DECISIONS, VARIATION, AND VISUALIZATION: A NOVEL INSTRUMENT FOR DECISION MAKING UNDER VARIABILITY

Zachary del Rosario, Erika Saur, Jin Ryu, Jessica Lin, and Jeremy Wilmer

ABSTRACT

To make statistically sound choices, decision makers must respond to the consequences of variability—to target variability. Previous work has shown that targeting is beneficial, but there are questions about its prevalence and contributing factors. With a novel instrument and a demographically representative, randomly assigned n=306 sample, we find that most U.S. adults target variability by default, even with minimal information (95% CI [78%, 92%]). Surprisingly, the same information presented via a standard bar graph inhibits targeting, while showing raw data enhances targeting. These results support a resource view of statistically sound decision making (over a deficit view), and contradict widespread ideas in data visualization (that raw data will overwhelm).

Keywords: *Statistical thinking, decision making, variability, instrument, data visualization*

1. INTRODUCTION

Variability is core to statistical thinking (Garfield & Ben-Zvi, 2005; C. Wild et al., 2018; C. J. Wild & Pfannkuch, 1999; Wood et al., 2018). Statistics educators have developed a variety of frameworks to articulate and ultimately teach statistical thinking. These frameworks include ideas about study design, data collection, and modeling (Peters, 2011), selecting inferential tests (Alacaci, 2004), having an intuitive sense for variability (Garfield & Ben-Zvi, 2005), and asking good statistical questions (Arnold & Franklin, 2021). However, with the emergence of data science, there has been a growing interest in the intersection of statistical thinking with other disciplines. This suggests a need for interdisciplinary work and new framings; as Hicks and Irizarry (2018) suggest, traditional statistics education is often misaligned with solving real-world problems that emerge in other disciplines. Within the statistics education community, the neglected, acknowledged, or targeted (NAT) taxonomy frames how a decision maker recognizes variability and incorporates it in a decision (del Rosario, 2024). However, this prior work was qualitative—studying a small number of engineers—leaving open questions about more general decision making under *variability*.

Two communities have pursued similar work under the related (but distinct) framing of decision making under *uncertainty*. The behavioral economics community has long studied decision making under uncertainty (Gigerenzer & Gaissmaier, 2011; Kahneman et al., 1982), but has two competing schools of thought. Borrowing language from the education community (Dewsbury, 2020; Findley & Lyford, 2019), the *deficit view* asserts that humans are generally flawed in their reasoning under uncertainty (Kahneman et al., 1982), while the *resource view* asserts that humans can make decisions that are valid for a given context (Gigerenzer & Gaissmaier, 2011). Meanwhile, the data visualization community has tested a variety of means to communicate uncertainty (Padilla et al., 2021), but there exists a pervasive idea that lay readers will feel "overwhelmed" if presented with information about uncertainty (Hullman, 2020). This idea encourages some authors to use misleading visuals, such as average bar charts (Kerns & Wilmer, 2021).

The goals of this study are twofold: 1. Develop a novel instrument to measure behavior according to the NAT taxonomy, and 2. Deploy the instrument to study how often U.S. adults target variability in everyday tasks, comparing presentations of data. Studying decision making in everyday tasks will provide foundational empirical results on how to regard targeting: either as a difficult behavior (following the deficit view) or as a natural behavior that is sometimes disrupted (following the resource view). We experimentally manipulate the presentation of data to test the capacity of lay readers to make use of information about variability.

2. BACKGROUND

In this section, we briefly review the two major lines of research that frame the present work. Decision making under uncertainty has long been studied in behavioral economics, while the data visualization community has investigated how to communicate information about uncertainty.

2.1 DECISION MAKING UNDER VARIABILITY

The behavioral economics community asserts that humans use simplified procedures to make decisions under uncertainty—heuristics. Tversky and Kahneman (1974) assert that “(p)eople rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations.” Heuristics themselves are "efficient cognitive processes, conscious or unconsciousness, that ignore part of the information" (Gigerenzer & Gaissmaier, 2011). Ultimately, heuristics are used to understand how humans make decisions under uncertainty.

Behavioral economics' use of "uncertainty" is only roughly aligned with statistics education's notion of "variability." Both refer to a lack of certainty (Tversky & Kahneman, 1974; C. J. Wild & Pfannkuch, 1999); however, the uncertainty of behavioral economics tends to emphasize scenarios where there is a single true value. For instance, Tversky and Kahneman (1974) provide examples such as "the outcome of an election, the guilt of a defendant, or the future value of the dollar", or the probability of heads in a coin flip. In a review of Kahneman and Tversky's work, Gigerenzer (1996) found 5 of 13 instances to consider single-event probabilities. Statisticians consider uncertainty in true values when practicing (frequentist) inference, but also consider situations that exhibit *real* variability (C. J. Wild & Pfannkuch, 1999), such as the inconsistency of manufactured items, the differing heights of individuals, and the constant fluctuation within a living organism. In this way, the variability of statistics education is a superset of the uncertainty generally considered by behavioral economics. We therefore focus on variability as it is conceived in statistics education.

Variability is a key feature of everyday life and human behavior. As Wild and Pfannkuch (1999) note, "Variation is an observable reality. It is present everywhere and in everything." Therefore, we chose to study decision making under variability. An important goal of many studies is to produce generalizable results; the principle of ecological validity (Kerns & Wilmer, 2021) states that designing real-world behavioral tasks will promote generalization to other scenarios. For investigating human behavior, this means picking tasks grounded in the everyday experiences of the target population that exhibit variability. For the present work, we have selected tasks that naturally exhibit real variability, such as an auction (variability in bids), a driving commute (variability in commute time), and a footrace (variability in race times). By studying tasks with real variability (C. J. Wild & Pfannkuch, 1999), we are broadening the space of empirical results beyond the single-event scenarios often considered in behavioral economics.

While researchers of decision making under uncertainty are united in their focus on uncertainty, there is deep disagreement about the decision making abilities of humans. Tversky and Kahneman (1974)

interpreted their collective results as evidence of persistent cognitive illusions, writing "(w)hat is perhaps surprising is the failure of people to infer from lifelong experience such fundamental statistical rules as regression towards the mean, or the effect of sample size on sampling variability." Cognitive illusions have similarly been identified in law (Saks, 1981), management science (Bazerman & Moore, 2012), and statistics (Cooper & Shore, 2008; Konold, 1989; Shaughnessy, 1977). However, Gigerenzer (1991) has reframed some of the single-event research questions of Kahneman and Tversky and shown that such cognitive illusions "disappear." Gigerenzer (1996) has further argued that the heuristics of behavioral economics "at once explain too little and too much… too much, because, post hoc, one of them can be fitted to almost any experimental result." More recent work has shown that supposed cognitive illusions generate accurate judgements in everyday contexts (Gigerenzer & Gaissmaier, 2011). Maintaining the language common in the education community (Dewsbury, 2020; Findley & Lyford, 2019), we adapt the terms *deficit view* (a belief in persistent cognitive illusions) and *resource view* (a belief in generally available, productive decision making processes) to contrast these competing schools of thought. An important goal of this work is to test whether targeting variability is better understood through a deficit or resource view.

The present work builds on the methodological insights and topical focus of these prior works. By designing everyday tasks exhibiting real variability, we contribute to statistics education's focus on variability (C. J. Wild & Pfannkuch, 1999). Given the documented difficulties applying traditional heuristics as an explanatory tool (Gigerenzer, 1996), we instead operationalize the neglected, acknowledged, targeted (NAT) taxonomy (del Rosario, 2024) from statistics education in a novel survey. The NAT taxonomy rungs are defined as:

1. Neglected: Participant's analysis neglects variability, usually by reporting a single value.
2. Acknowledged: Participant's analysis acknowledges variability, but does not respond to the consequences of variability.
3. Targeted: Participant's analysis acknowledges variability and responds to the consequences of that variability.

Finally, we assess whether the deficit or resource view is more appropriate for understanding targeting variability in the U.S. adult population.

2.2 DATA VISUALIZATION AND DECISION MAKING

Data visualization is a powerful means to make information available (Tukey, 1977) and thus connects to decision making under variability. There are many texts of practical advice for data visualization (Tufte, 2001; Wilke, 2019); however, these advice texts typically do not include empirical studies of human behavior. Such empirical work is important, as human behavior can run counter to prevailing visualization wisdom. For instance, error bars are ubiquitous in scientific communication, but both lay users (Hullman et al., 2015) and experts (Belia et al., 2005; Zhang et al., 2023) often misinterpret error bars. There is an ongoing need for empirical work on the challenges and opportunities for how data visualization can support decision making under variability.

The relevant empirical work has explored a wide variety of options for displaying, or withholding, variability. Some researchers have investigated highly specialized methods, such as iconic representations (Bisantz et al., 2005; Finger & Bisantz, 2002). Other researchers have used displays that are closer to standard practice in statistical graphics. For example, authors have tried animation (Hullman et al., 2015), ensemble displays (aka "hurricane plots") (Liu et al., 2017), error bars and their alternatives (Correll & Gleicher, 2014; Fernandes et al., 2018), and aesthetic edits (e.g., fuzziness or transparency) (MacEachren et al., 2012). However, few of these works assess decision making.

Additionally, the present work tackles a key, current area of debate in data visualization: if and when to plot raw data. Methods such as dotplots follow Tufte's (2001) principle "above all else show the data." However, there exists a common belief among data visualization professionals that any information on variability will "overwhelm" a lay reader, leading many practitioners to omit such information entirely (Hullman, 2020). Consequently, it is still common for data to be provided via "simple" charts that depict only average values—via bars, lines, or dots—occasionally with superimposed error bars to communicate statistical uncertainty of summary statistics, but rarely with any indication of the underlying data distribution (Correll & Gleicher, 2014; Kerns & Wilmer, 2021; Newman & Scholl, 2012).

Studies of these "simple" graphs of averages have revealed striking misinterpretations. For example, Kerns and Wilmer, using their novel Draw Datapoints on Graphs measure (2021), identified three common, severe miscommunications caused by bar charts of averages: the Bar-Tip Limit (BTL) error, in which 1 in 5 readers assumed that all data values fell within the bar tip (Kerns & Wilmer, 2021; Wilmer & Kerns, 2022); the Dichotomization Fallacy, where 1 in 3 readers assumed such minimal variability around averages that it implied zero overlap between groups or conditions (Wilmer & Kerns, 2022); and the Uniformity Fallacy, where 1 in 3 readers assumed a flat, uniform distribution—failing to recognize the standard tapering, or "tails," typical of real data (Wilmer & Kerns, 2022, 2025). Notably, all three of these miscommunications involve variability: its location, magnitude, and shape, respectively.

While these findings suggest major misconceptions, consistent with a deficit view, they all hinge on one class of visual—showing *summaries* rather than *showing the raw data*. Other empirical work suggests that humans are adept at interpreting raw data. Zhang et al. (2023) found that statistical confusions of experts disappeared when presented with raw data, while Wang et al. (2025) found that people were better at locating the averages in a graph when presented with raw data, compared with a treatment where the averages were *explicitly* shown using bar or line graphs. These studies support a resource view of graph interpretation with raw data; a key question we seek to answer is if sound graph interpretation translates to sound decision making under variability.

The present work builds on these results while contributing novel findings. We provide empirical assessment of the "variability will overwhelm" hypothesis by comparing three treatments: showing variability, simplifying with a bar chart, and avoiding visualization altogether with a text-only control. To display raw data with variability, we use sinaplots (Sidiropoulos et al., 2018). Drawing on studies of bar chart fallacies, we demonstrate strategies to detect both the BTL error (Kerns & Wilmer, 2021) and a variant of the Dichotomization Fallacy (Wilmer & Kerns, 2022) we refer to as *excessive precision*. Given that both fallacies constitute a misunderstanding of variability, they have direct connections to neglecting variability.

3. METHODS

This section describes the design and development of the survey instrument, data analysis approach, and sample. This research was determined to be IRB exempt by the Brandeis IRB, protocol #24274R-E. The full survey and analysis preregistration are available in open access via osf.io/u78bw. For convenient reference, all symbols used in this manuscript are defined inline but also reported in Table A1.

3.1 SURVEY DESIGN

The survey was designed according to the Measurement of Abstract Graph Interpretation (MAGI) design principles (Kerns & Wilmer, 2021). These principles have been shown to promote both general usability and measurement validity of graph interpretation survey items. The principles are summarized

in Table 1 and referenced throughout this section. The survey itself is available in open access form via osf.io/u78bw.

The primary body of the survey consists of five distinct tasks,[1] each with a baseline condition and a modified condition. Each task poses an everyday scenario and requests a decision (a number) for each condition, for ten total decisions. The task contexts are "everyday" in the sense that the target population should have some familiarity with the context: an auction, commuting to work, a footrace, a math test, and parking tickets. Choosing familiar contexts follows MAGI principle *ecological validity*, which promotes generalization to other real-world scenarios (Kerns & Wilmer, 2021). In particular, the contexts naturally exhibit real variability, promoting the external validity of findings to other everyday scenarios with variation.

Table 1. The MAGI design principles (Kerns & Wilmer, 2021) and their execution in this study.

MAGI Principle	**Goal**	**Execution**
Expressive Freedom	Allow unexpected responses	Elicited responses as a number, rather than a fixed choice.
Limited Instructions	Minimize effect of instructions	Iteratively refined text to minimize instructions.
Limited Mental Transformations	Promote accuracy by avoiding need for mental translation	Solicited responses in the same basic form as stimulus (numbers). Ensured that the units of requested responses match the units of data provided.
Ground truth Linkage	Enable objective assessment of participant responses	Selected stimuli and response for which a correct answer exists.
Ecological Validity	Support generalization of findings to other real-world scenarios; Minimize confounds	Used tasks that will be familiar to participants.
Information Richness	Enable higher-resolution analysis	Collected multiple responses per task: data perception (min and max), answers for baseline and modified conditions.

Aligned with the experimental goals of the project, we devised three treatment variations on the survey, presenting task information as: a sinaplot (Sidiropoulos et al., 2018), text alone, or a bar chart. Examples of all three treatments from the auction task are shown in Table 2. Participants were randomized into one treatment at the start of the survey—all tasks were presented in the same treatment once assigned. The order of tasks was also randomized (per individual) to prevent consistent learning effects.

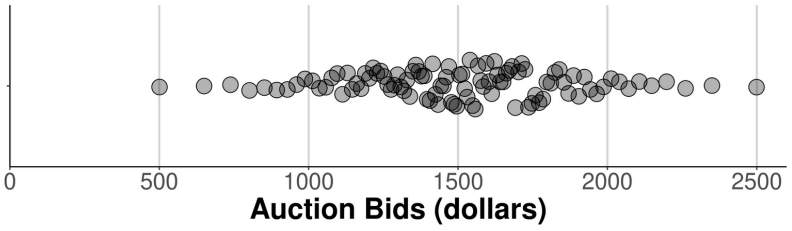
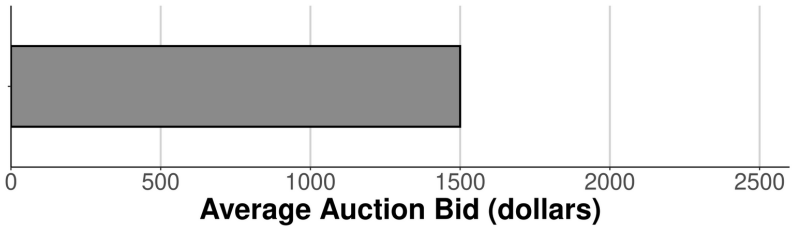
[1] The survey also included a sixth "Game" task that was designed to be more difficult than all other tasks. While participants indeed targeted at a lower rate on this task, the Game task also did not exhibit strong evidence of ecological validity. Therefore, we exclude the Game task from the present work.

Each task is designed with a data minimum l_d, data maximum u_d, and a symmetric distribution between this range. Participants report their perceived data minimum l_p, data maximum u_p, and chosen answer for the baseline $a_{p,0}$ and modified $a_{p,1}$ conditions. We chose to elicit numbers rather than present fixed choices following MAGI principle: *expressive freedom*, and requested all answers in the same units as the task data following MAGI principle: *limited mental transformations* (Kerns & Wilmer, 2021).

Tasks were designed to have a correct answer for both the baseline $a_{d,0}$ and modified conditions $a_{d,1}$, following MAGI principle: *ground truth linkage* (Kerns & Wilmer, 2021). This allows for an absolute assessment of participant answers. However, since each task presents two conditions, we are able to compare participant's baseline and modified condition answers. This serves as a judge of participant's task comprehension, following MAGI principle: *information richness* (Kerns & Wilmer, 2021).

We carried out more than ten rounds of piloting over a year and sought to maximize task comprehension while not reducing the survey to a catch trial. In revising, we paid particular attention to simplifying tasks, following MAGI principle: *limited instructions* (Kerns & Wilmer, 2021).

Table 2. Comparison of the treatment stimuli for the Auction task.

Sinaplot	In an auction, a painting receives 100 bids, with the bids shown by the dots on the following plot. 
Text	In an auction, a painting receives 100 bids, with an average bid of $1,500.
Bar	In an auction, a painting receives 100 bids, with the average bid shown by the following plot. 
All treatments	Guess the lowest bid: [Number entry: l_p] Guess the highest bid: [Number entry: u_p] Guess the bid that would guarantee you win the painting, without paying too much. [Number entry: a_p]

3.2 DATA ANALYSIS

To interpret survey results, we developed a null model to simulate responses according to a random guessing strategy. The null model does not include any bias towards targeting, but does respect the problem givens and ensures internal consistency of the simulated answer. The null model produces uniform random draws in two stages:

- Stage 1: Select a minimum $l_p \sim U(l_d, m_d)$ and maximum $u_p \sim U(m_d, u_d)$
- Stage 2: Select an answer $a_p \sim U(l_p, u_p)$

Both participant data and null model simulation data are analyzed following the same protocol.

To operationalize data analysis, we compute several quantities based on participant responses. The *perceived data width* $w_p = \frac{u_p - l_p}{u_d - l_d}$ is used to detect cases of excessive precision. We judge a participant's task comprehension by checking whether the ordering of their baseline and modified condition answers matches that of the ground truth answers. This is operationalized in terms of the sign function $sign(x)$, which is used to compute the signs of participant $g_p = sign(a_{p,0} - a_{p,1})$ and ground truth $g_d = sign(a_{d,0} - a_{d,1})$ answer differences. *Task comprehension* is judged if $g_p = g_d$.

Similarly, we operationalize a low-resolution comparison of participant answers a_p with ground truth answers a_d using the sign function. If a participant answer is on the same side of the data average m_d as the ground truth answer, then the signs of the differences $s_p = sign(a_p - m_d)$, $s_d = sign(a_d - m_d)$ will match $s_p = s_d$.

To operationalize a high-resolution comparison of participant answers a_p with ground truth answers a_d, we define *u-normalization* as an affine transformation

$$u_p = \frac{a_p - l_p}{u_p - l_p}, \; u_d = \frac{a_d - l_d}{u_d - l_d}$$

U-normalization transforms bound-respecting answers to the unit interval and is depicted schematically in Figure 1. Mapping participant and ground truth answers allows judgement of participant answers in terms of *their* perception of a task's variability, rather than holding them to a higher standard of absolute error. The u-normalized answers are used to compute a relative error $e_{rel} = |u_p - u_d|$, while the original answers are used to compute an absolute error $e_{abs} = |\frac{a_p - a_d}{u_d - l_d}|$.

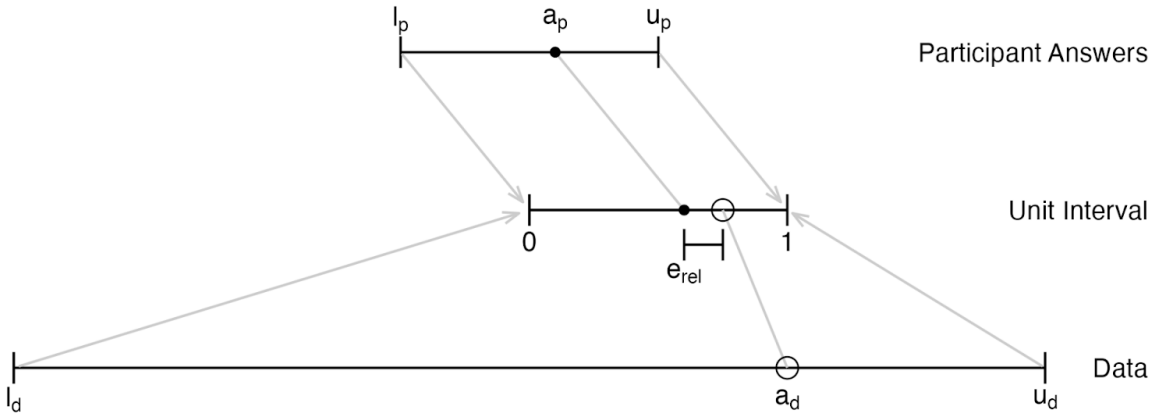


Figure 1. Schematic depiction of the relative error e_{rel} computation, based on u-normalization of participant answer and ground truth answer.

For linear modeling, we also define the *log-floored relative error* $l_f = log(e_f)$. To avoid undefined values, we first floor the absolute error e_f by replacing zero values of e_{rel} with the smallest nonzero value in the dataset (across all treatments).

Following our preregistration plan, we refined the Neglected, Acknowledged, Targeted (NAT) taxonomy using ideas from the background (Excessive Precision and the Bar Tip Limit error), and implemented it in terms of quantities designed into and elicited from the survey. The operationalization is given in Table 3.

Table 3. Operationalization of the NAT Taxonomy using survey data.

Step	**Definition**	**Description**	**Code**
1	$w_p < 0.1$	Perceived data width is small	Neglected: Excessive Precision (N:EP)
2	$u_p < m_d$	Perceived data max is less than data average	Neglected: Bar Tip Limit (N:BTL)
3	$e_{rel} < 0.1$	Participant answer is near ground truth answer	Targeted:Strict (T:Strict)
4	$s_p = s_d$	Participant answer is on correct side of data average, compared to ground truth answer	Targeted:Side (T:Side)
5	(True)	Otherwise	Acknowledged (A)

3.3 SAMPLE

We recruited a sample of U.S. adults, representative of the U.S. population according to age, sex, and ethnicity[2]. The sample demographics are reported in Table 4. The sample size was pre-determined based on precision analysis (Cumming & Calin-Jageman, 2024) and documented in our study pre-registration. Accounting for a training phase pass rate of 85% (determined from piloting), a target 95% margin of error at 0.3 units of standard deviation multiplied by three treatments yields a total sample size of n=306.

Within the sample, participants were randomized to one of the three treatments (see Tab. 5). According to a series of ANOVA tests, the treatment subsamples are not significantly different in age ($F[2,303] = 1.0$, $p = .371$), dummy-coded sex ($F[2,303] = 0.39$, $p = .678$), or dummy-coded ethnicity ($F[2,303] = 0.92$, $p = .401$).

Table 4. Sample demographics. Age figures do not add to 100% due to rounding.

Age	Male	Female	Ethnicity	
75-79	0.7%	0.3%	**Asian**	6.5%
70-74	0.7%	1.0%	**Black**	11.8%
65-69	2.9%	4.2%	**Mixed**	11.1%
60-64	6.9%	6.5%	**Other**	7.9%
55-59	6.5%	8.2%	**White**	62.7%
50-54	2.9%	3.3%		
45-49	5.2%	4.9%		
40-44	4.9%	3.3%		
35-39	3.3%	4.9%		
30-34	5.9%	3.6%		
25-29	3.3%	4.9%		
20-24	5.2%	4.9%		
18-19	0.7%	1.0%		

[2] According to the U.S. Census simplified categories.

Table 5. Sample sizes for each treatment, unequal due to randomization.

Treatment	Total	Passed Training
Sinaplot	109	95
Text	102	91
Bar	95	89
(Grand totals)	306	275

To promote data quality, we included a training phase that preceded the main body of the survey. This included the simple stimulus "A past data set has 100 values, with an average of 25", a few factual questions, and two attention checks. To pass the training phase, participants needed to pass both attention checks, provide a guess for the minimum (maximum) value in the dataset that was less than (greater than) or equal to the provided average, and to report the stated average of 25. Of our n=306 sample, 31 participants failed the training phase (~90% pass rate), for an analyzed sample size of n=275. There was no significant difference in the number of failed participants among the three treatments ($F[2,303] = 1.22$, $p = .296$).

4. RESULTS

The results section is structured around different levels of analysis, starting at the highest level of aggregation of the three treatments. We then drill down to a task level aggregation of the data, followed by an analysis at the numeric answer level. We conclude with an analysis of instrument reliability and validity.

4.1 TREATMENT LEVEL

Figure 2 shows the percentage of participants who targeted variability on a specific number of the tasks, combining strict and side sense targeting codes. Strikingly, over 45% of participants in the sinaplot treatment targeted variability on 10 out of 10 survey tasks, with a 95% CI [38%, 57%]. Simulated results from the null model indicate that 10 out of 10 targeting is exceedingly rare under random guessing (0.11%). We find that most U.S. adults target variability by default (on a majority of tasks), even with text alone (95% CI [78%, 92%])—high compared to the null model (41%). These results indicate that participants are responding to the variability in the tasks—the core idea of targeting variability (del Rosario, 2024).

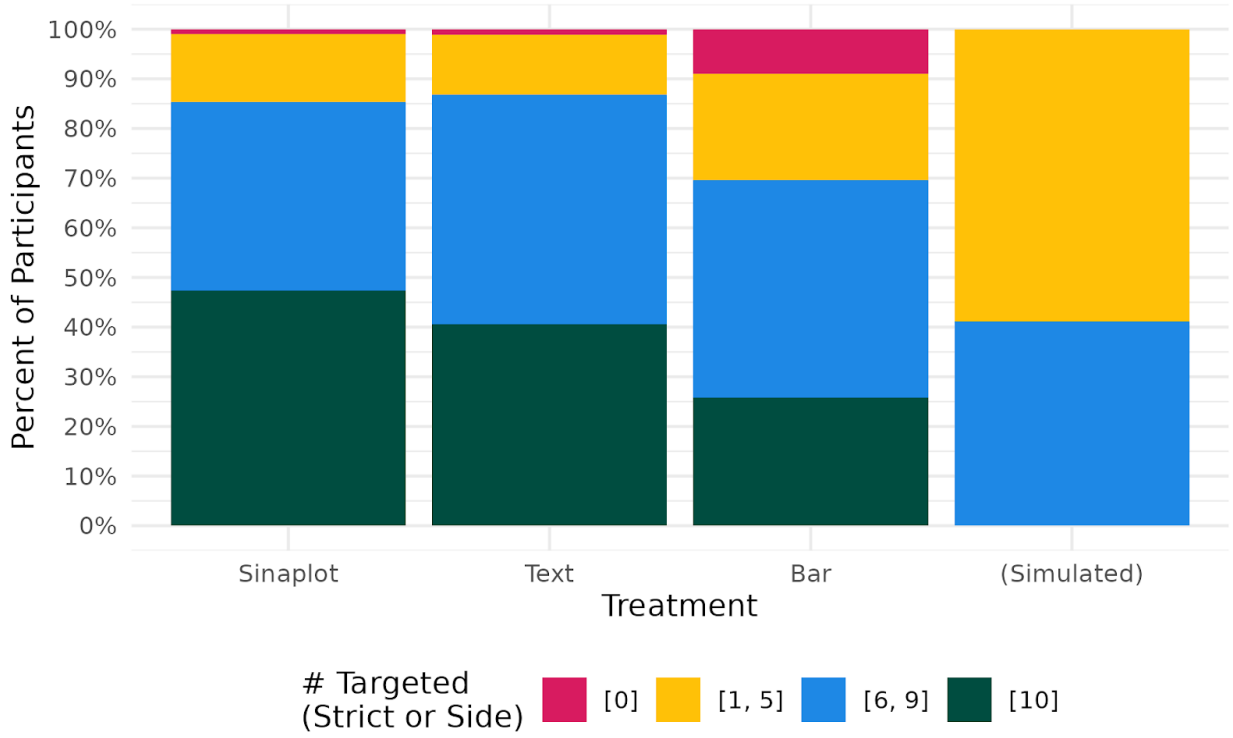


Figure 2. Number of targeted responses (combined Strict or Side) disaggregated by treatment. For comparison, results from 10,000 runs of the null model (Simulated) are provided.

The results corroborate key findings from previous studies. Prior qualitative analysis of interviews suggested that targeting variability depends on understanding the consequences of variability in the task. These are operationalized as boolean factors: combined strict & side targeting (Tab. 3) and task comprehension ($g_p = g_d$). A Chi-squared test for dependence between these factors is highly significant (χ^2 (2, N=275) = 255.1, p < .001) and yields a very strong association at $\Phi = 0.27$ (Akoglu, 2018).

Prior work also suggested that targeting variability should be thought of as a cognitive resource—a widespread and beneficial behavior, at least among engineering undergraduate students (del Rosario et al., 2024). Overall, nearly all participants targeted at least once in the sinaplot treatment (95% CI [94.2%, 99.8%]), the text treatment (95% CI [94.0%, 99.8%]), and the bar chart treatment (95% CI [91%, 95%]). At least in the right-side sense, most U.S. adults can target variability in everyday tasks, even when misled by the data presentation, e.g., as a bar chart.

4.2 TASK LEVEL

Table 6 shows consequence comprehension ($g_p = g_d$) rates disaggregated by treatment and task. Generally, a majority of participants understood the consequences of variability in each of the five tasks, excepting the Parking task in the text and bar treatments. These results demonstrate the ecological validity of the tasks.

Table 6. Consequence comprehension rates based on comparing baseline and modified conditions.

Task	**Sinaplot**	**Text**	**Bar**
Footrace	89.47%	82.42%	71.91%
Math Test	87.37%	85.71%	87.64%
Auction	81.05%	81.32%	79.78%
Commute	78.95%	80.22%	79.78%
Parking	69.47%	45.05%	49.44%

Figure 3 shows the percentage of participants responses according to NAT code (Tab. 3), disaggregated by treatment, task, and modified/baseline condition. These results show that bar charts do more harm than good for decision making. Overall, targeting variability (combined T:Strict and T:Side) is lower in the bar treatment (67%), compared to the text treatment (81%), with a significant difference (Welch's $t(1857.3) = -6.90$, $p < .001$). Given the randomized treatment assignment and comparability in demographic factors, this difference is fully accounted for by bar chart fallacies (BTL and EP). This is corroborated by the higher rate of N:BTL and N:EP codes in the bar chart condition.

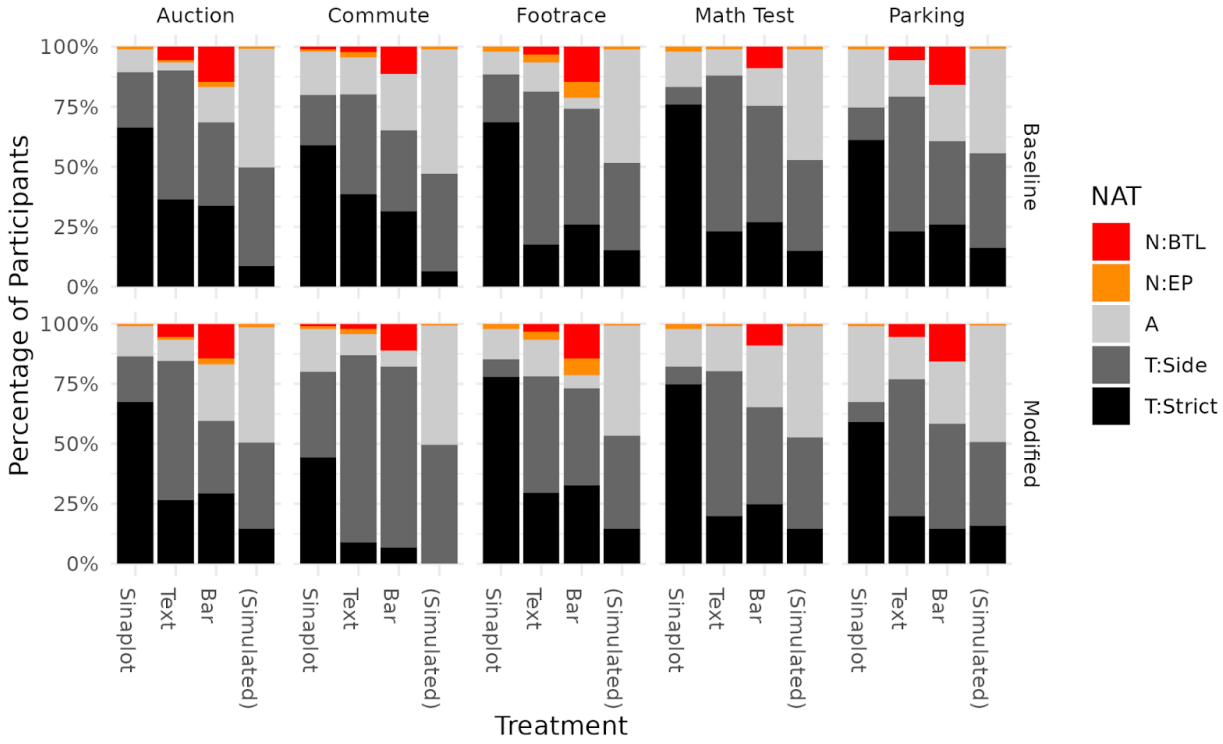


Figure 3. NAT coding of participant responses to baseline and nudged tasks. BTL = bar tip limit, EP = excessive precision, Side refers to an answer on the correct side (of data average), Strict refers to relative error less than 10%. A sample of 1000 simulated draws from the null model is also reported.

However, there is no significant difference between results in the text (24%) and bar (24%) treatments in terms of strict targeting alone (Welch's $t(1955.2) = 0.36$, $p = .721$). Given that T:Strict

corrects for perceived data range (see Fig. 1), this suggests that many participants are unable to fill in a realistically-shaped distribution to guide their decision. This stands in contrast with the sinaplot treatment results.

Sinaplots support sound decision making under variability. Figure 3 shows that strict targeting is more frequent in the sinaplot treatment (62%), compared to the text treatment (24%). In terms of T:Strict, participants in the sinaplot treatment targeted variability significantly more than those in the text treatment (Welch's $t(2098.1) = 19.1$, $p < .001$). In terms of combined T:Side and T:Strict codes, there is no significant difference between the sinaplot and text treatments (Welch's $t(2107.4) = -1.5$, $p = .122$). This contradicts the common notion that showing uncertainty information is confusing (Hullman, 2020)—sinaplot treatment participants did no worse in a side targeting sense, and did considerably better in a strict targeting sense.

4.3 NUMERIC ANSWER LEVEL

Figure 4 shows u-normalized participant responses (see Fig. 1), grouped by treatment, task, and baseline/modified condition. Ground truth answers are shown as a solid vertical line. Concentration around the ground truth answer is greatest in the Sinaplot treatment, corroborating the T:Strict coding results shown in Figure 3. However, concentration around the ground truth answer is greater in all treatments, compared with the null model results.

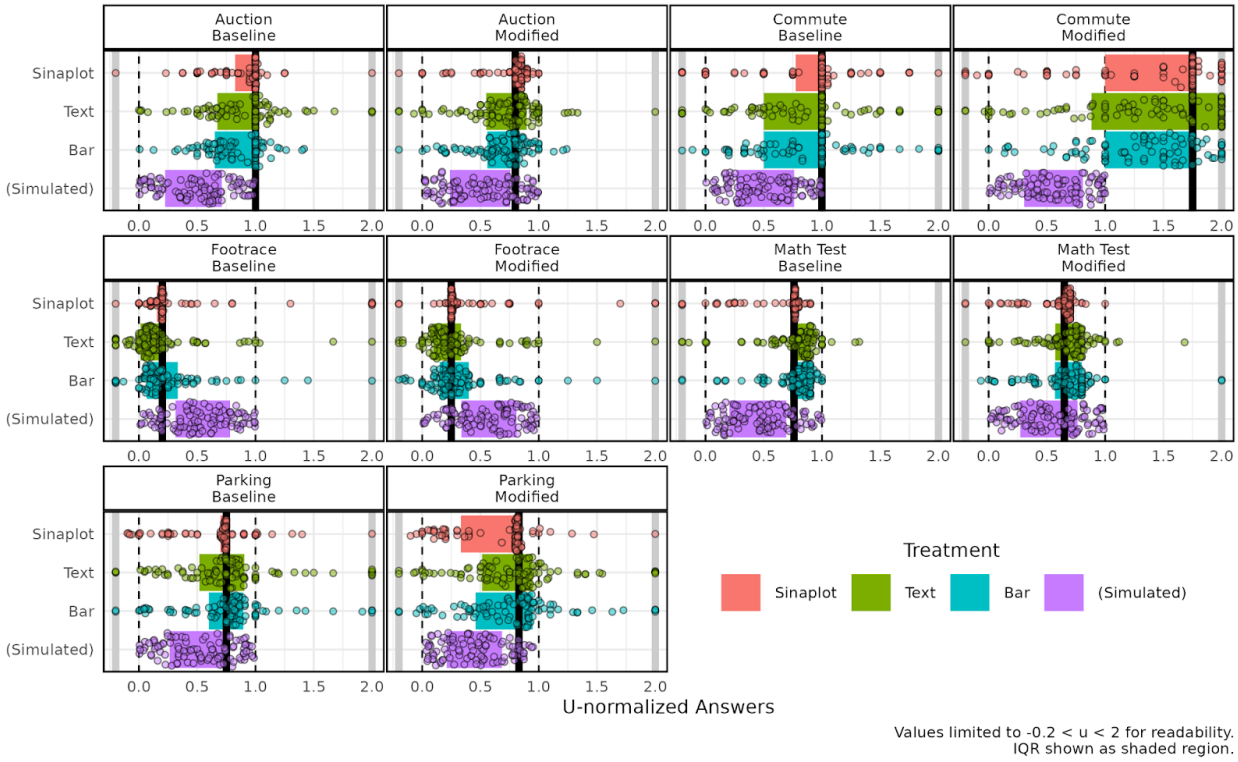


Figure 4. U-normalized answers for all tasks, conditions, and treatments. For readability, u values are limited to a minimum of -0.2 and maximum of +2.0 (Sinaplot: 34, Text: 62, Bar: 27). Additionally, 100 simulated responses are drawn from the null model. Ground truth answers are shown as a solid black vertical line, and the interquartile range is shown as a shaded rectangle.

Results in Figure 4 provide hints for the reasoning behind non-targeted answers. For instance, clusters of low u-normalized answers (e.g., in the Math Test and Parking tasks) suggest participants who mis-interpreted the task and mirrored the ground truth answer over the average value.

The presentation of data has a strong effect on (strict) targeting. Conducting a three-way ANOVA on the log-floored relative error (l_f, see A1), we find that, among preregistered variables, the treatment has the largest effect (Tab. 7), explaining 9.4% of the variance.

Table 7. Three-way ANOVA of log-magnitude-floored relative error on pre-registered factors.

Model Term	**Variance Explained (%)**	**DoF**	**p**
(Residuals)	80.7	3301	NA
Treatment	9.4	2	< .001
Task	6.8	5	< .001
Consequence Comprehension	3.0	1	< .001

4.4 INSTRUMENT RELIABILITY AND VALIDITY

Going beyond the preregistered analysis, we present results that support the reliability and validity of the instrument. Under classical test theory, the notions of reliability and validity rest upon an assumption of a latent variable that affects the measurement (DeVellis, 2017). For the present survey, this latent variable is *tendency to target* (T3) as an individual-level trait. Adding an individual identifier term to the ANOVA explains 11.8% of the variance—the highest fraction among all predictors (Tab. 8). This result supports the existence of T3.

Table 8. Multi-way ANOVA attributing variance in the floored relative error to pre-registered factors, plus an individual identifier.

Model Term	**Variance Explained (%)**	**DoF**	**p**
(Residuals)	68.9	3017	NA
Individual	11.8	272	< .001
Treatment	9.4	2	< .001
Task	6.7	5	< .001
Consequence Comprehension	3.1	1	< .001

To quantify the reliability of the instrument in measuring T3, we study the association between relative errors across tasks. To account for extreme outlying values (see Fig. 4), we compute Spearman correlations (Spearman, 2010) by first rank transforming the relative errors (grouped by treatment, task,

and condition). Combining these correlations across all baseline and modified conditions yields a Cronbach's $\alpha = 0.74$, suggesting good reliability of the instrument, commensurate with the early stages of a psychometric study (Nunnally & Bernstein, 1994).

Figure 5 shows a similar analysis where the relative error is first averaged between the baseline and modified condition answers before rank transformation (grouped by treatment and task). The resulting spearman correlations between survey tasks are all significant at the 95% level. Given that all tasks were designed to measure targeting variability, the pairwise agreement between relative errors supports the validity of the instrument in measuring T3.

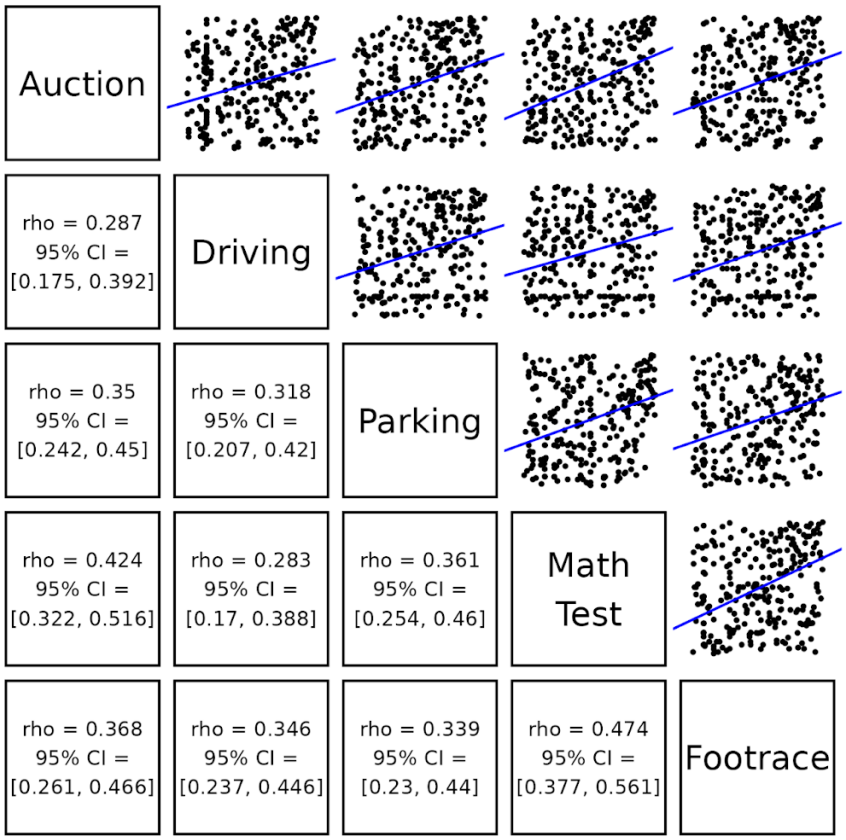


Figure 5. Scatterplot matrix showing the percentile ranks for the relative error. Lower panels report Spearman (rank) correlations and confidence intervals.

5. DISCUSSION

This study developed and deployed a novel survey instrument to measure decision making in everyday scenarios exhibiting variability. The instrument operationalized the neglected, acknowledged, targeted (NAT) taxonomy (del Rosario, 2024), which frames the universe of possible decisions in response to variability.

5.1 KEY FINDINGS

Nearly all participants targeted at least once in the sinaplot treatment (95% CI [94.2%, 99.8%]), the text treatment (95% CI [94.0%, 99.8%]), and the bar chart treatment (95% CI [91%, 95%]). This nearly

universal prevalence of targeting variability suggests that it should be understood as a cognitive resource among U.S. adults. This broadens previous findings with college engineering students (del Rosario et al., 2024), and sets expectations for future research with other populations. These results also conflict with the deficit view of behavioral economics—that people are generally flawed in their statistical thinking—and support the resource view—that people have access to sound decision making strategies that are well-adapted to their environment.

The experimental manipulation yielded a significantly lower rate of targeting in the bar chart treatment. This shows that bar charts do more harm than good: While some visualization authors feel that bar charts are a simple way to present data averages, prior work has shown that these lead to a variety of misreadings of visuals (Kerns & Wilmer, 2021; Wang et al., 2025; Wilmer & Kerns, 2022), and the present work shows that bar charts significantly degrade the quality of decision making under variability, even compared to a text-only description.

Conversely, participants were able to make use of the additional information afforded by showing variability: Participants in the sinaplot treatment targeted variability—in a strict sense—at a significantly higher rate. While some visualization creators believe that showing uncertainty information will "overwhelm" (Hullman, 2020), the present work shows that U.S. adults can generally use this additional information to make accurate decisions under variability.

5.2 LIMITATIONS

Treatment assignment was randomized and resulted in no significant difference in demographic factors between treatment groups. Thus, our evidence for the causal attribution of observed effects to the assigned treatment is sound. Other factors pose a greater threat to internal validity.

While treatments were randomized, participants were randomized once into a fixed treatment (text-only, bar chart, or sinaplot). This decision was deliberate to avoid the possibility of learning effects; for instance, it is unclear if a sinaplot disrupts the bar tip limit error. The single treatment randomization, along with task order randomization, mitigates systematic learning effects. However, this also reduces our statistical power and does not allow us to probe any potential learning effects.

Strict targeting is computed based on participant-reported perception of the variability. Since this corrects for a potentially incorrect interpretation of the variability, strict targeting (as computed in this work) is a somewhat generous operationalization. Therefore, results based on strict targeting may overestimate the "real" quality of decisions made under variability. Notably, the rate of strict targeting in the sinaplot condition was not substantially inflated by this generous operationalization, given to the highly accurate participant-reported variability in this condition, and any inflation of the text and bar conditions were insufficient to make up for the substantially higher strict targeting rate in the sina plot condition.

Our sample is demographically representative of the target population: U.S. adults. While this supports inference to the target population, some features of data collection may prevent unbiased inference. Our recruitment vendor provides representative sampling according to age, sex, and ethnicity, but not, for instance, education. Combined with the computerized administration of the survey, this suggests our sample could conceivably be skewed, e.g., towards more technologically inclined individuals. Given the novelty of the instrument, it is unknown what effect this may have on the external validity of results. In the event that targeting positively correlates with technological inclination, our rates of targeting could conceivably be overestimated.

5.3 IMPLICATIONS FOR FUTURE RESEARCH

We presented initial evidence for the reliability and validity of the targeting instrument. This evidence positions the tendency to target (T3) as an individual-level trait, suggesting possible investigations into covariates (e.g., Does T3 vary with demographic factors?) or development (e.g., How does T3 develop with age?).

Our experimental work also suggests that contextual features modulate T3. We experimentally manipulated the presentation of data, but other factors may matter. For instance, a classic finding from behavioral economics is that task framing in terms of gains or loss can significantly affect decisions, a phenomenon known as loss aversion (Schmidt & Zank, 2005). Future work adapting the instrument could investigate other experimental manipulations of scenario wording.

Finally, this work aimed to characterize decision making in everyday life, but was originally inspired by a study of domain-specific decision making (del Rosario, 2024). Having established that targeting is a cognitive resource, future work could investigate whether domain-specific practices inhibit a person's natural inclination to target variability. Such findings are important from an education perspective, as statistics educators seek to train students from other disciplines to carry a statistical mindset into their domain-specific work. Armed with an instrument to measure targeting, instructors can design and evaluate teaching interventions that encourage students to make sound decisions under variability.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under grant no. 2138463.

REFERENCES

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Alacaci, C. (2004). Inferential Statistics: Understanding Expert Knowledge and its Implications for Statistics Education. *Journal of Statistics Education*, *12*(2), 6. https://doi.org/10.1080/10691898.2004.11910737

Arnold, P., & Franklin, C. (2021). What Makes a Good Statistical Question? *Journal of Statistics and Data Science Education*, *29*(1), 122–130. https://doi.org/10.1080/26939169.2021.1877582

Bazerman, M. H., & Moore, D. A. (2012). *Judgment in managerial decision making*. John Wiley & Sons.

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

Bisantz, A. M., Marsiglio, S. S., & Munch, J. (2005). Displaying Uncertainty: Investigating the Effects of Display Format and Specificity. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(4), 777–796. https://doi.org/10.1518/001872005775570916

Cooper, L. L., & Shore, F. S. (2008). Students' Misconceptions in Interpreting Center and Variability of Data Represented via Histograms and Stem-and-Leaf Plots. *Journal of Statistics Education*, *16*(2), 1. https://doi.org/10.1080/10691898.2008.11889559

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. https://doi.org/10.1109/TVCG.2014.2346298

Cumming, G., & Calin-Jageman, R. (2024). *Introduction to the new statistics: Estimation, open science, and beyond* (Second edition). Routledge, Taylor & Francis Group.

del Rosario, Z. (2024). Neglected, Acknowledged, or Targeted: A Conceptual Framing of Variability, Data Analysis, and Domain Consequences. *Journal of Statistics and Data Science Education*. https://doi.org/10.1080/26939169.2024.2308119

del Rosario, Z., Ryu, J., & Saur, E. (2024). *Targeting Consequences of Variability as a Data Science Resource*. IASE Roundtable Conference, Auckland, NZ. https://doi.org/10.52041/iase24.505

DeVellis, R. F. (2017). *Scale development: Theory and applications* (Fourth edition). SAGE.

Dewsbury, B. M. (2020). Deep teaching in a college STEM classroom. *Cultural Studies of Science Education*, *15*(1), 169–191. https://doi.org/10.1007/s11422-018-9891-z

Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018). Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. *Proceedings of the*

2018 CHI Conference on Human Factors in Computing Systems , 1–12. https://doi.org/10.1145/3173574.3173718

Findley, K., & Lyford, A. (2019). INVESTIGATING STUDENTS' REASONING ABOUT SAMPLING DISTRIBUTIONS THROUGH A RESOURCE PERSPECTIVE. *Statistics Education Research Journal*, *18*(1).

Finger, R., & Bisantz, A. M. (2002). Utilizing graphical formats to convey uncertainty in a decision-making task. *Theoretical Issues in Ergonomics Science*, *3*(1), 1–25. https://doi.org/10.1080/14639220110110324

Garfield, J., & Ben-Zvi, D. (2005). A Framework for Teaching and Assessing Reasoning About Variability. *Statistics Education Research Journal*, *4*(1), 92–99. https://doi.org/10.52041/serj.v4i1.527

Gigerenzer, G. (1991). How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases." *European Review of Social Psychology*, *2*(1), 83–115. https://doi.org/10.1080/14792779143000033

Gigerenzer, G. (1996). On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky. *Psychological Review*, *103*(3), 592.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Hicks, S. C., & Irizarry, R. A. (2018). A Guide to Teaching Data Science. *The American Statistician*, *72*(4), 382–391. https://doi.org/10.1080/00031305.2017.1356747

Hullman, J. (2020). Why Authors Don't Visualize Uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 130–139. https://doi.org/10.1109/TVCG.2019.2934287

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. https://doi.org/10.1371/journal.pone.0142444

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

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. https://doi.org/10.1167/jov.21.12.17

Konold, C. (1989). Informal Conceptions of Probability. *Cognition and Instruction*, *6*(1), 59–98.

Liu, L., Boone, A. P., Ruginski, I. T., Padilla, L., Hegarty, M., Creem-Regehr, S. H., Thompson, W. B., Yuksel, C., & House, D. H. (2017). Uncertainty Visualization by Representative Sampling from Prediction Ensembles. *IEEE Transactions on Visualization and Computer Graphics*, *23*(9), 2165–2178. https://doi.org/10.1109/TVCG.2016.2607204

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual Semiotics & Uncertainty Visualization: An Empirical Study. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2496–2505. https://doi.org/10.1109/TVCG.2012.279

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, *19*(4), 601–607. https://doi.org/10.3758/s13423-012-0247-5

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3. ed., [Nachdr.]). McGraw-Hill.

Padilla, L., Kay, M., & Hullman, J. (2021). Uncertainty Visualization. In R. S. Kenett, N. T. Longford, W. W. Piegorsch, & F. Ruggeri (Eds.), *Wiley StatsRef: Statistics Reference Online* (1st ed., pp. 1–18). Wiley. https://doi.org/10.1002/9781118445112.stat08296

Peters, S. A. (2011). Robust Understanding Of Statistical Variation. *Statistics Education Research Journal*, *10*(1), 52–88.

Saks, M. J. (1981). Human Information Processing and Adjudication: Trial by Heuristics. *Law & Soc'y Rev*, *15*(1), 123–160.

Schmidt, U., & Zank, H. (2005). What is Loss Aversion? *Journal of Risk and Uncertainty*, *30*(2),

157–167. https://doi.org/10.1007/s11166-005-6564-6

Shaughnessy, J. M. (1977). Misconceptions of Probability: An Experiment with a Small-Group, Activity-Based, Model Building Approach to Introductory Probability at the College Level. *Educational Studies in Mathematics*, *8*, 295–316.

Sidiropoulos, N., Sohi, S. H., Pedersen, T. L., Porse, B. T., Winther, O., Rapin, N., & Bagger, F. O. (2018). SinaPlot: An Enhanced Chart for Simple and Truthful Representation of Single Observations Over Multiple Classes. *Journal of Computational and Graphical Statistics*, *27*(3), 673–676. https://doi.org/10.1080/10618600.2017.1366914

Spearman, C. (2010). The proof and measurement of association between two things. *International Journal of Epidemiology*, *39*(5), 1137–1150. https://doi.org/10.1093/ije/dyq191

Tufte, E. R. (2001). *The visual display of quantitative information* (Second edition, tenth printing, April 2018). Graphics Press.

Tukey, J. W. (1977). *Exploratory Data Analysis* (Repr.). Addison-Wesley.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

Wang, Y., Kerns, S. H., Brady, T. F., & Wilmer, J. B. (2025). The Paradox of Certainty: When Graphed Ensembles Convey Averages Better than Graphed Averages. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*. https://doi.org/10.31234/osf.io/73ywp_v1

Wild, C. J., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Wild, C., Utts, J., & Horton, N. (2018). What Is Statistics? In D. Ben-Zvi, J. Garfield, & K. Makar (Eds.), *International Handbook of Research in Statistics Education* (1st ed. 2018, pp. 5–36). Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-319-66195-7

Wilke, C. (2019). *Fundamentals of data visualization: A primer on making informative and compelling figures* (First edition). O'Reilly.

Wilmer, J. B., & Kerns, S. H. (2022). *What's really wrong with bar graphs of mean values: Variable and inaccurate communication of evidence on three key dimensions* . https://doi.org/10.31219/osf.io/av5ey

Wilmer, J. B., & Kerns, S. H. (2025). The Uniformity Fallacy: A Second Common, Severe Misinterpretation of Bar Graphs of Averages. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Wood, B. L., Mocko, M., Everson, M., Horton, N. J., & Velleman, P. (2018). Updated Guidelines, Updated Curriculum: The *GAISE College Report* and Introductory Statistics for the Modern Student. *CHANCE*, *31*(2), 53–59. https://doi.org/10.1080/09332480.2018.1467642

Zhang, S., Heck, P. R., Meyer, M. N., Chabris, C. F., Goldstein, D. G., & Hofman, J. M. (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences*, *120*(33), e2302491120. https://doi.org/10.1073/pnas.2302491120

APPENDICES

A. GLOSSARY OF SYMBOLS

For reference, we include a glossary of symbols used in this work in Table A1.

Table A1. Names, mathematical symbols, and their use / definition in this work.

Name	**Symbol / Calculation**	**Use / Description**
Data minimum, maximum; ground truth answer	$l_d, u_d; a_d$	Used to define a ground truth for a task
Perceived minimum, maximum; numeric answer	$l_p, u_p; a_p$	Used to elicit a participant's task perception and decision

Perceived data width	$w_p = \frac{u_p - l_p}{u_d - l_d}$	Used to detect cases of Neglected: Excessive Precision
U-normalization	$u_p = \frac{a_p - l_p}{u_p - l_p}, u_d = \frac{a_d - l_d}{u_d - l_d}$	Used for standardized data visualization and calculating relative error
Relative error	$e_{rel} = \lvert u_p - u_d \rvert$	Used to judge correctness of participant answer, accounting for their task perception
Absolute error	$e_{abs} = \lvert \frac{a_p - a_d}{u_d - l_d} \rvert$	Used to judge correctness of participant answer, irrespective of their task perception
Floored relative error	e_f	Zeros in e_{rel} are replaced with the smallest e_{rel} across all treatments Used to compute l_f
Log-floored relative error	$l_f = log(e_f)$	Used in 3-way ANOVA to study factors that affect participant answer accuracy
Data average	$m_d = (u_d + l_d)/2$	Used to compute the flipped ground truth answer NB Formula is valid due to symmetry of each task's data
Uniform random variable X on a, b	$X \sim U(a, b)$	Used in null model definition
Sign function	$sign(x)$	Returns the sign $\{-1, +1\}$ of x Used to detect Targeted:Side and to judge task understanding
Sign of participant answer difference, sign of ground truth answer difference	$g_p = sign(a_{p,0} - a_{p,1})$, $g_d = sign(a_{d,0} - a_{d,1})$	Used to judge participant's task understanding (via $g_p = g_d$)
Side of participant answer, side of ground truth answer	$s_p = sign(a_p - m_d)$, $s_d = sign(a_d - m_d)$	Used to detect Targeted:Side (via $s_p = s_d$)