the drawing instructions, (k) one of the four stimulus graphs, (l) the graph caption, and (m) the upload instructions. See Methods for further procedural details.

differences in results based on question wording were observed. Participants were provided a text box for answers with no character limit. Eighty-three percent of responses were correct, with credit being given for responses that were correct either conceptually (e.g., "a calculated central value of a set of numbers") or mathematically (e.g., "sum divided by the total number of respondents").

The Draw Datapoints on Graphs (DDOG) measure: Method

Selection of DDoG measure stimulus graphs

For the present investigation, four mean bar graph stimuli, shown in Figure 11, were taken from popular Introductory Psychology textbooks. This source of graphs was selected for three main reasons. First, Introductory Psychology is among the most popular undergraduate science courses, serving two million plus students per year in the United States alone (Peterson & Sesma, 2017); and this course tends to rely heavily on textbooks, with the market for such texts estimated at 1.2 to 1.6 million sales annually (Steuer & Ham, 2008). The high exposure of these texts gives greater weight to the data visualization choices made within them. Second, Introductory Psychology attracts students with widely varying interests, skills, and prior data experiences, meaning that data visualization best practices developed in the context of Introductory Psychology may have broad applicability to other contexts involving diverse, nonexpert populations. Third, given the relevance of psychological research to everyday life, inaccurate inferences fueled by Introductory Psychology textbook data portrayal could, in and of themselves, have important negative real-world impacts.

The mean bar graph stimuli were chosen to exhibit major differences in form, and to reflect diversity of content (Figure 11). They convey concrete scientific results, rather than abstract theories or models, so that understanding can be measured relative to the ground-truth of real data (*Ground-truth linkage*). Introductory texts were chosen because of their direct focus on conveying scientific results to nonexperts (*Ecological Validity*). Stimuli were selected to show "independent groups" (aka between-participants) comparisons because these comparisons are one step more straightforward, conceptually and statistically, than "repeated-measures" (aka within-participants) comparisons.

We refer to the four stimulus graphs by their independent variables: as AGE, CLINICAL, SOCIAL, and GENDER. The stimuli were selected, respectively, from texts by Kalat (2016), Gray and Bjorklund (2017), Grison and Gazzaniga (2019), and Myers and DeWall (2017), which ranked 7, 22, 3, and 1 in median Amazon.com sales rankings across eight days during March and April of 2019, among the 23 major textbooks in the Introductory Psychology market. Further details about these graphs are shown in Table 3.

Stimuli were chosen to represent both meaningful content differences (four areas about which individual participants might potentially have strong personal opinions), and form differences (differing relationship of bars to the baseline) to evaluate replication of results

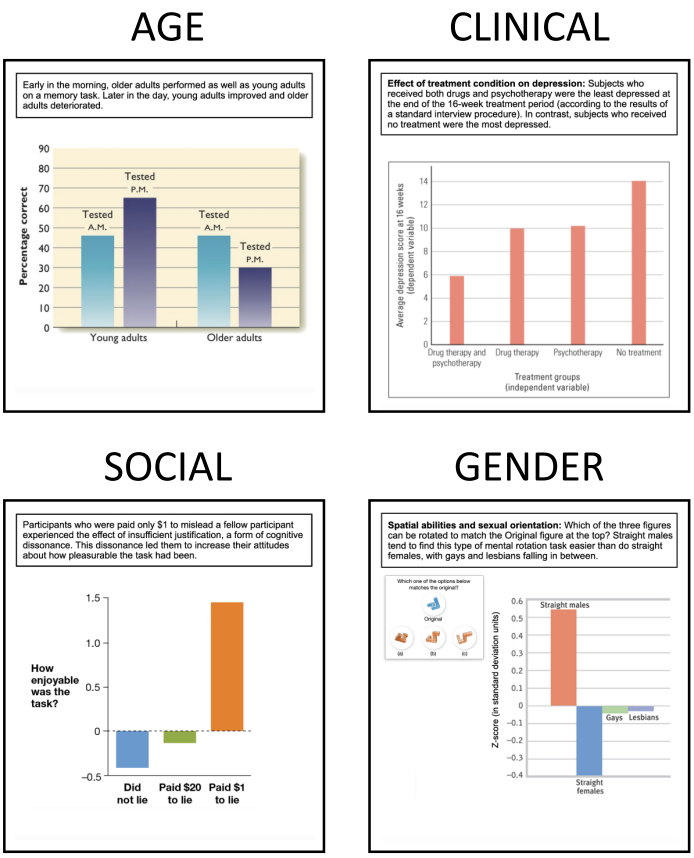

Figure 11. **The four stimulus graphs used in this study.** These stimulus graphs were taken from popular Introductory Psychology textbooks to ensure the direct real-world relevance of our results (see *ecological validity* MAGI principle). Figure legends were adapted as necessary for comprehensibility outside the textbook. Stimulus graph textbook sources are AGE: Kalat, 2016; CLINICAL: Gray and Bjorklund, 2017; SOCIAL: Grison & Gazzaniga, 2019; and GENDER: Myers & DeWall, 2017.

across graphs despite differences that might reasonably be expected to change graph interpretation.

The specific form difference we examined was the distinction between unidirectional bars (all bars emerge from the same side of the baseline) and bidirectional bars (bars emerge in opposite directions from the baseline). Two of the four graphs, AGE and CLINICAL were unidirectional (Figure 11, top), and the other two, SOCIAL and GENDER, were bidirectional (Figure 11, bottom). While unidirectional graphs were more common than bidirectional graphs in the surveyed textbooks, two examples of each were selected to set up an internal replication mechanism.

Replication of results could potentially be demonstrated across graphs in terms of mean values (the mean result of different graphs could be similar), or in terms of individual differences (an individual participant's result on one graph could predict their result on another graph). Both types of replication were observed.

Graph drawing task: The DDoG measure

The DDoG measure was designed using the MAGI principles (Table 1), and it was modeled after the patient drawing tasks mentioned above (Landau et al., 2006; Agrell & Dehlin, 1998). Ease of administration and incorporation into an online study were further key design considerations.

The *Expressive freedom* inherent in the drawing medium avoids placing artificial constraints on participant interpretation, and it has multiple potential benefits: it helps to combat the "observer effect," whereby a restricted or leading measurement procedure impacts the observed phenomenon; it lends salience to a consistent, stereotyped pattern (such as the Bar-Tip Limit response shown in Figures 2d, 4c (bottom), 5f, 6b (bottom), and 13 (right)); and it allows unexpected responses, which render the occasional truly inattentive or confused response (Figure 8d) clearly identifiable.

Twenty drawn datapoints per bar was chosen as a quantity small enough to avoid noticeable impacts of fatigue or carelessness, while remaining sufficient to gain a visually and statistically robust sense of the imagined distribution of values. Recent research suggests that graphs that show 20 datapoints enable both precise readings (Kay et al., 2016) and effective decisions (Fernandes et al., 2018).

Before beginning the drawing task, participants were instructed to divide a page into quadrants, and they were shown a photographic example of a divided blank page (Figure 10i). They were then presented with task instructions (Figure 10j) with one of the four bar graph stimuli and its caption (Figure 10k, 10l). This process—instructions, graph stimulus, caption—was repeated four total times, with the order of the four graph stimuli randomized to balance order effects such as learning, priming, and fatigue. After completing the four drawings on a single page, each participant photographed and uploaded their readout (Figure 10m) via Qualtrics' file upload feature. Few technical difficulties were reported, and only a single submitted photograph was unusable for technical reasons (due to insufficient focus). All original photographs are posted to the *Open Science Framework (OSF)*.

	AGE	CLINICAL	SOCIAL	GENDER
Text source	Kalat, 2016	Gray & Bjorklund, 2017	Grison & Gazzaniga, 2019	Myers & DeWall, 2017
Paper source	May et al, 1993	DiMascio et al, 1979	Festinger & Carlsmith, 1959	Rahman et al, 2004
Independent variable 1	old vs. young	therapy vs. none	lie vs. truth	male vs. female
Independent variable 2	early vs. late	drug vs. none	paid $1 vs. $20	gay vs. straight
Dependent variable	memory performance	self-reported depression symptoms	self-reported enjoyment	mental rotation performance

Table 3. **Stimulus graph sources and variables.**

DDoG measure data

Collection of DDoG measure readouts

One hundred ninety participants nominally completed the study. Of these, 149 (78%) followed the directions sufficiently to enable use of their drawings, a usable data percentage typical of online studies of this length (Litman & Robinson, 2020). Based on predetermined exclusion criteria, participant submissions were disqualified if most or all of their drawings were unusable for any of the following:

(1) Zero datapoints were drawn (18 participants).
(2) Datapoints were drawn but in no way reflected either the length or direction of target bars, thereby demonstrating basic misunderstanding of the task (16 participants).
(3) Drawings lacked labels or bar placement information necessary to disambiguate condition or bar-tip location (five participants).
(4) Photograph was insufficiently focused to allow a count (one participant).

Each of the 149 remaining readouts contained four drawn graphs, for 596 graphs total. Of these, 30 individual drawings were excluded for one or more of the above reasons, and 15 were excluded for the additional predetermined exclusion criterion of a grossly incorrect number of datapoints, defined as >25% difference from the requested 20 datapoints.

The remaining 551 drawn graphs (92.4% of the 596) from 149 participants were included in all analyses below. Ages of these 149 participants ranged from 18 to 71 years old (median 31). Reported genders were 96 male, 52 female, and one nonbinary. Locations included 26 countries and 6 continents. The most common countries were United Kingdom ($n = 58$), United States ($n = 34$), Portugal ($n = 8$), and Greece/Poland/Turkey (each $n = 4$).

Coding of DDoG measure readouts via BTL index

As a systematic quantification of datapoint placement in DDoG measure readouts, a Bar-Tip Limit (BTL) index was computed. The BTL index estimated the within-bar percentage of datapoints for each graph's two target bars by dividing within-bar datapoints (i.e., datapoints drawn on the baseline side of the bar-tip) by the total number of drawn datapoints for the two target bars and multiplying by 100. Written as a formula: [(# datapoints on baseline-side of bar-tips) / (total # datapoints)] × 100

The highest possible index (100) represents an image in which all drawn datapoints are on the baseline sides of their respective bar-tips (i.e., within the bar; a Bar-Tip Limit response). An index of 50 represents a drawing with equal numbers of datapoints on either side of the bar-tip (i.e., balanced distribution across the mean line, or a Bar-Tip Mean response). BTL index values in this study ranged from 27.5 (72.5% of points drawn outside of the bar) to 100 (100% of points drawn within the bar).

The straightforward coding procedure was designed to yield a reproducible, quantitative measure of datapoint distribution relative to the bar-tip, however, one ambiguity existed. Datapoints drawn directly on the bar-tip could reasonably be considered to either use the bar-tip as a limit (if the border is considered part of the object) or not (if a datapoint on the edge is considered outside the bar). This ambiguity was handled as follows: (1) if, at natural magnification, a drawn datapoint was clearly leaning toward one or the other side of the bar-tip, it was assigned accordingly, and (2) datapoints on the bar-tip for which no clear leaning was apparent were alternately assigned first inside the bar-tip, then outside, and continuing to alternate thereafter.

This procedure maximizes reproducibility by minimizing the need for the scorer to interpret the drawer's intent—and it was successful (see Reproducibility of BTL index coding procedure). Yet reproducibility may have come at the cost of some minor conservatism in quantifying drawings where the participant placed no datapoints beyond the bar-tip—arguably displaying a complete Bar-Tip Limit interpretation—yet placed some of the datapoints directly on the bar-tip. For example, if six of the 20 datapoints for each bar were placed on the bar-tip, the drawing would get a BTL index of only 85, when it arguably represented a pure BTL conception of the graph.

Reproducibility of BTL index coding procedure

Reproducibility of the BTL index coding was evaluated by comparing independently scored BTL indices of the two coauthors for a substantial subset of 201 drawn graphs. Figure 12 shows the data from this comparison where the x coordinate is JBW's scoring and the y coordinate is SHK's scoring of each drawing. The correlation coefficient between the two sets of ratings is extremely high ($r(199) = 0.991$, 95% CI [0.988, 0.993]). Additionally, the blue best-fit line is nearly indistinguishable from the black line representing hypothetically equivalent indices between the two raters.

The fact that the line of best fit is so closely aligned with the line of equivalence is important. In theory, even with a correlation coefficient close to 1.0, one rater might give ratings that are shifted higher/lower, or that are more expanded/compressed, compared to the other. This would show up as a best-fit line that was either vertically shifted relative to the line of equivalence, or of a different slope compared to the line

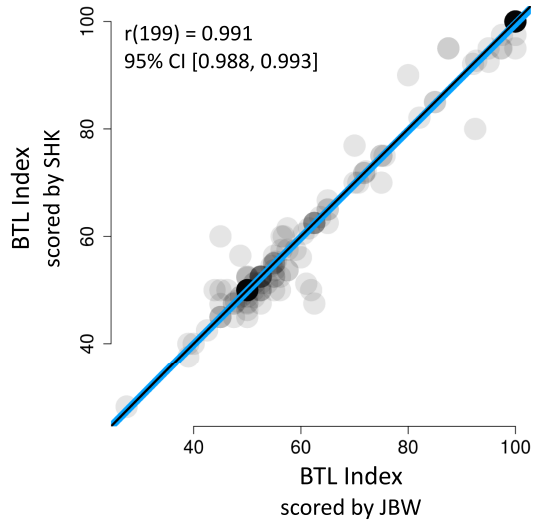

Figure 12. **Interrater reliability of Bar-Tip Limit (BTL) index coding demonstrates high repeatability of coding method.** Coauthors SHK and JBW independently computed the BTL index for a subset of 201 drawn graphs (of 551 total). Each dot represents both BTL indices for a given drawn graph (y-value from SHK, x-value from JBW). Semitransparent dots code overlap as darkness. The black line is a line of equivalence, which shows x = y for reference. The line of best fit is blue. The close correspondence of these two lines, and the close clustering of dots around them, demonstrate high repeatability of the BTL index coding procedure.

of equivalence. The absence of such a vertical shift or slope difference is therefore particularly strong evidence for repeatability. This strong evidence is echoed by a high interrater reliability statistic (Krippendorf's alpha) of 0.93 (Krippendorf, 2011); Krippendorf's alpha varies from 0 to 1, and a score of 0.80 or higher is considered strong evidence of repeatability (Krippendorf, 2011). These analyses therefore demonstrate that the BTL index coding procedure is highly repeatable, and that variations in coding—a potential source of noise that could theoretically constrain the capacity of a measure to precisely capture Bar-Tip Limit interpretation—can be minimized.

Establishing cutoffs and prevalence

For purposes of estimating the prevalence of the BTL error in the current sample, a BTL index cutoff of 80 was selected. This was the average result when five clustering methods (k-means, median clustering, average linkage between groups, average linkage within groups, and centroid clustering) were applied via SPSS to the 551 individual graph BTL indices (cutoffs were 75, 75, 80, 85, and 85, respectively). 80 was additionally verified as reasonable via visual inspection of the data shown in Figure 14.

Due to the strongly bimodal distribution of BTL index scores (see Results), only 14 of the 551 total drawings (2.5%) fell within the entire range of computed cutoffs (75 to 85). Therefore, though the computed confidence intervals around prevalence estimates, reported in Results, do not include uncertainty in selecting the cutoff, this additional source of uncertainty is small (at most, perhaps ± 1.25%).

For Pentoney and Berger's (2016) data, analyzed below in Prior work reflects the same phenomenon: Prevalence, the respective cutoffs from the same five clustering methods produced BTL error percentages of 19.5, 19.5, 19.5, 19.5, and 24.6, for an average percentage of 20.5.

Prevalence of bar graphs of means versus bar graphs of counts

The methods discussed in this subsection pertain to Results section Ecological exposure to mean versus count bar graph. In that investigation, we assessed the likelihood of encountering mean versus count bar graphs across a set of relevant contexts by tallying the frequency of each bar graph type from three separate sources: elementary educational materials accessible via Google Image searches, college-level Introductory Psychology textbooks, and general Google Images searches. It was hypothesized that these sources would provide rough, but potentially informative, insights into the relative likelihood of exposure to each bar graph type in these areas. The methods used for each source were:

The elementary education Google image search used the phrase: "bar graph [X] grade," with "first" to "sixth" in place of [X]. The first 50 grade-appropriate graph results, from independent internet sources, for each grade-level, were categorized as either count bar graph, mean bar graph, or other (histogram or line graph).

The college-level Introductory Psychology textbook count tallied all of the bar graphs of real data across the following eight widely-used Introductory textbooks: Ciccarelli & Berstein, 2018; Coon, Mitterer & Martini, 2018; Gazzaniga 2018; Griggs 2017; Hockenbury & Nolan, 2018; Kalat, 2016; Lilienfeld, Lynn & Namy, 2017; and Myers & DeWall, 2017. The bar graphs were then categorized as dealing with either means or counts (histograms excluded).

The general Google Image search used the phrase "bar graph." The first 188 graphs were categorized as count bar graphs, mean graphs, or were excluded for being not bar graphs (tables, histograms), or for containing insufficient information to determine what type of bar graph they were.

Plotting and evaluation of data from previous studies

In our reexamination of prior results, we compare previous study results to the present results. Data

is replotted from Newman and Scholl (2012) and Pentoney and Berger (2016) as Figure 20. For Newman and Scholl (2012), data from Study 5 is plotted categorically as "0" (no difference between in-bar and out-of-bar rating) versus "not 0" (positive or negative difference between in-bar and out-of-bar rating). The directionality of the latter differences, while desired, are not obtainable from the information reported in the original paper. For Pentoney and Berger (2016), the data plotted in their Figure 3 are pooled across the three separate conditions whose stimuli included bar graphs. A similar pattern of bimodality is evident in all three of those conditions. We used WebPlotDigitizer to extract the raw data from Pentoney and Berger's (2016) Figure 3.

Statistical tools and packages

The graphs shown in Figures 12, 16, and 18 were produced by ShowMyData.org, a suite of web-based data visualization apps coded in R by the second author using the R Shiny package. The statistics and distribution graphs shown in Figures 14, 17, 19, and 20 were produced via the ESCI R package. The bimodality simulations and graphs shown in Figure 15 were produced in base R, and the numerical analyses of bimodality using Hartigan's Dip Statistic (HDS) and Bimodality Criterion (BC) (Freeman & Dale, 2013; Pfister et al., 2013) were computed via the mousetrap R package. CIs on HDS, BC, and Cohen's d were computed with resampling (10,000 draws) via the boot R package. The clustering analyses that estimated BTL error cutoffs and prevalence were conducted via SPSS.

Results

This paper contains three separate sets of results: first, the core quantitative study of the BTL error (subsections Overview of core quantitative study through Educational and demographic correlates); second, the study of the prevalence of mean versus count bar graphs (subsection Ecological exposure to mean versus count bar graphs); and third, a reanalysis of prior studies of mean bar graph accessibility, the results of which suggest that the previously reported "within-the-bar bias" was actually not a bias at all, but instead was caused by the BTL error we report here (subsection A reexamination of prior results).

Overview of core quantitative study

From the Draw Datapoints on Graphs (DDoG) measure readouts shown in Figures 4, 5, and 6, one can already see the severe, stereotyped nature of the Bar-Tip Limit (BTL) error. It is additionally evident that this error is well-explained as a conflation of mean bar graphs with count bar graphs; and the effectiveness of the DDoG measure in revealing the BTL error is clear.

Our main study sought to add quantitative precision to these qualitative observations by testing a large, education-diverse sample. Examination of this larger sample confirms the BTL error's categorical nature, high prevalence, stability within individuals, persistence despite thoughtful, intentional responding, and independence from foundational knowledge and graph content. Together, these results demonstrate that mean bar graphs are subject to a common, severe error of interpretation that makes them a potentially unwise choice for accurately communicating empirical results, particularly where general accessibility is a priority, such as in education, medicine, or popular media. These results also establish the DDoG measure and MAGI principles as effective tools for gaining powerful insights into graph interpretation.

The core study utilizes a set of 551 DDoG measure response drawings, or readouts, produced by an educationally and demographically diverse sample of 149 participants. These readouts, of four mean bar graph stimuli that vary in form and content, are accompanied by an array of control and demographic measures (see Methods).

Initial observations of the Bar-Tip Limit error

Figure 13 shows the four stimulus graphs and, for each, an example of the two common response types submitted. The correct Bar-Tip Mean response has datapoints distributed across the bar-tip (Figure 13, center column), and the incorrect Bar-Tip Limit response has datapoints on only the baseline side of the bar-tip (Figure 13, right column).

The severe inaccuracy shown in the Bar-Tip Limit readouts, relative to the Bar-Tip Mean readouts, raises the possibility that the Bar-Tip Limit readouts represent a categorical difference in thinking: an error. Yet the few isolated examples shown in Figures 4, 5, 6, and 13 are insufficient to conclusively distinguish common errors of thinking from outliers. To clearly demonstrate a common error requires a nontrivial degree of bimodality in the response distribution (Related works: Defining thought inaccuracies).

The analysis that follows, evaluates and plots all 551 readouts on a continuous scale of datapoint placement (Figure 14). The Bar-Tip Limit response is revealed to be a separate mode, eliminating the possibility that early observations were outliers and identifying the incorrect Bar-Tip Limit response as a common error rather than a bias or confusion.

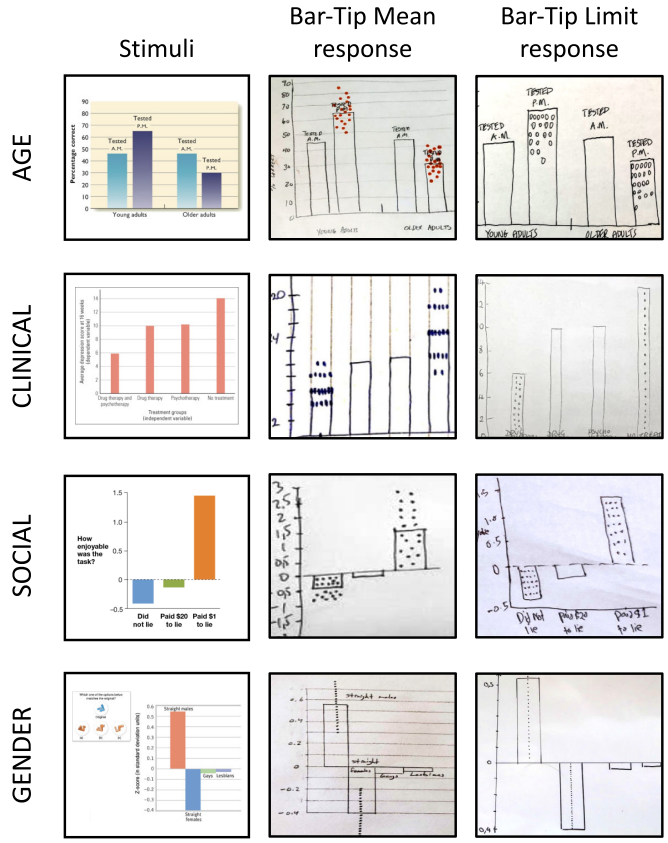

Figure 13. **DDoG measure readouts illustrating the two common, categorically different response types for each of the four stimulus graphs.** The left column shows the four stimulus graphs (AGE, CLINICAL, GENDER, SOCIAL). For each stimulus graph, the center column shows a representative correct response (Bar-Tip Mean), and the right column shows an illustrative incorrect response (Bar-Tip Limit).

The nature and severity of the BTL error

A minimum requirement to differentiate between errors, biases, and confusions is a response scale with sufficient granularity to produce a detailed, finely contoured response distribution. The DDoG measure readouts provide such granularity when analyzed via a relatively continuous scale such as the BTL index (Methods: Coding of DDoG measure readouts via BTL index).

The BTL index is the percentage of drawn datapoints that did not exceed the bar-tip (i.e., those that remain within the bar). On this scale, a response that treats the bar-tip as mean has a value of 50 (or 50% of datapoints within the bar), while a response that treats the bar-tip as limit has a value of 100 (or 100% of datapoints within the bar). Values between 50 and 100 represent gradations between the pure "mean" versus "limit" uses of the bar-tip.

Figure 14e plots the BTL indices for each of the 551 drawn graphs produced by 149 participants. Keep in mind that any degree of bimodality—even barely detectable rises amidst a sea of responses—would provide substantial evidence that measured inaccuracy in the responses resulted from common errors of interpretation, as opposed to biases or confusions. Yet the modes here in Figure 14e are not merely detectable; they are tall, sharp peaks with a deep valley of rare response values in between. Equally remarkable to the height and sharpness of the peaks is the trough between them, which bottoms out at a frequency approaching zero. In the context of the DDoG measure's *Expressive freedom* and *Information richness*, which enable a very wide array of possible responses, it is striking to find such a clear dichotomy.

The sharp peak on the right side of Figure 14e includes 86 drawn graphs (15.6%) with BTL index values of exactly 100. The left-hand mode is similarly sharp, with a peak at 50 that contains 169 (30.7%) drawn graphs. Taken together, the peaks of the two modes alone—BTL index values of exactly 50 and 100—account for nearly half of all drawn graphs (46.3%).

Notably, the incorrect right peak—which treats the bar-tip as limit rather than mean—is evident despite two factors that, even given starkly categorical thinking, could easily have blunted or dispersed it: first, our conservative scoring procedures (Methods: Coding of DDoG measure readouts via BTL index), and second, potential response variation based on differences in graph stimuli form and content.

In addition to the distinct bimodality in the data, the particular locations of the modes are revealing: one mode is at the correct Bar-Tip-Mean location, and the other is at the precise incorrect Bar-Tip-Limit location that is expected when mean bar graphs are conflated with count bar graphs.

The sharpness and location of these peaks not only provide clarity with regard to mechanism—an error, a conflation—but they also underscore the precision of the measure that is able to produce such clear results. Notably, the DDoG measure, using the BTL index, is fully capable of capturing patterns of bias (quantitative leaning, Figures 14c, 8c), or confusion (nonsystematic perplexity, Figures 14d, 8d), or values that do not correspond to any known cause. Yet here we see it clearly reveal an error (Figures 14b, 8b) with conflation as its apparent cause.

Panels f through i, on the right side of Figure 14, show BLT indices plotted separately for each stimulus graph. These plots are nearly indistinguishable from each other, and from the main plot (14e), despite substantial differences between the four stimuli (Figure 11). These plots thereby demonstrate four internal replications of the aggregate results shown in the main plot (14e). Such replications speak to the generality of the BTL error phenomenon.

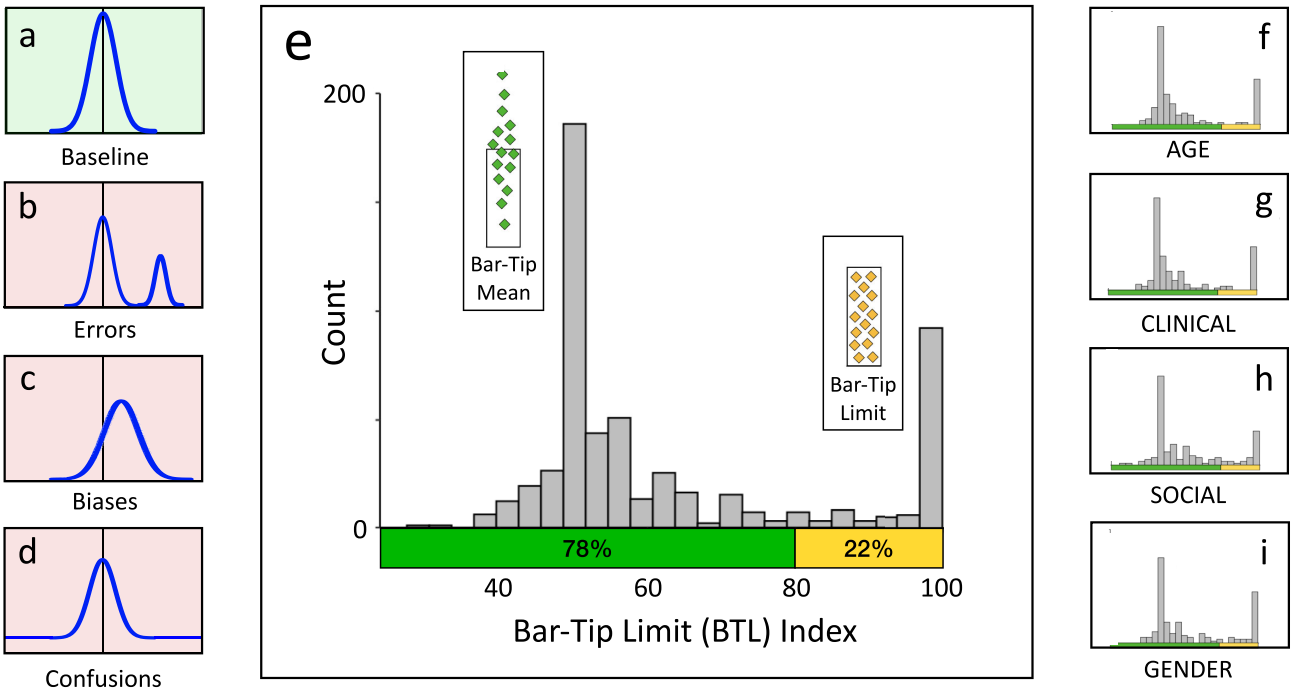

Figure 14. **Bimodal distribution of Bar-Tip Limit (BTL) index values reveals that the BTL error represents a categorical difference.** The main graph (e) plots the distribution of BTL index values for all 551 DDoG measure drawings. The left inset graphs show visual definitions of: (a) baseline (no systematic inaccuracy), (b) errors, (c) biases, and (d) confusions (Figure 8). Note how closely e (our data) matches b (data pattern for errors). The right inset graphs (f, g, h, i) plot BTL indices by graph stimulus (AGE, CLINICAL, GENDER, SOCIAL), which provide four internal replications of the aggregate result (e). The colors on the x-axes indicate the location of the computed cutoff of 80, taken from our cluster analyses, between the two categories of responses: Bar-Tip Mean responses (green) and Bar-Tip Limit (BTL) responses (yellow).

Critically, these analyses demonstrate that Bar-Tip Limit drawings like those shown in Figure 13 are in no way rare or atypical. They are not outliers. Nor are they the tail of a distribution. Rather, they represent their own mode; a categorically different, categorically incorrect interpretation, or error. We next turn to a more formal assessment of BTL error prevalence.

Prevalence of the BTL error: One in five

Having first characterized the BTL error as a severe error in graph interpretation (Figures 13, 14b, 14e), we then assessed its prevalence in our sample using a classification-based approach. A multi-method cluster analysis yielded the BTL index of 80 as a consensus cutoff between Bar-Tip Limit and Bar-Tip Mean responses (Methods: Establishing cutoffs and prevalence). We marked Figure 14 (panels e to i) with green and yellow x-axis colors to capture this distinction. At this cutoff, 122 of 551 graphs, or 22.1%, 95% CI [18.9, 25.8], demonstrated the BTL error.

Given the strength of bimodality in these data, and the sharpness of the BTL peak, BTL error prevalence values are remarkably insensitive to the specific choice of cutoff. For example, across the full range of cutoffs (75 to 85) produced by the multiple cluster analyses we carried out, these estimates remain around one in five (see Methods: Establishing cutoffs and prevalence). As we will see below, this one in five prevalence also turns out to be highly consistent with a careful reanalysis of past results (A reexamination of prior results).

Analytic quantification of bimodality

A next step, complementary to Figure 14's graphical elucidation of the BTL error's categorical nature, was to analytically quantify the degree of bimodality in the data. We do this via two standard analytical bimodality indices: Hardigan's Dip Statistic (HDS) and the Bimodality Coefficient (BC). HDS and BC are complementary measures that are often used in tandem (Freeman & Dale, 2013; Pfister et al., 2013). HDS and BC utilize idiosyncratic scales that range from 0.00 to 0.25, and from 0.33 to 1.00, respectively. To illustrate how these scales quantify bimodality, Figure 15 plots simulated distributions of 10,000 datapoints (middle), with both their HDS (top, blue) and BC (bottom, purple) scores. The HDS and BC scores for this dataset (the 551 drawn graphs) are indicated with arrows, accompanied by 95% confidence intervals (CIs). Clearly, the uncertainty in the data is small relative to the exceptionally high magnitude of bimodality. Again,

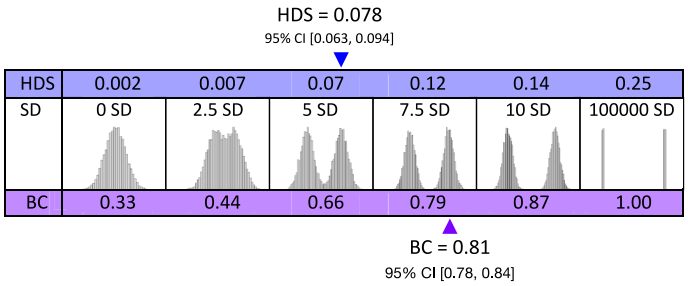

Figure 15. **Two analytic approaches confirm that the BTL index data are strongly bimodal.** Distribution graphs show reference data sets where varying degrees of separation were imposed on independently generated normal distributions. Separation is quantified in units of standard deviation (SD). Two standard bimodality statistics for each reference distribution were computed: Hardigan's Dip Statistic (HDS) is shown above in blue; the Bimodality Coefficient (BC) is shown below in purple (Freeman & Dale, 2013; Pfister et al., 2013). HDS and BC for our full data set, marked with arrows at top and bottom, along with 95% CIs, confirm strong BTL index bimodality (Table 4 bimodality statistics provide four internal replications of this overall result).

the clarity of this result necessarily reflects not only the bimodal nature of the BTL error, but also the high resolution of the DDoG measure and the efficacy of the MAGI principles used in its design.

Consistent with the replications we see in the plotted data (Figures 14e–i), the analytic results per stimulus graph, collected in Table 4, show clear internal replications of the aggregate results, verifying the consistency of the finding of bimodality across substantial differences in form and content between the four stimuli.

Cognitive mechanisms of the BTL error

Even before our systematic investigation, Figures 1 through 6 suggested a conflation of mean and count bar graphs as the apparent cognitive mechanism for the BTL error. The case for conflation is heightened by the clear second mode shown in Figures 14e–i. A second mode indicates a common, erroneous thought process (Figure 8b, Figure 14b), and that mode's position at the Bar-Tip Limit response, with a BTL index of 100 signals a conflation with count bar graphs in particular. Here we evaluate potential sources of this conflation.

BTL error stability within individuals

A first question is whether the BTL error represents a relatively stable understanding, or whether it occurs ad hoc, reflecting a more transient thought process. To answer this question, we first investigate consistency of the BTL error in individuals.

We find substantial consistency. The scatterplot matrix shown inFigure 16 plots BTL indices using pairwise comparison. Each scatterplot represents the comparison of two graph stimuli, and each pale gray dot plots a single participant's BTL indices for those two graphs: the x-coordinate is an individual's BTL index for *one* graph, and the y-coordinate is their BTL index for *a second* graph.

If the individual's BTL index is exactly the same for both graphs, their dot will fall on the (black) line of equivalence (the diagonal line where x = y). As a result, the distance of a dot from the line of equivalence demonstrates the difference in BTL index values between the two compared stimulus graphs. The dots cluster around the line of equivalence in every scatterplot in Figure 16, indicating that individuals show substantial consistency of interpretation across graphs.

The dots are, additionally, semitransparent, so that multiple dots in the same location show up darker. For example, the very dark dot at the top-right corner of the top-right scatterplot represents the 18 participants whose BTL index values on both the GENDER graph (x-axis) and AGE graph (y-axis) are 100. Thus, the presence of dark black dots at (x = 100, y = 100) and (x = 50, y = 50) in all of the scatterplots is a focused visual indicator of consistency within participants.

Together, these scatterplots confirm what the two DDoG measure readouts in Figure 5 suggested: that an individual's interpretation of one graph is highly predictive of their interpretation of the other graphs, and that those who exhibit the BTL error for one graph are very likely to exhibit it in most or all other graphs, regardless of graph form or content.

Further supportive of this conclusion of consistency, we found that a Cronbach's alpha internal consistency statistic (Cronbach, 1951), computed on the entire dataset shown in Figure 16, was near ceiling, at 0.94. This high internal consistency indicates that an individual's BTL indices across even just these four graphs can capture their relative tendency toward the BTL error with near-perfect precision. In sum, the BTL error appears to represent not just an ad hoc, one-off interpretation, but rather, a more stable thought process that generalizes across graphs.

	AGE	CLINICAL	SOCIAL	GENDER
Hardigan's Dip Statistic (HDS) [and 95% CIs]	0.077 [0.054, 0.109]	0.076 [0.052, 0.112]	0.067 [0.042, 0.099]	0.094 [0.063, 0.125]
Bimodality Coefficient (BC) [and 95% CIs]	0.859 [0.811, 0.900]	0.833 [0.782, 0.875]	0.704 [0.611, 0.800]	0.811 [0.767, 0.859]

Table 4. **Data bimodality analysis, per stimulus graph.**

Independence of the BTL error from graph content or form

Conceivably, despite the substantial stability of erroneous BTL interpretations within individuals, there might still exist global differences, from one stimulus to another, in levels of erroneous BTL thinking.

Remember that we chose our four graphs to vary substantially in both form (unidirectional versus bidirectional bars) and content (independent variables of age, therapeutic intervention, social situation, and gender, and dependent variables of memory, depressive symptoms, enjoyment, and visuospatial performance). If BTL interpretations were to differ substantially between pairs of these stimuli, it could suggest that such form or content differences played a role in generating or moderating the BTL error.

Figure 16 plots global mean BTL index values as red dots. The further from the black line of equivalence (where x = y) a red mean dot lies, the more divergent the average interpretation is between the two graphs. All red mean dots are close to the line of equivalence, indicating that participants, as a group, exhibit similar levels of BTL interpretation for all graph stimuli. The small distances between the red dot and black line on the graph are echoed numerically in the small Cohen's d effect sizes (Figure 16, "d="). Despite a slight tendency for bidirectional graph stimuli (SOCIAL, GENDER) to produce greater BTL interpretation than unidirectional graph stimuli (AGE, CLINICAL), the Cohen's d values do not exceed the 0.20 threshold which defines a small

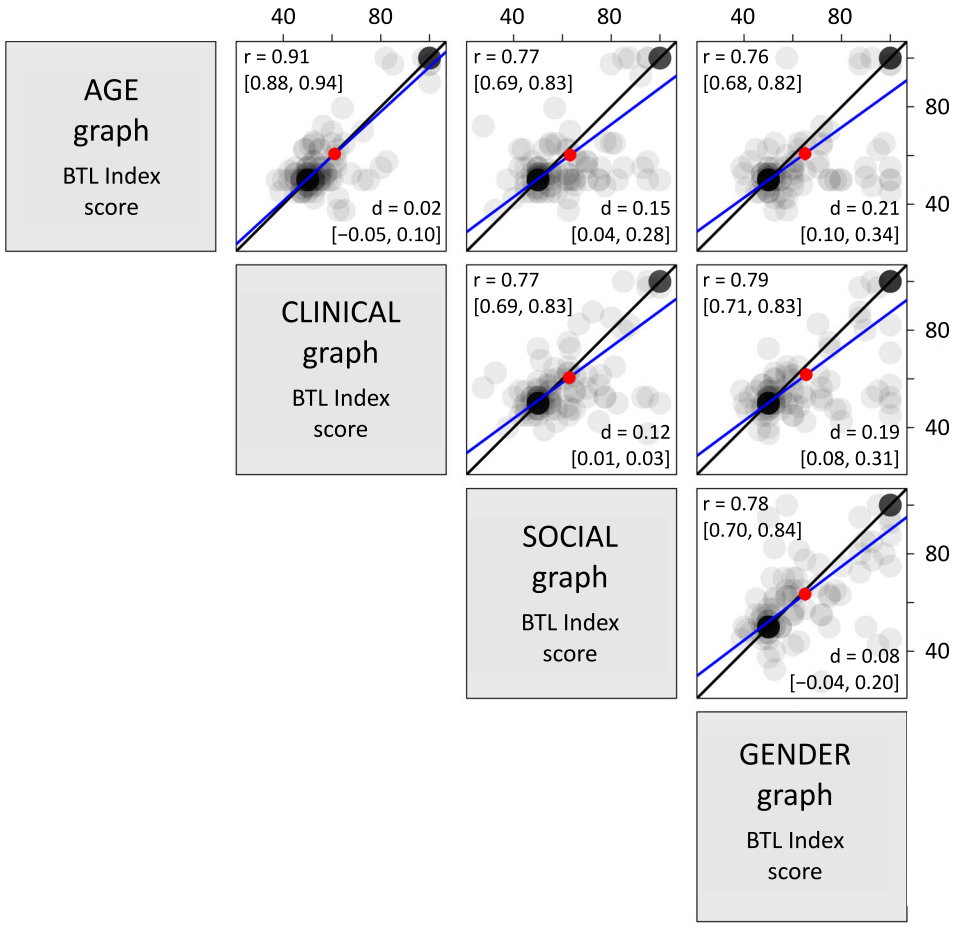

Figure 16. **The Bar-Tip Limit (BTL) error persists across differences in graph form and content.** Each of the six subplots in this figure compares individual BTL index values for two graph stimuli, one plotted as the x-value and the other as the y-value. Each gray dot represents one participant's data. Mean values are shown as red dots. Gray dots are semitransparent so that darkness indicates overlap. High overlap is observed near BTL index values of 50 (Bar-Tip Mean) and 100 (Bar-Tip Limit) for both stimulus graphs, indicating persistence of interpretation between compared stimuli, despite differences in graph form and content. This persistence is reflected numerically in the high correlations among (Pearson's r), and the small mean differences between (Cohen's d), graph stimuli. The high correlations are echoed visually by steep lines of best fit (blue), and the small differences are echoed visually by the proximity of mean values (red dots) to lines of equivalence (black lines, which show where indices are equal for the two stimuli). In brackets are the 95% CIs for r and d.

effect (Cohen, 1988). The BTL error thus appears to occur relatively independently of graph content or form.

Independence of the BTL error from foundational knowledge

A next step in isolating the thought processes that underlie the apparent conflation of mean bar graphs with count bar graphs is to ask what relevant foundational knowledge the participants have.

We assessed foundational knowledge in two ways: (1) a definition task: asking participants to define the concept of average/mean in their own words immediately after the drawing task (Definition task: Define the average/mean), and (2) a graph reading task: asking participants to read a particular average/mean value from stimulus graphs that they later sketched for the DDoG measure drawing task (Graph reading task: Find the average/mean). In the definition task, 83% of definitions were mathematically or conceptually correct, and in the graph reading task, 93% of readings were correct.

As Figure 17 shows, filtering responses for a correctly *defined* average/mean, a correctly *identified* average/mean, or both, has little effect on the percentage of BTL interpretations. The rate of the BTL error remains near one in five, at 96/486 = 19.8% [16.5, 23.5], 87/413 = 20.7% [17.1, 24.9], and 71/367 = 19.3% [15.6, 23.7], respectively. Together, these analyses suggest that the BTL error persists despite foundational knowledge about the average/mean and, moreover, despite sufficient care to the foundational knowledge questions to answer them correctly.

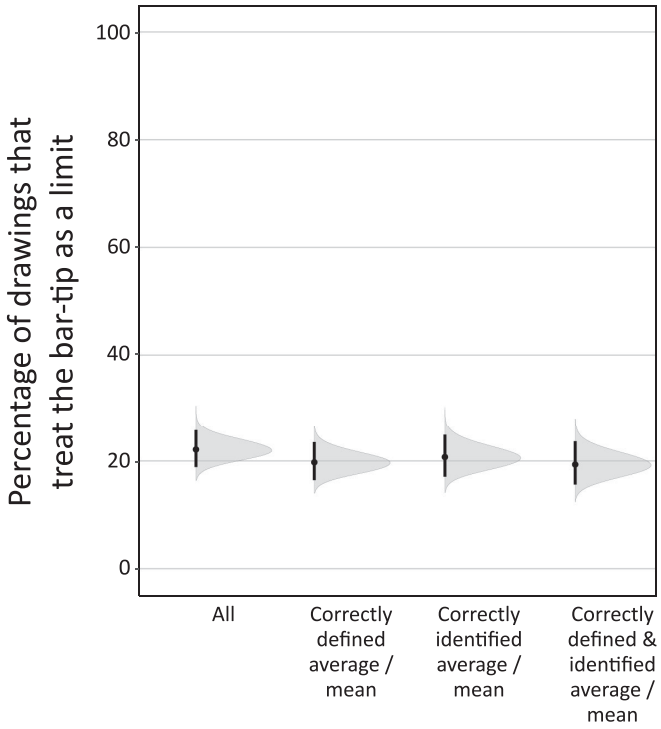

Figure 17. **The Bar-Tip Limit (BTL) error occurs despite correctly defining "mean" and correctly locating it on the graph.** Percentage of readouts that showed the BTL error (defined as a BTL index over 80). "All" is the full dataset ($n = 551$ readouts). "Correctly defined average/mean" is restricted to participants who produced a correct definition for the mean ($n = 486$ readouts). "Correctly identified average/mean" is restricted to participants who correctly identified a mean value on the same graph that produced the readout ($n = 413$ readouts). "Correctly defined & identified average/mean" is restricted to participants who both correctly defined the mean and correctly identified a mean value on the graph ($n = 367$ readouts). In all cases, the proportion of BTL errors hovers around one in five. Vertical lines show 95% CIs, and gray regions show full probability distributions for the uncertainty around the percentage values.

The BTL error occurs despite thoughtful, intentional responding

It was also relevant to establish that erroneous BTL interpretations were thoughtful and intentional by exploring whether carelessness, rushing, or task misunderstanding were major factors in BTL error. In addition to the results shown in Figure 17, two further factors suggest that the BTL error persists despite thoughtful attention specifically to the DDoG drawing task itself.

First, drawn graphs for which the instructions were not followed sufficiently to provide usable data were excluded from the start (Methods: Collection of DDoG measure readouts). Here, the *Expressive freedom* of the DDoG measure worked to our advantage. More constrained measures—for example, button presses or rating scales—often render carelessness and task misunderstanding difficult to detect and segregate from authentic task responses. The DDoG measure, in contrast, makes it much harder to produce even a single usable response without a reasonable degree of intentional, thoughtful responding, thus guaranteeing a certain intentionality in the resulting data set.

Second, rushing was evaluated as a possible contributing factor to erroneous BTL interpretations by computing the correlation of drawing time of usable graphs with BTL indices, on a graph-by-graph basis. This correlation was very close to zero: r(371) = –0.01, 95% CI [–0.11, 0.09]. Due to skew in the drawing time measure, we additionally computed a nonparametric (Spearman) correlation, which was also close to zero: rho (371) = 0.004, 95% CI [–0.10, 0.11]. The lack of correlation between drawing speed and BTL index provides further evidence that lack of intentional, thoughtful responding was not a major contributor to erroneous BTL interpretations.

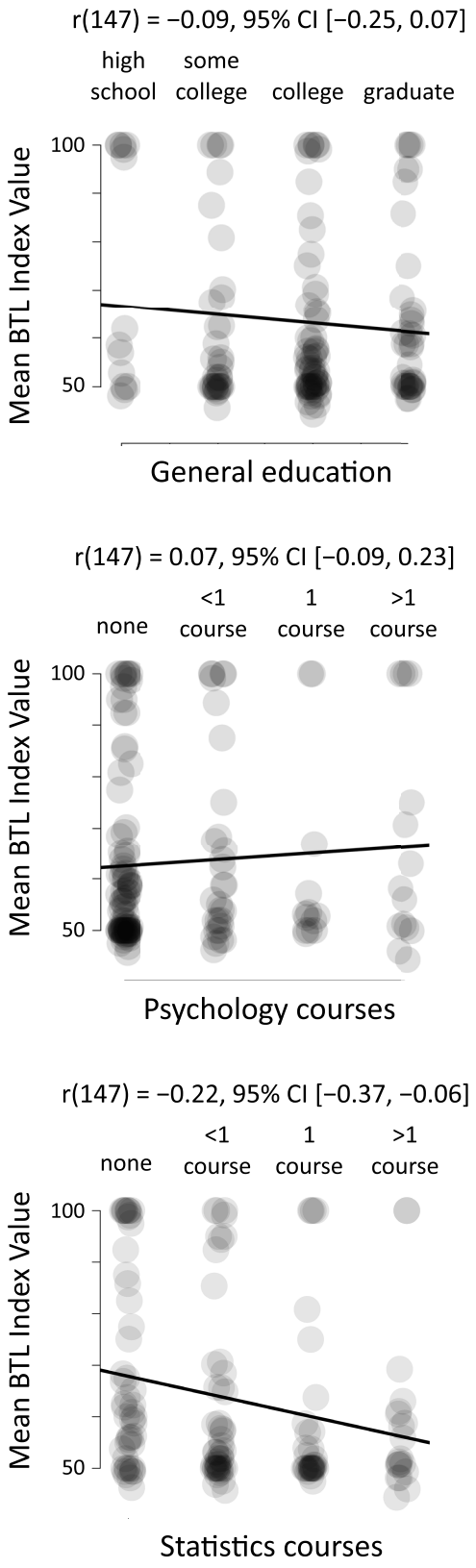

Figure 18. **The Bar-Tip Limit (BTL) error is substantially independent of education.** Mean BTL indices for each participant plotted against general education level, number of psychology courses taken, and number of statistics courses taken (each reported via the four-point scale shown on the respective graph). The black line is the least-squares regression line, computed with rated responses treated as interval-scale data. Axis ranges and graph aspect ratios were chosen by ShowMyData.org such that the physical slope of each regression line equals its respective correlation coefficient (*r*).

Educational and demographic correlates

BTL error correlates of individual differences: Education

The existence of individual differences in the likelihood of making the BTL error (stability within individuals) raises questions about their correlates and potential origins. With this in mind, we evaluated the dataset for demographic predictors. Of particular interest is whether prior educational experiences, general or specific, correlated with BTL readouts. The stability that the BTL index exhibited within individuals across graphs made it possible to use each individual's mean BTL index to evaluate educational correlates. Figure 18 shows that while general education and psychological coursework did not robustly predict mean BTL indices, statistics coursework did predict a slightly more correct (lower) mean BTL index.

This pattern of correlations raises several questions about education's relationship to the BTL error. First, what causes the nonzero relationship to statistics coursework, and does this relationship indicate that the BTL error is malleable and amenable to training? Second, why are all of these correlations not higher? Does this mean that the BTL error is such a tenacious belief that it is resistant to change? And, alternatively, are there aspects of standard statistical, psychological, or general education that could be improved to more effectively ameliorate the BTL error?

BTL error correlates of individual differences: Age, gender, country

In addition to education, we examined potential correlates: age, gender, and location (i.e., the country from which one completed the study). Age showed a modest correlation with mean BTL indices, with older individuals demonstrating somewhat more BTL error than younger individuals ($r(146) = 0.21$, 95% CI [0.05, 0.36]). Mean BTL indices were slightly lower among males, but not statistically significantly so ($r(146) = 0.10$, 95% CI [−0.06, 0.26]). While modest numerical differences in mean BTL indices were found between countries (United States $M = 66.1$, $SD = 20.3$, $n = 36$; United Kingdom $M = 61.2$, $SD = 15.6$, $n = 61$; Other countries $M = 64.0$, $SD = 19.6$, $n = 52$), none of these differences reached statistical significance (all p values > 0.05).

Ecological exposure to mean versus count bar graphs

As we have seen, the apparent conflation of mean and count bar graphs demonstrated by the BTL error does not appear to be a thoughtless act or careless misattribution. The data show, rather, that one in five participants capable of understanding a mean, identifying it on a graph, and producing attentive, intentional, consistent drawings make this error.

Broadly speaking, such thoughtful conflation could arise from either or both of "nature" and "nurture." Influences of nature might include fundamental, inborn aspects of perception, such as compelling visual interpretations of bars as solid objects (containers, stacks). Influences of nurture might include real-world experiences as varied as educational curricula and mass media.

While a full investigation of such developmental influences was beyond the scope of the current investigation, we were able to probe a key avenue made salient by our understanding of early graph pedagogy (Figure 1): putative exposure to mean and count bar graphs in early education and beyond.

As an investigation of the relative prevalence of these two bar-graph types, we counted bar graphs from three age-correlated sources. First, we evaluated a set of Google Image searches using the phrase "bar graph [X] grade" where X was "1st" through "6th" (Methods: Prevalence of bar graphs of means versus bar graphs of counts). This search yielded an array of graphs embedded in elementary pedagogical materials. Such graphs were deemed interesting due to early education's role in forming assumptions that may carry forward into adulthood. Second, we performed a general Google Image search for "bar graph," included as a window into the prevalence of count versus mean bar graphs on the internet as a whole. Our third and final source was a set of eight widely used college-level Introductory psychology textbooks, included as a common higher-level educational experience (Ciccarelli & Berstein, 2018; Coon et al., 2018; Gazzaniga 2018; Griggs 2017; Hockenbury & Nolan, 2018; Kalat, 2016; Lilienfeld et al., 2017; and Myers & DeWall, 2017).

As Figure 19 shows, in all three sources, mean graphs constituted only a minority of bar graphs. The effect was nearly unanimous among the elementary education materials, where mean bar graphs were almost nonexistent (1/81 = 1.2%, 95% CI [0.2, 6.7]). Mean bar graphs were slightly more prevalent in the general Google Image search, with 17% (17/100 = 17.0, 95% CI [10.9, 25.5]). Compared to the other sources, Introductory Psychology textbooks had a higher proportion of mean bar graphs, at 36% (53/149 = 35.6%, 95% CI [28.3, 43.5]). This higher proportion was predictable, given that in psychology research, analyses of mean values across groups or experimental conditions are often a central focus. Yet even in the psychology textbook sample, mean bar graphs still constituted only about a third of bar graphs.

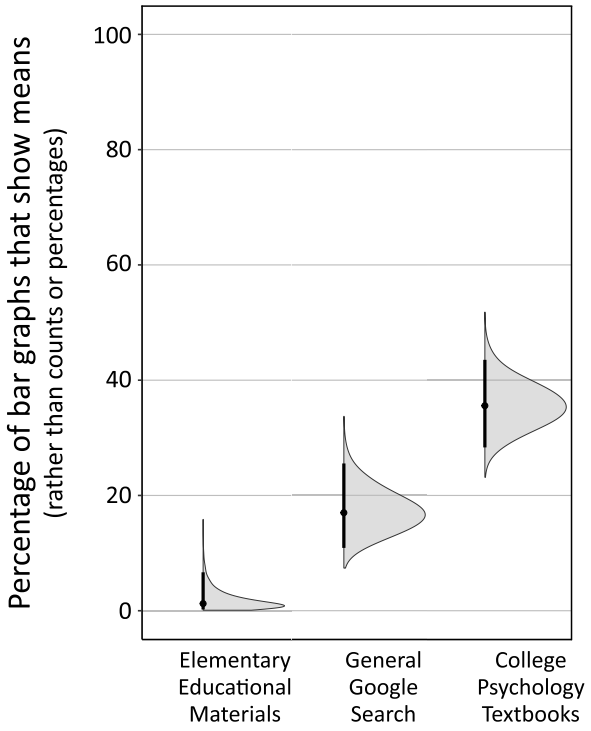

Figure 19. **Count bar graphs are more common than mean bar graphs across three major domains.** Mean bar graphs plotted as a percentage of total bar graphs (i.e., mean bar graphs plus count bar graphs). "Elementary Educational Materials": Google Image search for grades 1–9 ($n = 81$). "General Google Search": Google search for "bar graph" ($n = 100$). "College Psychology Textbooks": bar graphs appearing in eight widely used college Introductory Psychology textbooks ($n = 149$). Vertical lines show 95% CIs, and gray regions show full probability distributions for uncertainty around the percentage values. In all three cases, the percentage of mean bar graphs remains substantially, and statistically robustly, below 50%.

This analysis identifies elementary education both as a plausible original source of erroneous BTL thinking and as a potential lever for remediating it. Further, the preponderance of count bar graphs across all three of these sources suggests a pattern of iterative reinforcement throughout life. One's understanding of bar graphs, limited to count bar graphs in elementary school (Figures 1, Figure 19, and reinforced through casual exposure (Figure 19, Google), may represent a well-worn, experience-hardened channel of thinking that is modified only slightly by higher education

(Figure 18; BTL error correlates of individual differences: Education).

In this context, count bar graph interpretation plausibly becomes relatively effortless: a count graph heuristic, or mental shortcut (Shah & Oppenheimer, 2008; Gigerenzer & Gaissmaier, 2011). In the context of real-world exposure, where mean bar graphs are in the vast minority, such a count graph heuristic would generally produce the correct answer. Only occasionally, when misapplied to a mean bar graph, would it lead one astray.

Moreover, the relative abstraction, ambiguity, and complexity of mean bar graphs relative to count bar graphs, added to the visually identical forms of these two bar graph types (Figure 2), provides ideal fodder for such a heuristic-based conflation. Therefore, the apparent conflation that we observe may result from an experience-fueled heuristic, or cognitive shortcut.

A reexamination of prior results

As discussed in Related Works (Literature Related to the BTL Error), the probability rating scale approach used in prior studies of mean bar graph accessibility had successfully identified a reproducible asymmetry: higher rated likelihood of data in-bar than out-of-bar for the average person (Newman & Scholl, 2012; Correll & Gleicher, 2014; Pentoney & Berger, 2016; Okan et al., 2018). For nearly a decade, that asymmetry—dubbed the "within-the-bar bias"—arguably provided the strongest existing evidence against the accessibility of mean bar graphs.

With our current results in hand, we now reexamine those prior reports of asymmetry, offering four key insights: First, the prior work observed a prevalence remarkably similar to ours. Second, the bimodality and accurate left peak in prior results indicate that the prior work misidentified an error as a bias. Third, a flatter, wider right data cluster in Pentoney and Berger's study (2016), compared to ours, highlights a key limitation of the probability rating scale approach used in prior work. Fourth, recent changes in methodological and statistical practices shed light on choices that may have affected result characterization in prior work.

The first two observations—similar prevalence and strong bimodality—confirm that the previously reported "within-the-bar bias" was likely not a bias at all, but rather the same BTL error, and the same apparent count bar conflation, that is revealed here by our data. The latter two observations—right-mode differences and changing statistical practices—help to explain how both error and conflation could have remained undiscovered in prior work despite over a decade's worth of relevant, robust, self-consistent evidence from numerous independent labs.

Prior work reflects the same phenomenon: Prevalence

Our data demonstrate the Bar-Tip Limit (BTL) error in about one in five persons, across educational levels, ages, and genders, and despite thoughtful responding and relevant foundational knowledge. Prior studies differed from ours in that they were conducted by different research groups, at different times, with different participant samples, using a different measure (Newman & Scholl, 2012; Correll & Gleicher, 2014; Pentoney & Berger, 2016; Okan et al., 2018). If, despite all these differences, a prevalence reasonably comparable to our one in five was detectible in one or more of the prior studies, it would provide striking evidence that: (1) those past results were likely caused by the same BTL error phenomenon and (2) the underlying BTL error is robust to whatever differs between our study and the past studies.

Pentoney and Berger (2016, hereafter "PB study") provide a particularly rich opportunity to reexamine prevalence, because, unique among the past studies cited above, the PB study produced graphs that showed raw data for each participant. A careful look (see PB study data replotted as our Figure 20b) demonstrates a familiar bimodality. Additionally, the same analytic tests used to quantify the bimodality in our own data show a high degree of bimodality in the PB study data (HDS = 0.068 [0.042, 0.098]; BC = 0.66 [0.56, 0.75]), which strongly supports the principled separation of those data into two groups.

We applied to the PB study the same five cluster analysis techniques used to define the correct/error cutoff in our own data (Methods: Establishing cutoffs and prevalence). These analyses yielded an average prevalence estimate of 20.5%. Based on visual inspection of the PB study data in Figure 20b, this cutoff does a good job of separating the two apparent clusters of responses. Additionally, the resulting 20.5% prevalence estimate closely replicates the approximate one in five estimate from our own sample and suggests the BTL error as the source of the asymmetry reported in the PB study.

While none of the other three prior probability rating scale studies (Newman & Scholl, 2012; Correll & Gleicher, 2014; Okan et al., 2018) provided raw underlying data of the sort that would enable a formal analysis of bimodality and prevalence, a single parenthetical statement in Newman and Scholl's (2012) Study 5 (hereafter NS5) is suggestive. It notes:

> "In fact, 73% of the participants did rate the [in-bar vs out-of-bar] points as equally likely in this study. But ... the other 27% of participants — i.e., those in the minority who rated the two points differently — still reliably favored the point within the bar."

In other words, 73% of participants showed no asymmetry, which is the expected response for an

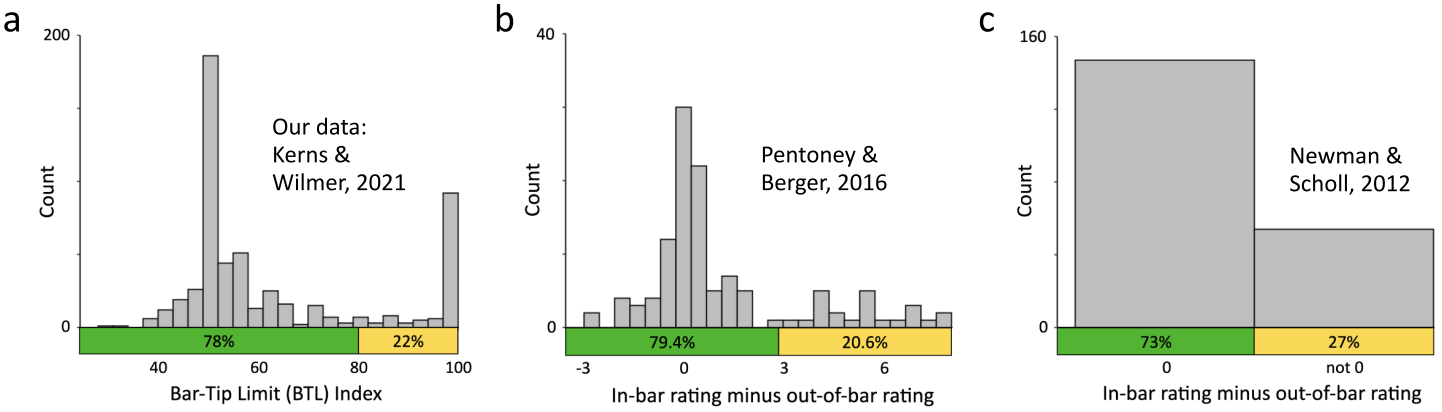

Figure 20. **A reexamination of prior results reveals both consistency with our current results and overlooked evidence for the Bar-Tip Limit (BTL) error.** Shown side by side for direct comparison are plotted data from: (a) The present study using the DDoG measure; (b) Pentoney and Berger (2016) (PB); (c) Newman and Scholl (2012) Study 5 (NS5). All graphs label correct Bar-Tip Mean response values in green and computed (or inferred, in the case of NS5) Bar-Tip Limit (BTL) response values in yellow. Notably, the PB study used the exact same 9-point rating scale and graph stimulus (of hypothetical chemical freezing temperature data) as the NS5 study (Figure 9 shows the scale and stimulus) but added ratings of four additional temperatures (−15, −10, 10, 15) to NS5's original two (−5, 5). Comparing PB's results (b) to ours (a) suggests that while PB's version of the probability rating scale measure achieved a fairly high degree of accuracy and precision at values near zero, it still suffered from apparently irreducible inaccuracy and/or imprecision as values diverged from zero (see text for further discussion).

interpretation of the bar-tip as a mean. The asymmetry shown by the other 27% of participants, while not numerically broken down by direction and magnitude, "reliably favored the point within the bar"; and the average result, including the 73% of participants with zero asymmetry, was a robust in-bar asymmetry. So, the notion that around 20% may have shown a rather large in-bar asymmetry, as in the PB study seems highly plausible. We plotted the zero asymmetry and nonzero asymmetry results for the NS5 study as Figure 20c.

In sum, prevalence estimates gleaned from past studies are highly consistent with our own prevalence estimates, suggesting both that the asymmetry found in past studies was caused by the same BTL error, and that this error of interpretation is highly robust to differences between studies.

Prior work reflects the same phenomenon: Error, rather than bias

In addition to providing prevalence, the distinct clusters of responses and strong bimodality in the PB study's data demonstrate a key feature of erroneous graph interpretation: two discernible peaks, or modes (Figures 8b, 21a). A second key feature of erroneous interpretation is that one of the two modes is on the correct interpretation.

Both the DDoG and PB data plots demonstrate two peaks, and in each plot, one of those peaks corresponds exactly to the correct value for a Bar-Tip Mean interpretation (left mode in Figures 8b, 21a and 20b, 21b).

In the PB study, this correct value is 0. The PB study's analyses used a difference score: the mean in-bar rating minus the mean out-of-bar rating. This score ranged from −8 to 8, in increments of 1/3. The task was taken directly from the NS5 study (same stimulus, same response scale, see Figure 9a). However, the PB study had participants rate three in-bar locations (rather than NS5's one) and three out-of-bar locations (rather than NS5's one). The study therefore had more datapoints, and greater continuity of measurement, than that of NS5 (Figure 20b).

A difference score near 0 indicates similar in-bar and out-of-bar ratings, which is the expected response for a Bar-Tip Mean interpretation. Approximately 25% of the PB study's participants (30 of 118) gave a response of exactly 0, and more than half (64 of 118, or 54%) gave a response in the range –0.67 to 0.67, creating a tall, sharp mode at the correct Bar-Tip Mean value.

Results from the NS5 study tell a similar story. Despite the NS5 study's relatively low-resolution measure (the single rated in-bar location and out-of-bar location), and limited reporting of individual values (only the proportion of response scores that were zero versus nonzero were reported), we can still see that the NS5 data contained a mode on the correct value (Figure 20c), replicating a feature of erroneous interpretations. Indeed, a full 73% of participants gave 0 as their response, particularly strong evidence *against* bias (which tends to shift a peak away from the correct value) and *for* error (which tends to preserve a peak at the correct value; see Figure 8).

Therefore, a careful reexamination of the available information from the PB and NS5 studies shows strong evidence for a causative, systematic error in participant interpretation. The one in five prevalence of asymmetry, strong bimodality, and a response peak at the correct value are three further indications that the previously observed effects were likely caused by the same BTL error that we observed via the DDoG measure.

How the DDoG measure supported the identification of the BTL error

If the BTL error truly is at the root of these prior reports of asymmetry, why did its existence and apparent mechanism—conflation of mean bar graphs with count bar graphs—elude discovery for over a decade?

Though our reexamination of past results (Newman & Scholl, 2012; Correll & Gleicher, 2014; Pentoney & Berger, 2016; Okan et al., 2018) has so far focused on ways in which the probability rating scale data aligns closely to our DDoG measure findings (Figure 20), the answer to our question may lie in the areas of difference.

Figure 20 shows a salient way in which the DDoG measure response distribution (Figure 20a) *differs* from that of the PB study (Figure 20b): the shape of the rightmost (erroneous) error peak. DDoG measure's BTL error peak is sharp, and the location corresponds exactly to the expectation for a count bar graph interpretation (i.e., a BTL index of 100, representing a perfect Bar-Tip Limit response Figure 20a). The PB study's probability rating scale data, in contrast, shows no such sharp right peak. The right data cluster is flatter and more dissipated (Figure 20b).

To understand why these response distributions would differ so much if they measure the same effect, consider a core difference between measures. The probability rating scale requires participants to translate their interpretation of a graph into common English words ("likely," "unlikely," "somewhat," and "very") (see Figure 9a). This location-to-words adjustment is a prime example of the sort of mental transformation that we sought to minimize via the MAGI principle of limited mental transformations. Such translations infuse the response with the ambiguity of language interpretation (i.e., does every participant quantify "somewhat" in the same way?) A relevant analogy here is the classic American game of "telephone," where translations of a message, from one person to the next, may yield an output that bears little resemblance to the input. Similarly, the greater the number of necessary mental transformations that exist between viewing and interpreting a graph stimulus and generating a response, the less precise the translation, and the more difficult to reconstruct the interpretation from the response. In such cases, it becomes difficult to use a response to objectively judge the interpretation as correct or incorrect, violating the MAGI principle of *Ground-truth linkage*.

Further, though rating scale method responses can indicate the presence and directionality of an asymmetric response, they lack the necessary specificity (e.g., "somewhat," "very") to delineate, in concrete terms, the *magnitude* of asymmetry in the underlying interpretation. As we have seen, it is the magnitude of the error—the position of the second mode revealed by the DDoG measure—that so clearly indicates its mechanism.

But why does this translation issue only pertain to the right peak of the PB study's distribution? This is because the (left) Bar-Tip Mean peak represents a mean value. In a DDoG measure readout, this mean value is concretely expressed by equal distribution of datapoints on either side of the bar-tip. In the probability rating scale measure, equal distribution is expressed by using the same word to describe the likelihood of both the in-bar and out-of-bar dot position. In this single case of equality, the language translation issue is obviated: so long as the person uses the same words (e.g., "somewhat likely," or "very unlikely") to describe both in-bar and out-of-bar locations, it matters little what those words are. In this specific case only, when the difference score is zero, the probability rating scale has *Ground-truth linkage.* This, in turn, results in a tall, sharp, interpretable left (Bar-Tip Mean) peak in the asymmetry measure.

How changing statistical and methodological practices supported the identification of the BTL error

We have seen that the mental transformations required by previous rating scale methods impaired identification of the BTL error despite compelling evidence for a categorical (bimodal) distinction between accurate and erroneous thinking. Here, we explore possible analysis-based reasons for the BTL error's previous mischaracterization as a bias, and we reflect on recent changes in methodological and statistical practices that make the distinction clearer in hindsight.

Methodologically, a traditional best practice was to make a single, strong theoretical prediction and then examine the data for consistency with that prediction. While the existence of strong theory can avoid the risks of post hoc analytic degrees of freedom, it also risks constraining data collection and data analysis in ways that miss unexpected patterns (Cumming, 2013). Notably, prior reports of an asymmetry in mean bar graph interpretation (Newman & Scholl, 2012; Correll & Gleicher, 2014; Pentoney & Berger, 2016; Okan et al., 2018) focused on the theory of an "automatic" and "irresistible" perceptual bias (Newman & Scholl, 2012). The lack of such a strong theory in the early stages of our own work was a choice that aimed to preserve our openness to a variety of phenomena and mechanisms.

More specifically, we used a relatively atheoretical piloting process to hone the DDoG measure and isolate phenomena of interest. The iterative nature of this process serves as a tool for establishing reproducibility and removes the need for a strong theory to protect against "seeing faces in the clouds" (aka seeing an apparent pattern in the data that is due to random variation rather than due to a stable signal; Cumming, 2013). We designed the DDoG measure using the MAGI principles: its Expressive *freedom* imbues it with the capacity to expose unpredicted patterns, while *Ground-truth linkage, Limited mental transformations*, and *Information richness* impart direct and detailed interpretability of the patterns it might present us with.

Statistically, the traditional approach focuses on comparing mean values via null hypothesis significance tests (NHST) (Cumming, 2013). Prior studies reported statistical analyses primarily in terms of NHSTs of mean values (Newman & Scholl, 2012; Correll & Gleicher, 2014; Pentoney & Berger, 2016; Okan et al., 2018). Yet as we have seen, it is the shape of the distribution, rather than its mean value, that reveals the core insights reported here: prevalence, bimodality, a correct mode (at the Bar-Tip Mean value), and an error mode (at the Bar-Tip Limit value).

A mean-focused Null Hypothesis Significance Testing approach has at least three potential limitations. First, any mean-focused approach potentially obscures important differences in distributional shape. For example, consider Figure 21: though the mean lines (red) produced by erroneous and biased interpretations are comparable, the distribution shapes are entirely different.

Second, the usage of the mean value as a "summary statistic" implicitly assumes that it effectively "summarizes," or represents, the response distribution. Yet for strongly bimodal distributions like those we observe here, the mean value rarely, if ever occurs, and it therefore poorly represents the response distribution. In such a case, results reported in terms of mean values, while arguably justifiable on pure statistical grounds, risk theoretical misinterpretation. For example, if one were to fall into assuming that the mean value was representative (e.g., via the standard "normal distribution" assumption), one would, effectively, be assuming the existence of a bias (Figure 21b) and the absence of an error (Figure 21a).

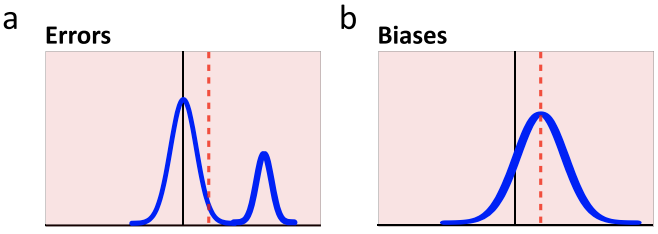

Figure 21. **Even when data clearly indicate the presence of an error, analyses based purely on mean values could misidentify that error as a bias.** Shown are visual definitions of (a) errors and (b) biases, using prototypical response distributions (see discussion of Figure 8). Black lines represent the correct response; blue lines represent alternate distributions; red lines represent mean values. Since the mean values do not differ between the two shown distributions, mean value alone is insufficient to distinguish between error and bias.

Third, the Null Hypothesis Significance Testing approach draws a dichotomous distinction between a "null" value (typically zero) and all other values (Cumming, 2013). The probability rating scale approach does this also, offering a clear theoretical distinction between zero, which is concretely interpretable as an accurate Bar-Tip Mean interpretation, and all other values. Where the rating scale falls short, however, is in the interpretability of nonnull values (see How the DDoG measure supported the identification of the BTL error). The rating scale is therefore well-matched to the particular, null-focused distinction that NHSTs were designed to draw. Yet it was precisely the interpretability of the DDoG measure's non-Bar-Tip Mean (nonnull) values that supported each core insight reported here.

It is difficult to know for sure whether these three potential disadvantages of a mean-focused, Null Hypothesis Significance Testing statistical approach impacted the major conclusions drawn in the four prior reports of asymmetry in mean bar graph interpretation (Newman & Scholl, 2012; Correll & Gleicher, 2014; Pentoney & Berger, 2016; Okan et al., 2018). At minimum, it speaks to the influence of the mean-focused statistical paradigm that these prior papers would feel compelled to report their data primarily in terms of mean values—even though their investigative focus was the inaccurate conceptions of raw data that underlie mean values.

Our investigation, in contrast, intentionally prioritized a focus on individual responses: through the creation of a measure that made each response as informative as possible; the analysis of individual responses; and the choice of analyses and visualizations that retained a clear connection to individual responses. This individual-focused approach made possible the key insights and contributions reported here.

Discussion

Three mutually supportive contributions

This paper makes three contributions to the scientific study of abstract-graph cognition: a set of principles, a

measure, and the discovery of a common severe error. These contributions each have separate and substantial uses and implications for research and practice, even while they support and illustrate aspects of the other two.

We call the first contribution the Measurement of Abstract Graph Interpretation (MAGI) principles. The MAGI principles aim to guide the collection of graph interpretation data for graph types and markings that abstract away from individual-level data to present aggregate-level information. They center around the dual aims of general usage and valid measurement. We have here illustrated their utility as a metric for the creation, evaluation, and comparison of graph interpretation measures.

During this study, the MAGI principles guided the development of the DDoG measure, helped to evaluate and make sense of why and how the measure was able to show erroneous BTL thinking, and structured the comparison of the DDoG measure to other measures, pointing to important areas of overlap as well as notable differences. Each of the MAGI principles also constitutes its own testable hypothesis, whereby the principles could be supplemented, revised, or edited in the future based on the utility they do (or do not) demonstrate.

We call the second contribution the Draw Datapoints on Graphs (DDoG) measure. The DDoG measure was designed as an implementation of the MAGI principles and modeled after classic neuropsychological drawing-based tasks. It collects drawn readouts (concrete, detailed, visuospatial records of thought) to reveal the viewer's interpretation of a graph. We have here demonstrated that DDoG's readout-based measurement approach, and more specifically its elicited graph approach (whereby the elicited response is, itself, a graph) distinguish it from prior measures of graph interpretation. We offer the DDoG measure as exemplar and impetus for the development of further elicited graph and readout-based measures.

In this study, the DDoG measure demonstrates its utility by readily revealing a pattern (the BTL error) that is so salient as to be evident even at the level of a single bar, on a single elicited graph, drawn by a single participant (Figures 4, 5, 6). Given this salience, it is worth underscoring just how nonobvious the BTL error interpretation was before the DDoG measure. Conflation of mean and count bar graphs was not even on the list of theoretical concerns raised by those who critiqued mean bar graphs on logical (Tufte & Graves-Morris, 1983, p. 96; Wainer, 1984; Drummond & Vowler, 2011; Weissgerber et al., 2015; Larson-Hall, 2017; Rousselet, Pernet, & Wilcox, 2017; Pastore, Lionetti, & Altoe, 2017; Weissgerber et al., 2019; Vail & Wilkinson, 2020) or empirical grounds (Newman and Scholl, 2012; Correll and Gleicher, 2014; Pentoney and Berger, 2016).

While the existence and direction of an asymmetry in interpretation of mean bar graphs was evident in prior empirical work (Newman and Scholl, 2012; Correll and Gleicher, 2014; Pentoney and Berger, 2016; Okan et al., 2018), the DDoG measure's readouts give a disambiguating clarity that is entirely new, providing a clearer picture of its categorical nature, high prevalence, stability within individuals, likely developmental influences, and independence from foundational knowledge and graph content. Yet elucidation of the BTL error is just one of many possible applications of the DDoG measure. By investigating an increasingly wide array of graph gestures, topics, contexts, and readers, the potential insights offered by the DDoG measure are broad and varied.

We call the third contribution the Bar-Tip Limit (BTL) error. The BTL error is a common, severe, categorical misinterpretation of mean bar graphs, whereby the bar-tip is incorrectly interpreted as an outside limit rather than as a balanced center point. The basic cognitive mechanism for the BTL error appears to be a conflation of mean bar graphs with the visually identical count bar graphs. A series of control analyses suggests that this conflation is the result of a heuristic, or cognitive shortcut, as it persists even when logically inconsistent with other knowledge. The BTL error provides a case in point that using the same graphical gesture (bars) to represent two fundamentally different graphs risks confusion—hence our title "Two graphs walk into a bar."

The implications of the BTL error relate both to mean bar graphs specifically and, more generally, to abstraction as a route to visual simplification. Mean bar graphs have long been among the most common tools for communicating data to both expert and nonexpert audiences (Mogull & Stanfield, 2015; Weissgerber et al., 2015; Chen et al., 2017; Larson-Hall, 2017; Weissgerber et al., 2019). Their visual simplification is commonly believed to enhance understandability (Zubiaga & MacNamee, 2016; Barton & Barton, 1987; Angra & Gardner, 2017), and they have therefore persisted despite the criticism that they hide useful information (Tufte & Graves-Morris, 1983, p. 96; Wainer, 1984; Drummond & Vowler, 2011; Weissgerber et al., 2015; Larson-Hall, 2017; Rousselet, Pernet, & Wilcox, 2017; Pastore, Lionetti, & Altoe, 2017; Weissgerber et al., 2019; Vail & Wilkinson, 2020). By cutting at the heart of the primary presumed benefit of mean bar graphs (i.e., accessibility), the BTL error underscores the merit of these criticisms and provides a new urgency to calls for improved best practices. More generally, the BTL error provides a case in point that simplification, when accomplished via abstraction, can yield unintended negative consequences for comprehension.

Future directions

Future directions related to the BTL error

The existence of the Bar-Tip Limit (BTL) error as a severe, high-prevalence misinterpretation of mean bar graphs suggests a number of avenues for future work. Three we focus on are: mitigation through concrete interventions, further elucidation of detailed cognitive mechanisms, and further investigation of the BTL error's relationship to expertise.

Lever points for concrete intervention could include publishing, education, and point-of-contact "nudges." An example of changing publication practices comes directly from the datavis community. Careful attention to the prior reports of asymmetry in bar graph interpretation (Newman and Scholl, 2012; Correll and Gleicher, 2014; Pentoney and Berger, 2016; Okan et al., 2018) has led this community, in its published work, to largely abandon mean bar graphs in favor of dot plots, whereby the mean value is indicated by a dot rather than a bar. Similar practices could be encouraged, or explicit standards introduced, in other scholarly, educational, and/or mass communication publishing contexts. Key questions include which specific alternate graphing practices to support, and how to incentivize or enforce the resulting standards.

In education, various curricular lever points may exist. For example, students could be taught how to recognize and think about both mean and count bar graphs, and the data that underlies them. Key questions include what level(s) to target (college-level statistics courses, elementary school science) and which entities to work with to enact curricular change (governments, schools, teachers, students).

Point-of-contact "nudges" might come in different forms for different audiences. For graph producers, a nudge might consist of software defaults that automatically show raw underlying data, such as those implemented at ShowMyData.org (created by the second author), EstimationStats.org, or in the ESCI plug-in to Jamovi. For graph consumers, such a nudge might involve standard annotations to clarify the type of bar graph or give clarification about the nature of the distribution.

The second set of future directions we suggest involves deeper investigation into the BTL error's detailed cognitive mechanisms. For instance, although conflation of mean and count bar graphs seems to be the root cause, the persistence of this conflation despite logically inconsistent knowledge and thoughtful responding suggests the operation of a heuristic, or cognitive shortcut (Shah & Oppenheimer, 2008; Gigerenzer & Gaissmaier, 2011); a default, automatic, tenacious propensity to interpret any bar graph as a count bar graph. Further examination of the detailed nature and operation of such a heuristic could be a productive line of inquiry that might further inform the search for effective interventions.

A third set of future directions would follow up on our analyses of educational attainment, courses taken, and educational materials to explore the high end of expertise. Is BTL error entirely confined to nonexperts, or does it manifest, in some form, even among experts? We hypothesize that there may be inherent costs to using a single graphical gesture, the bar, to represent two mutually exclusive forms of data, that manifest even in experts. Perhaps such effects would be confined to measures of speed or perceived clarity, as opposed to frank errors. Conceivably, extensive experience with mean bar graphs might actually reverse such an effect, producing hesitation, momentary confusion, or even actual errors of interpretation when viewing count bar graphs.

Future directions related to the DDoG measure

Like any general-purpose tool, the Draw Datapoints on Graphs (DDoG) measure has a wide array of potential future uses. The shortest leap would be to study non-BTL error aspects of mean bar graph interpretation. Another straightforward next step would be to chase down the subtle, but nonzero differences observed here between mono- and bidirectional mean bar graphs. The DDoG measure could also be applied to nonmean bar graphs, such as count and stacked bar graphs, or histograms, or be used to study nonbar representations of the mean such as lines or dots.

A wide range of graphical depictions of other statistical entities that abstract beyond the raw underlying values, such as nonmean central tendency, variability, uncertainty, or distribution shape could be explored using the DDoG measure. The measure could also be used to examine the impact of graph design choices such as stretching, framing, annotation, coloring, or selection of axis ranges.

The DDoG measure's isolation of the BTL error provides clear validation of its utility. An important distinction that remains, however, is between *could* and *did*. The DDoG measure *did* isolate the BTL error. But *could* probability rating scales, or balls-and-bins—or completely different techniques, such as written verbal report, or in-person interviews—be used in a way that they would isolate the same BTL error with the same degree of accuracy and precision? While we have laid out our own analytic predictions, grounded in the MAGI principles, this is, ultimately, an empirical question.

Efforts to see phenomena through the eyes of multiple measurement approaches offers high potential benefit. We therefore call for future work that seeks both to uncover new phenomena and to study them using a multimethod approach that includes readout-based methods such as the DDoG measure along with the

full richness of methodology offered by vision science (Elliott et al., 2020), datavis (Hullman et al., 2018), psychology (Wilmer et al., 2012), and other disciplines.

While the DDoG measure has demonstrated substantial utility in its current form, there undoubtedly remains room for optimization and expansion, for instance: automations to scoring, strategic variations in numbers of drawn datapoints used for different purposes, or the development of novel indices to extract further aspects of graph interpretation from DDoG measure readouts.

Future directions related to the MAGI principles

A key contribution of the MAGI principles to the current investigation was to inspire and guide the readout-based measurement approach that we took. While the potential applications of the MAGI principles are broad, we suggest just a single future direction for the readout-based approach in particular: that readout-based artifacts, themselves, become a topic of study as communication tools. Such artifacts can have unusual clarity and persuasive power. The clock drawing shown in Figure 3, for example, is commonly featured in textbooks because it so powerfully illustrates pathology. Likewise, the DDoG measure response drawings shown in Figures 4, 5, and 6 powerfully illustrate erroneous thinking. These artifacts have immediacy, charisma, and accessibility, giving them high impact. Above and beyond their value for drawing scientific, evidence-based conclusions, readout-based methods may be useful for communicating science with immediacy and impact to institutional decision makers, practitioners, and laypersons alike.

Costs and benefits of data abstraction

We now come full circle, back to the overarching question of costs and benefits of simplification via abstraction, in graph design, but also more generally. The DDoG measure is used to assess how graphs that abstract away from their raw underlying data are comprehended, and our BTL error finding provides evidence that such abstraction can be very costly indeed. Our results thus motivate this as their broadest question: when is simplification via abstraction, per se, beneficial, and when is the cost in comprehension too high to support its use? We believe that existing evidence may justify a systematic reexamination of the presumption that central tendencies, absent underlying distributional information, enhance the comprehension of quantitative data.

Conclusions

This paper establishes our Measuring Abstract-Graph Interpretations (MAGI) principles and Draw Datapoints on Graphs (DDoG) measure as effective tools for gaining powerful insights into graph interpretation. Our work highlights the value of a readout-based approach, while providing tools and model examples to enable the greater adoption of such an approach in future work. With these tools, we demonstrate that mean bar graphs are subject to a common, severe error of interpretation—the Bar-Tip Limit (BTL) error—which makes them a potentially unwise choice for accurately communicating empirical results, particularly where general accessibility is a priority, such as in education, medicine, or popular media.

Keywords: bar graph of means, data visualization, DDoG measure, Bar-Tip Limit error, MAGI principles

Acknowledgments

The authors thank Heesu Kim for early piloting that led to the DDoG measure, and the peer reviewers for their helpful suggestions.

Funded in part by NSF award #1624891 to JBW, a Brachman Hoffman grant to JBW, and a subaward from NSF grant #1837731 to JBW.

Open Science Framework link containing raw drawings and raw data: https://osf.io/e2zh5.

Author Contributions: SHK and JBW contributed equally to this project in both effort and insight, collaborating closely on all its major aspects, including:
SHK and JBW developed the DDoG measure and the MAGI principles.
SHK and JBW observed and probed the BTL error.
SHK and JBW designed pilots and final studies.
SHK and JBW performed pilots and final studies.
SHK and JBW analyzed data from pilots and final studies.
SHK and JBW wrote and revised the paper.

Patrick Wilmer colored Figure 1b.

Commercial relationships: none.
Corresponding authors: Sarah R.H. Kerns; Jeremy B. Wilmer.
Emails: sarah.kerns@wellesley.edu; jwilmer@wellesley.edu.
Address: Department of Psychology, Wellesley College, Wellesley, MA, USA.

References

Agrell, B., & Dehlin, O. (1998). The clock-drawing test. *Age and Ageing, 41* Suppl 3, iii41–iii45.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392–398.

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

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

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

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

Choy, S. L., O'Leary, R., & Mengersen, K. (2009). Elicitation by design in ecology: Using expert opinion to inform priors for Bayesian statistical models. *Ecology, 90*(1), 265–277.

Ciccarelli, C., & Berstein, B., (2018). *Essentials of psychology* (7th ed.). Boston, MA: Cengage.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J: L. Erlbaum Associates.

Coon, D., Mitterer, J., & Martini, T. (2018). *Introduction to psychology* (15th ed.). New York, NY: Cengage.

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.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty. *Cognition, 58*(1–73)

Crilly, N., Blackwell, A. F., & Clarkson, P. J. (2006). Graphic elicitation: Using research diagrams as interview stimuli. *Qualitative Research, 6*(3), 341–366.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Abingdon, Oxon: Routledge.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*:4, 299–311.

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition, 126*(1), 87–100.

Deveney, C. M., Chen, S. H., Wilmer, J. B., Zhao, V., Schmidt, H. B., & Germine, L. (2018). How generalizable is the inverse relationship between social class and emotion perception? *PloS One, 13*(10), e0205949.

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

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature Communications, 10*(1), 1–10.

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

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology General, 123*(2), 161–177.

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

Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018, April). Uncertainty displays using quantile dotplots or CDFs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–12)

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology, 58*(2), 203.

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

Fortenbaugh, F. C., DeGutis, J., Germine, L., Wilmer, J. B., Grosso, M., Russo, K., . . . Esterman, M. (2015). Sustained attention across the life span in a sample of 10,000: Dissociating ability and strategy. *Psychological Science, 26*(9), 1497–1510.

Franzblau, L. E., & Chung, K. C. (2012). Graphs, tables, and figures in scientific publications: The good, the bad, and how not to be the latter. *The Journal of Hand Surgery, 37*(3), 591–596.

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods, 45*(1), 83–97.

Fyfe, E.R., McNeil, N.M., & Son, J.Y. et al. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review, 26*, 9–25

Gazzaniga, M. (2018). *Psychological science* (6th ed.). New York, NY: W.W. Norton & Co.

Germine, L., Russell, R., Bronstad, P. M., Blokland, G. A., Smoller, J. W., Kwok, H., Anthony, S. E., Nakayama, K., Rhodes, G., . . . Wilmer, J. B. (2015). Individual aesthetic preferences for faces are shaped mostly by environments, not genes. *Current Biology, 25*(20), 2684–2689.

Godau, C., Vogelgesang, T., & Gaschler, R. (2016). Perception of bar graphs—a biased impression? *Computers in Human Behavior, 59*, 67–73.

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

Gray, D., & Bjorklund, D. (2017). *Psychology* (8th ed.). New York, NY: Worth.

Griggs, R. (2017). *Psychology: A concise introduction* (5th ed.). New York, NY: Worth.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62(1)*451–48

Grison, S., & Gazzaniga, G. (2019). *Psychology in your life* (2nd ed.). New York, NY: W.W. Norton & Co.

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences, 109*(28), 11116–11120.

Hansen, C., Zidowitz, S., Ritter, F., Lange, C., Oldhafer, K., & Hahn, H. K. (2013). Risk maps for liver surgery. *International Journal of Computer Assisted Radiology and Surgery, 8*(3), 419–428.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 203–212.

Hockenbury, S., & Nolan, S. (2018). *Discovering psychology* (8th ed.). New York, NY: Worth.

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

Hullman, J. (2016). Why evaluating uncertainty visualization is error prone. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (pp. 143–151).

Hullman, J., Kay, M., Kim, Y., & Shrestha, S. (2017). Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE Transactions on Visualization and Computer Graphics, 24*(1), 446–456.

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.

Jenkinson, D. (2005). *The elicitation of probabilities: A review of the statistical literature*. BEEP working paper, Univ. Sheffield.

Kalat, J. (2016). *Introduction to psychology* (11th ed.). Boston, MA: Cengage.

Kang, H., Ji, J., Yun, Y., & Han, K. (2021). Estimating bar graph averages: Overcoming within-the-bar bias. *i-Perception, 12*(1), 2041669520987254

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016, May). When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5092–5103).

Kerns, S. H. (2019). Evaluating anecdotal Lyme disease treatments for Borrelia-inhibitory properties. Honors Thesis, Wellesley College. https://repository.wellesley.edu/object/ir858.

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.

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

Kim, Y. S., Reinecke, K., & Hullman, J. (2017b). Data through others' eyes: The impact of visualizing others' expectations on visualization interpretation. *IEEE Transactions on Visualization and Computer Graphics, 24*(1), 760–769.

Kimchi, R., Yeshurun, Y., & Cohen-Savransky, A. (2007). Automatic, stimulus-driven attentional capture by objecthood. *Psychonomic Bulletin & Review, 14*(1), 166–172.

Krippendorff, K. (2011). "*Computing Krippendorff's alpha-reliability*." Philadelphia: Annenberg School for Communication Departmental Papers.

Landau, B., Hoffman, J. E., & Kurz, N. (2006). Object recognition with severe spatial deficits in Williams syndrome: Sparing and breakdown. *Cognition, 100*(3), 483–510.

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

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America, 99*(14), 9596–9601.

Lilienfeld, S., Lynn, S., & Namy, L. (2017). *Psychology: From inquiry to understanding* (4th ed.). New York, NY: Pearson.

Litman, L., & Robinson, J. (2020). *Conducting online research on Amazon Mechanical Turk and beyond* (Vol. 1). Thousand Oaks, CA: Sage.

Marino, A. C., & Scholl, B. J. (2005). The role of closure in defining the "objects" of object-based attention. *Perception & Psychophysics, 67*(7), 1140–1149.

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

Mogull, S. A., & Stanfield, C. T. (2015, July). Current use of visuals in scientific communication. In *2015 IEEE International Professional Communication Conference (IPCC)* (pp. 1–6). IEEE.

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. *The Wiley Blackwell Handbook of Judgment and Decision Making, 2*, 182–209.

Myers, D., & DeWall, N. (2017). *Psychology* (12th ed.). New York, NY: Worth.

Natter, H., & Berry, D. (2005). Effects of active information processing on the understanding of risk information. *Applied Cognitive Psychology, 19*(1), 123–135.

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.

Nguyen, F., Qiao, X., Heer, J., & Hullman, J. R. (2020). Exploring the effects of aggregation choices on untrained visualization users' generalizations from data. *Computer Graphics Forum, 39*.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., . . . Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities.* Hoboken, NJ: John Wiley & Sons.

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.

Oxford University Press (n.d.) Error. In Google's English Dictionary. Retrieved January 9, 2021, from https://www.google.com/search?q=error+definition.

Oxford University Press (n.d.) Bias. In Google's English Dictionary. Retrieved January 9, 2021, from https://www.google.com/search?q=bias+definition.

Oxford University Press (n.d.) Confusion. In Google's English Dictionary. Retrieved January 9, 2021, from https://www.google.com/search?q=confusion+definition.

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

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

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.

Pfister, R., Schwarz, K. A., Janczyk, M., Dale, R., & Freeman, J. B. (2013). Good things peak in pairs: A note on the bimodality coefficient. *Frontiers in Psychology, 4*, 700.

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

Richler, J. J., Wilmer, J. B., & Gauthier, I. (2017). General object recognition is specific: Evidence from novel and familiar objects. *Cognition, 166*, 42–55.

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

Seipel, S., & Lim, N. J. (2017). Color map design for visualization in flood risk assessment. *International Journal of Geographical Information Science, 31*(11), 2286–2309.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin, 134*(2), 207–222.

Smith, A. D. (2009). On the use of drawing tasks in neuropsychological assessment. *Neuropsychology, 23*(2), 231.

Stern, E., Aprea, C., & Ebner, H. (2003). Improving cross-content transfer in text processing by means of active graphical representation. *Learning and Instruction, 13*(1), 191–203.

Steuer, F. B., & Ham, K. W. (2008). Psychology textbooks: examining their accuracy. *Teaching of Psychology, 35*(3), 160–168.

Sutherland, C. A., Burton, N. S., Wilmer, J. B., Blokland, G. A., Germine, L., Palermo, R., Collova, J. R., … Rhodes, G. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences, 117*(19), 10218–10224.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520–522.

Tufte, E. R., & Graves-Morris, P. R. (1983). *The visual display of quantitative information*. Cheshire, Conn: Graphics Press.

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

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

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

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

Wilmer,J. B., & Nakayama, K. (2007). Two distinct visual motion mechanisms for smooth pursuit: Evidence from individual differences. *Neuron, 54*(6), 987–1000.

Wilmer, J. (2008). How to use individual differences to isolate functional organization, biology, and utility of visual functions; with illustrative proposals for stereopsis. *Spatial Vision, 21*(6), 561–579.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology, 29*(5-6), 360–392.

Wolfe, J., Kluender, K., Levi, D., Bartoshuk, L., Herz, R., Klatzky, R., … Merfeld, D. (2020). *Sensation and Perception* (6th ed.). Oxford, UK: Oxford University Press.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233–235.

Zubiaga, A., & MacNamee, B. (2016) Graphical perception of value distributions: an evaluation of non-expert viewers' data literacy. *Journal of Community Informatics, 12*(3).