2023 Annual Conference & Exposition
Baltimore Convention Center, MD | June 25 - 28, 2023
The Harbor of Engineering
Education for 130 Years
ASEE
Paper ID #36726

Board 198: A Mixed-Methods Investigation of Engineers Targeting the Consequences of Variability

Prof. Zachary Riggins del Rosario, Olin College

Zachary del Rosario is an Assistant Professor of Engineering and Applied Statistics at Olin College. His goal is to help scientists and engineers reason under uncertainty. Zach uses a toolkit from data science and uncertainty quantification to address a diverse set of problems, including reliable aircraft design and AI-assisted discovery of novel materials.

A Mixed-Methods Investigation of Engineers Targeting the Consequences of Variability

Abstract

Variability is an unavoidable reality. Physical phenomena such as loading conditions, material properties, and human behavior all exhibit variability. Engineers must deal with this variability when designing solutions. Unfortunately, an extensive body of human subjects research suggests that people—including engineers—consistently fail to understand variability. This deficit view of working with data is focused on statistical inference; identifying stable patterns in data. However, engineering concerns are not identical to statistical concerns! In this paper, we report results from two studies: Qualitative Study 1 of practicing engineers (n=24) identified the behavior of *targeting the consequences of variability*. Mixed-methods Study 2 of engineering students (n=22) developed a survey to measure the prevalence of this targeting behavior. These two studies suggest that targeting is a broadly-accessible behavior, *so long as the decision-maker recognizes the consequences of variability*. In this work we describe the process of targeting variability, highlight factors that affect an engineer's proclivity to target (or not), and discuss implications for engineering education.

Background & Motivation

Variability is ubiquitous, but often misunderstood. Prior research on human behavior has demonstrated that people often use crude heuristics to make decisions in the presence of variability, such as the *representativeness heuristic* [1] and the *outcome approach* [2]. This mistreatment of variability extends to engineering, with an example from interface design providing an illuminating example.

In the 1940s the US Air Force had serious issues transitioning their fleet to jet fighters. At the height of this calamity 17 pilots crashed in a single day [3]. While the Air Force initially blamed individual pilots and instructors, the researcher Gilbert Daniels investigated the aircrafts' human interfaces. The standard at the time was to design for "the average man," with non-adjustable controls assuming fixed human dimensions. Daniels studied the measurements of 4063 pilots, and found that precisely zero were average [4]. The solution to this design error was dramatic: The Air Force effectively "banned" the average by requiring its aircraft suppliers to design for the variability observed among its pilots [3].

While variability across humans is now acknowledged in aerospace engineering, other sources of variability are still mistreated. The standard practice in aerospace design is to quantify certain material properties in terms of sample averages [5], a practice that has been in-use since at least the 1960's [6]. This practice similarly ignores sources of variability, and exposes aircraft passengers to elevated levels of risk.

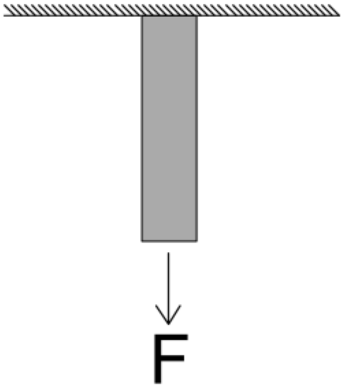

Figure 1. A rod with uniform cross-section, loaded in uniaxial tension. This image is relevant to the example problem in this section, and was used to illustrate the rod design scenario considered in Q7 for Study 1.

To illustrate the problem, we consider a simple structural design task: Consider a rod (Fig. 1) with uniform cross-section of area A, subject to a simple uniaxial tensile load F. Assuming deterministic input quantities, this structure will survive loading if the following inequality is met

$$F \,/\, A \;<\; \sigma,$$

where σ is the strength of the material. In practice, real materials exhibit (real) variability in their properties; for instance, Table 4 in Appendix A1 reports tensile yield strength values for a cast steel alloy. These values exhibit variability, which demands a statistical treatment of the data in order to make sound decisions.

In the context of design, an engineer would use the inequality above to select the cross-sectional area A based on the loading conditions F and chosen material σ. If the engineer were to use the median strength, this would effectively design for a 50% failure rate: an unacceptable level of safety for any application where failure may result in human injury. Using the mean strength would result in a similarly deficient design; this would apply the same faulty reasoning as the "average man" fallacy from interface design. In aerospace engineering, using either the mean or median for the material strength would actually be *unlawful*; federal regulations require that more conservative values be used for design [5].

While the treatment of material strength is tightly regulated in Aerospace Engineering, the treatment of other properties is far more lax. Common practice in Aerospace Engineering when treating variability in material elasticity is to simply take the average value. In cases where failure depends on elasticity (such as with buckling) [7], the use of the average results in elevated risk for structures and any users of those systems [5].

This research was motivated initially by these observed trends in Aerospace Engineering. However, the goals of this project are to understand how engineers across disciplines react to variability. The following sections describe the frameworks used to frame the research, results from two studies under this project umbrella, and implications from across the studies. The goal of this paper is to describe the behavior of *targeting the consequences of variability*, and to begin development of an instrument to measure population-level rates at which engineers target variability.

Frameworks

We use several frameworks to focus this work. In this section we report the theoretical framework underpinning our work, as well as the knowledge-base (conceptual framework) layered on this theoretical foundation [8].

Theoretical framework: Knowledge-in-Pieces (KiP)

To frame our work, we adopt the Knowledge-in-Pieces (KiP) theoretical framework, originally developed by diSessa and colleagues [9], [10]. This theory was originally developed to serve as an "epistemology of physics"—to explain how students transition from an intuitive to formal sense of physics [9]. KiP has also been used in conceptual change research in the learning sciences [10]. KiP articulates how a person's intuitive sense of the physical world interacts with learning formal reasoning methods; therefore, it is an appropriate framework for our studies.

KiP posits that knowledge of a reasoning agent is not monolithic, but rather composed of relatively small elements called *phenomenological primitives* (p-prims) [9]. These knowledge elements act by being recognized by the agent, called *cuing*. P-prims cue on perceived configurations in the real world, have different relevant stimuli, and have different sensitivities to activation called *cuing priority*.

A useful example of a p-prim from physics is "Ohm's p-prim"—that more effort begets more result, and more resistance begets less result [9]. This p-prim is named for the fundamental electronic circuit concept of Ohm's law, but serves as a much broader reasoning tool: a larger force begets larger acceleration, a larger mass begets less acceleration; a larger pressure begets larger flow, a larger pipe resistance (e.g., smaller diameter) begets less flow.

KiP predicts that different contextual features will cue different p-prims, and that certain p-prims will be more sensitive to activation. This implies that presenting slight variations on the same problem may elicit different behaviors from reasoning agents, and that these variations will have some stability across persons. Based on the background and our motivation, we have designed our studies to vary relevant problem features, but hold constant the exposure to variability. Our aim at this stage of the work is not to identify p-prims *per se*, but rather to carefully observe and understand the variation in participants' behavior with task-contextual features.

Conceptual Framework: Cause and Source of Variability

Our focus in this work is on engineers' reactions to variability. Put simply, variability arises when multiple observations on "the same" quantity result in different values. In reality, variability is a complex and multifaceted concept: The reason for the existence of statistics as a discipline is the problematic nature of variability [11]. To sufficiently articulate the aspects of variability we seek to study, we introduce the *cause and source axes* of variability.

The dichotomy of assignable and chance cause was introduced by Shewhart [12] to frame the process of reducing variability in manufacturing. In short, an assignable cause of variability is practical to understand and eliminate, while a chance cause is considered impractical to control. A process is thought to be due to chance cause alone if it exhibits sufficient statistical regularity; this is usually assessed using a *control chart*, shown schematically in Figure 2 [12], [13]. If no assignable causes are detected, then the process is provisionally said to be under statistical control and can be pragmatically modeled as a random process.

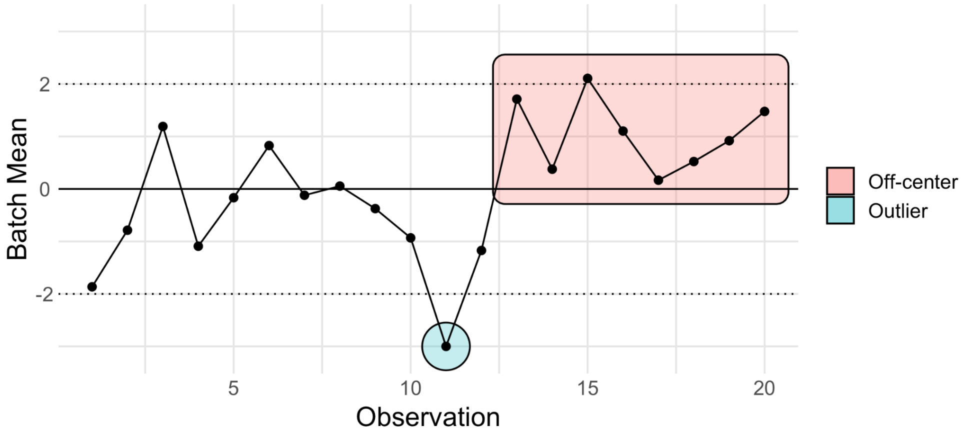

Figure 2. An example control chart. The highlighted observations are different detections of assignable causes.

While the cause dichotomy helps determine what to treat as random, it does not describe the *consequences* of observed variability. We use the dichotomy of real and erroneous sources to study the consequences of variability[1]. To understand this dichotomy, we must first draw a distinction between the quantity of interest (QOI) we seek to study and a measurement of that QOI. Real variability affects the QOI, while erroneous variability affects the measurement *only*. Figure 3 schematically depicts sources of real and erroneous variability.

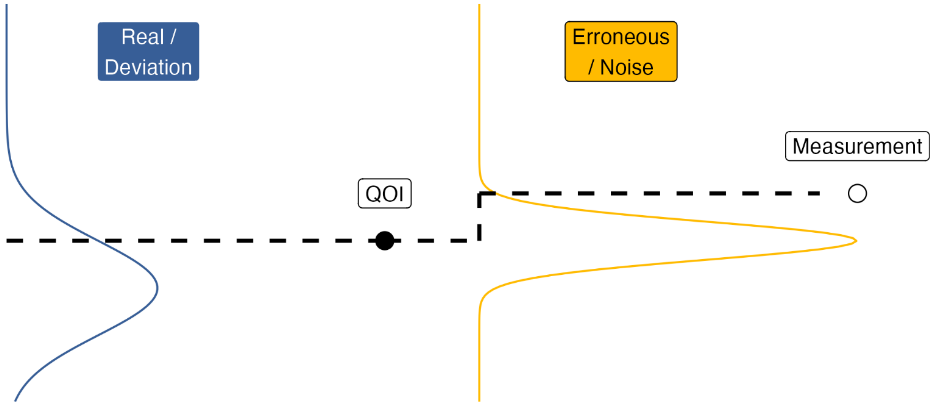

Figure 3. A schematic depiction of the real vs erroneous dichotomy.

The real vs erroneous source dichotomy elucidates the consequences of variability: Real variability can affect the outcomes of a system, such as the failure probability of a mechanical structure. Erroneous variability cannot change outcomes on its own; however, erroneous variability can affect human decisions. A smaller sample will lead to larger erroneous variability, which in turn increases the probability of a poor decision.

As a brief aside, note that it is common to classify uncertainties (hence, variability) as either *aleatory* or *epistemic* [14]. We reject this framing in our work, as the term "aleatory" has connotations of both "inherent randomness" (what we call *chance*) and "natural variability" (what we call *real*) [15], [16]. As we will see below, separating cause and source enables one to understand and eliminate supposedly "irreducible" sources of variability.

The dichotomies of cause and source are orthogonal, and can be used as axes to organize different sources of variability. Figure 4 depicts the *cause-source variability quadrants*, while Figure 5 provides examples of each quadrant in the context of manufacturing and testing metallic components.

[1] Wild and Pfannkuch [11] introduced the distinction of real and induced variability, which we have developed further. We use the term "erroneous" rather than "induced" to avoid overloading phenomena engineers and physicists would consider to be real, such as induced drag and induced current.

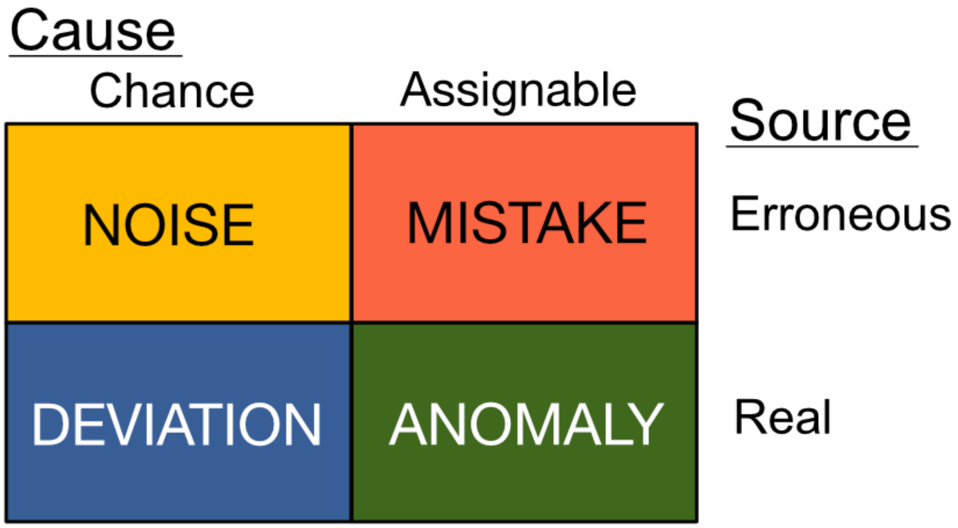

Figure 4. The cause/source axes, with names for each quadrant.

The manufacture and testing of metallic components has opportunities for all four quadrants of variability to manifest (Fig. 4). Figure 5 illustrates four such examples: In this case, (1) an anomaly has occurred in mixing the alloy precursors, such that the chemical composition does not match the desired levels. This will certainly affect the real material properties, but can be controlled and prevented in future manufacturing runs. However, (2) deviations occur as cracks form in the material and the microstructure settles into an intrinsic realization of disorder. This will also affect the material properties, but it is impractical to eliminate all cracks in all manufactured components.

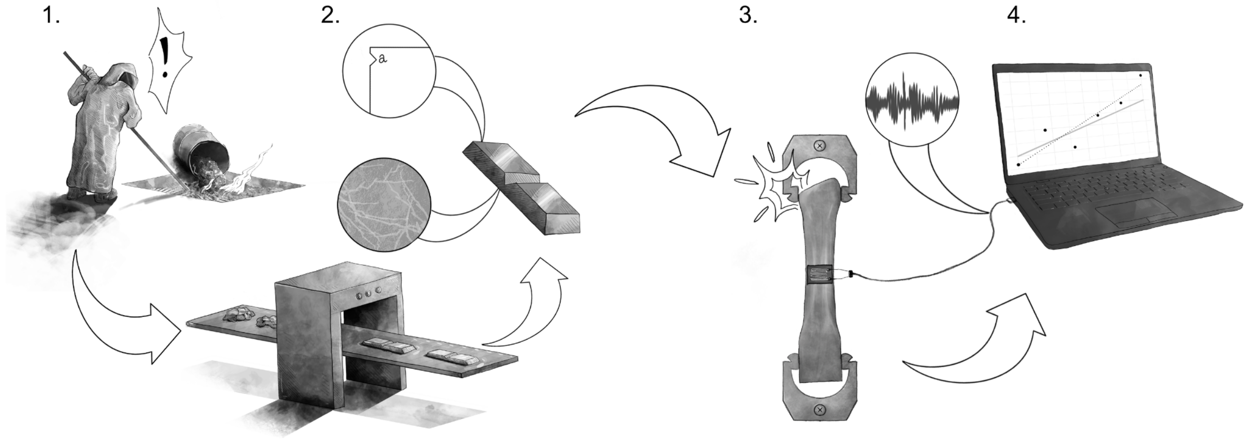

Figure 5. Examples of an anomaly (1), deviation (2), mistake (3) and noise (4) in the context of manufacturing and testing metallic components. Image drawn by Alana Huitric.

After manufacturing, coupons are extracted from the components in order to determine material properties. During testing (3) a mistake occurs: the grips on the mechanical testing jig are not sufficiently tight, and slipping occurs. This leads to an incorrect level of force applied to the specimen, and an incorrect measurement of the material's elasticity and strength. Such an occurrence can be corrected once identified. However, (4) noise enters into the transduction of strain into an electrical signal due to ambient electromagnetic radiation. While proper shielding can reduce this noise, it is not practical to eliminate entirely.

The cause/source axes enable us to articulate the focus of our research. While the concepts of assignable and chance cause have seen broad adoption in manufacturing, to the best of our knowledge, the concepts of real and erroneous sources have been under-studied in literature. Across the studies in this project, we have designed tasks to emphasize the possibility of both real and erroneous variability (e.g., Appendix A1. Fig. 10). The source dichotomy allows us to analyze consequences of variability; the concept of *reification* (described next) is a behavioral pattern that is hypothesized to contribute to mishandling of consequential variability.

Conceptual Framework: Reification

Steven Jay Gould [17] defines reification as "the mental conversion of a person or abstract concept into a thing." Originally introduced in Marxist theory by Georg Lukács [18], reification describes not just "thing-ification," but also a "forgetting" of other analyses or interpretations. In his book *Full House: The Spread of Excellence from Plato to Darwin*, Gould [17] treats reification as the central mechanism by which people misunderstand variability, focusing on narratives from evolutionary biology and American baseball. Gould writes,

> The common error lies in failing to recognize that apparent trends can be generated as by-products, or side consequences, of expansions and contractions in the amount of variation within a system, and not by anything directly moving anywhere.

For instance, Gould notes that the question of "Why does no one hit 0.400 anymore in baseball?" is commonly answered as an *overall* decay of skill among baseball athletes. Gould presents the case that reification of an extreme value (the batting average of the best players) has led to this incorrect interpretation of a population phenomenon: a decay in the variability over time.

Reification has been reported in other disciplines such as environmental science, where averages are commonly used to summarize variation [19]. Hoffman and Thiessen [20] describe a case where expert modelers sought to predict Cesium 137 concentrations in humans using data from the Cherynobyl disaster; they document that even domain experts neglected interindividual variability in their models, effectively reifying the population average. We hypothesize that reification has contributed to the aircraft allowables issue described above.

Reification of the average in particular is dangerous for sources of real variability. If a decision-maker focuses on the average, then any source of variability is unaccounted. Since real variability has the potential to affect outcomes, any consequences of variability are ignored when the average is used as the sole decision input.

While prior authors treat reification as the core of a "grand narrative" (e.g., past incarnations of Marxist theory [18]), combined with KiP we interpret reification as just one of many possible behaviors emerging from p-prim cuing. While reification is an important and potentially dangerous behavior, we do not necessarily expect an individual to commit reification in all scenarios. Rather, our studies have been designed to investigate what contextual features lead to reification among engineers.

Studies and Results

In this section we report results from two studies. Study 1 is a qualitative study where we first identified the behavior of *targeting the consequences of variability*. Study 2 is a mixed-methods study where we prototyped a survey instrument to assess the rate at which individuals would target. We conducted follow-up interviews to understand cases where participants did not target.

Study 1: Qualitative Study of Practicing Engineers

This study was conducted under a protocol approved by Brandeis University's IRB, number #22134R-E. Full details on this study are reported elsewhere [21]; what follows is a short description to support understanding Study 2.

One of our goals in Study 1 was to understand what analysis choices practicing engineers would make in response to variability. Potential participants were recruited via the author's professional network. Potential participants were incentivized to participate through a 6-week professional development course offered by the author. Participants were then selected to have an engineering background, at least 2 years of professional experience, and to balance representation across race, gender, and subfield. Compared with degrees awarded in 2020 [22], our sample is relatively diverse in gender (sample Female 29% vs 2020 degree share 24%), race (sample white 33% vs 2020 degree share 56%), and nationality (including participants residing in Canada, Turkey, and the Philippines). Aligned with the goals of the larger project, participants were drawn from Aerospace, Civil, and Mechanical engineering disciplines. Participant demographics are summarized in Table 1.

Table 1. Summary of participant demographics.

Experience	2 years: 3	3 years: 2	4 years: 8	5+ years: 11
Race	Asian: 10	Black: 2	White: 8	Other: 4
Subfield	Aerospace: 5	Civil: 9	Mechanical: 9	Other: 1
Gender	Male: 17	Female: 7		

Participants were interviewed by a researcher: either the author or a research assistant on the project. Interviews were conducted and recorded on Zoom, lasted 45 to 90 minutes, and followed a common semi-structured protocol.

The interview protocol was structured to expose participants to variability in material properties, in order to study their interpretation and analysis choices in response to variability. Prompts were designed to not be prescriptive; we do not ask for specific numbers, but rather ask participants to describe their process under a given scenario. We designed the interview task prompts to provide contextual features that were hypothesized to cue different behaviors. For instance, Q1.7 from late in the protocol presented participants with a dataset of strength values (Appendix A1. Tab. 4), an image of the structural element (Appendix A1. Fig. 1), and asked the following prompt,

> **Researcher (Q1.7):** Imagine you were going to design a rod to withstand a tensile load, using the cast alloy described by this dataset. How would you use this dataset to help design the rod? Please just describe your process; you don't need to do any calculations.

This task is abstracted from any specific application but carries a clear design context: there is a specific structural element that will be loaded in tension, and the task is to avoid failure of the structure. This is in contrast with an earlier question (Q1.2) that presented the same dataset but asked a different prompt,

> **Researcher (Q1.2):** These are measured tensile yield strength values for a cast steel alloy. How would you use the data to describe this alloy?

In contrast, this task has far fewer contextual features: no structural element, no application, and no specific design goal. The full interview protocol varied the material property under consideration, the goal of the task presented (description or design), and additional engineering knowledge artifacts (e.g., diagrams and equations). For this report, we focus on a subset of the protocol and results to introduce the concept of *targeting variability*.

From Study 1, we observed that some participants would, on certain occasions, make analysis choices that would prevent adverse consequences of variability from occurring. We dubbed this behavior *targeting*, and developed a closed coding scheme to identify this behavior from interview transcripts. We report the development and validation of this coding scheme in Reference [21]; below, we provide examples and non-examples of targeting.

Reification of the mean. Use of the mean was widespread, as 100% of participants used the mean at least once in their interview. In response to *all* interview tasks, Participant 1.1 took the mean of the data. For instance, for Q1.7 he responded

> **Participant 1.1 (Q1.7):** Same thing as before. [chuckles] Just assumptions. Just go with assumptions first and after that, I'll take the mean of these 10 samples. After that, based on the values that I have, I'll look into the tables or any related websites to see what the value represents or interprets.

For Participant 1.1, the mean is strongly-reified; he presents no other possible analyses of the data. However, other participants responded to the context-specific factors. For instance, Participant 1.23 initially used the average alone to describe the strength data (Q1.2),

> **Participant 1.23 (Q1.2):** … through experience I know that's a very high yield strength. This material is really strong, and it takes a lot of energy to break it. Looking at the data itself, again, it seems mostly uniform average-- There's one outlier from what I can see sample four, but if you exclude the outlier, the average is about 156 ksi.

Following the protocol, the interviewer asked why they picked that specific approach,

> **Researcher:** Then, why did you use that quantity of choosing the middle, middle like median or mean?
>
> **Participant 1.23 (Q1.2):** Mean.
>
> **Researcher:** Why did you, yes?
>
> **Participant 1.23 (Q1.2):** Why? I just did it. At least in engineering, I don't know if I ever use median or mode.

Use of the mean is reflexive for Participant 1.23 ("I just did it.") and to the exclusion of other approaches ("I don't know if I ever use median or mode."). Note that the participant does not initially mention any of these alternative analyses; these are prompted by the researcher. This illustrates the "forgetting" nature of reification: the tendency to shut out other potential analyses of the data. However, the same participant used a very different analysis in response to Q7,

> **Participant 1.23 (Q1.7):** Okay. Pretty much the same thing in the previous example [Q1.6, design of a column]. I would take the minimum recorded measurements of the

> tensile yield strength, and I would use that to inform the design of the rod. Not the average, not the maximum. I would want to design for worst case scenario.

In this case, the design context of the problem cues a different knowledge element, which results in a different approach to the data analysis. Here, Participant 1.23's approach *targets* the potential consequences of variability: the possibility of failure. Furthermore, the participant has overcome reification of any specific value; he mentioned several alternative analyses and dismisses them as less appropriate for the task at hand.

Targeting variability. Across participants and tasks, we observed a plethora of analyses that target variability. A common approach was to use the minimum observed value,

> **Participant 1.6 (Q1.2):** To be conservative, I would almost use always the min value from this set.

Participants who used one-sided extreme values did so in response to consequences of the variability—these are instances of targeting variability. Cases where participants used upper and lower extreme values usually did so to quantify variability; these were not in response to any consequences, and hence were not targeting.

Other participants used quantiles in order to target variability in a way that controls probabilities,

> **Researcher:** You also mentioned quartiles. Could you tell me a little bit more about what those quartiles are and what they would tell you about the alloy?
>
> **Participant 1.2 (Q1.1):** If we had a larger like, I mean we were manufacturing lots of like samples of aluminum, I don't know, bars or whatever you'd want to, like aim for some low incident rate of breaking or something.

While some participants sought to control probabilities (rates) directly, not all instances of targeting follow this approach. Other participants combined multiple summaries to construct a conservative value. For instance, Participant 1.13 did so starting from the first interview task due to common practice,

> **Participant 1.13 (Q1.1):** Mean minus one standard deviation, right? This is a standard procedure actually for the approaches. When you look at the codes, or let's say actually in Turkey it's a law, actually, you have to obey it.

This legal requirement prevents reification; since engineers are required to compute both mean and standard deviation, their required practice involves computing multiple statistics of the same dataset. This may promote targeting variability; Participant 1.13 targeted in 6/7 interview tasks.

Since Participant 1.4 trained as a Civil Engineer in Canada, she is accustomed to using the Canadian practice of computing the 5th percentile strength [23],

> **Participant 1.4 (Q1.2):** We don't want to overestimate strength, we want to have those safety factors. We might be interested, say, for example, the fifth percentile strength, in which case we need to close the whole distribution curve.

From Study 1, we found that 21/24 participants targeted variability at least once in their interview: a large fraction. However, the nature of our participant recruitment likely biases this statistic. To produce a reasonable population-level estimate of targeting rate among practicing engineers, we must recruit a representative sample and deploy a scalable survey instrument. While we have future plans for such sampling, Study 2 was designed to begin development of this instrument.

Study 2: Mixed-methods Study of Student Engineers

This study was conducted under a protocol approved by Brandeis University's IRB, number #23053R-E. Study 1 allowed us to identify and describe the behavior of targeting; the goal of Study 2 was to develop a survey instrument that could help measure targeting of variability in non-engineering contexts and with a larger sample. Study 1 was designed to investigate engineering reasoning under variability in the context of material properties; however, these tasks require specialized knowledge. To measure targeting outside the context of material properties, we aimed to develop tasks that use more common knowledge.

To test this survey, we recruited undergraduate student participants from a small, private engineering college in the US Northeast. Potential participants were contacted via internal email listservs and incentivized via free food in a public location on-campus. Approximately 10% of the student body initiated the survey, though only ~60% of the initiators completed the survey. At the close of survey data collection we obtained n=22 valid responses. Among these responders, we selected m=7 participants to conduct a follow-up interview, selecting participants based on cases of *non*-targeting in their survey response.

The survey itself was designed to be maximally intelligible to participants. All questions required numeric input, which limits user answers to a single value, but facilitated scalable coding of participant responses. Since we required numeric input, we presented quantile dotplots (rather than raw data) to visualize variability; these have been shown to improve graphical perception of uncertainty [24]. Two classes of tasks were designed: Dotplot Comprehension questions tested participant understanding of the quantile dotplots, while Variability Targeting asked participants to make choices using dotplots in various scenarios. The Variability Targeting tasks were designed to have clear directionality, responses can then be coded as targeting if their provided numerical value satisfies an inequality. Our analysis of the survey data and follow-up interviews were primarily aimed at assessing whether participants interpreted the survey questions as intended.

Survey tasks. This section briefly overviews the survey used in Study 2, a more complete description is given in Appendix A2. Figure 6 illustrates an example quantile dotplot that visualizes a normal distribution; this was also the first image presented to survey participants.

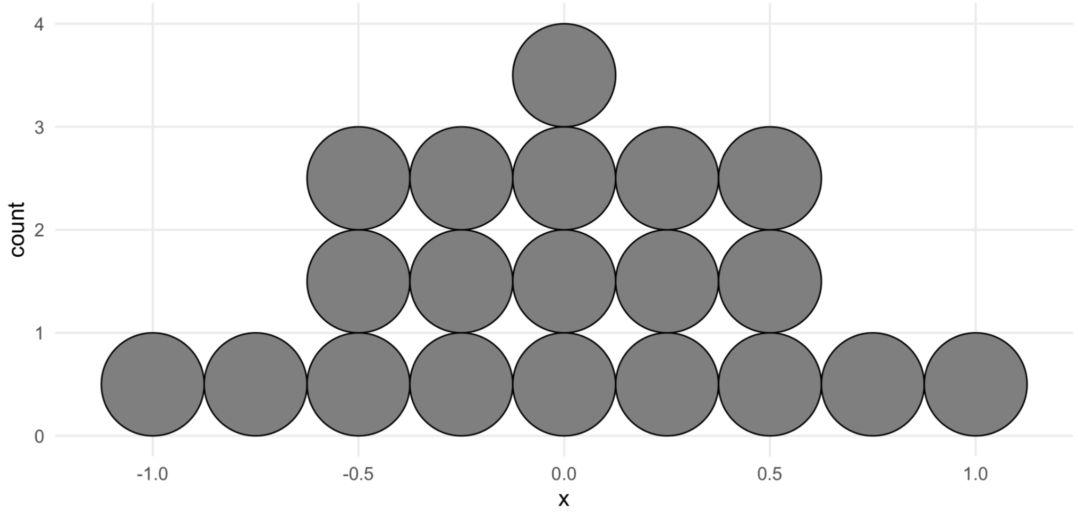

Figure 6. Image used for Dotplot Comprehension questions.

Participants were given the following explanation for Figure 6,

> This plot represents a set of observed values of x; each dot represents a single observation. Vertical stacks of dots represent multiple observations that have nearly the same value of x.

Participants were then asked four Dotplot Comprehension questions; precise wordings are given in Appendix A2. In short, these four questions tested participants' ability to identify the mean (0.0), max value (1.0), 25th percentile (-0.5), and 75th percentile (+0.5) according to the dotplot. These questions were designed to assess whether participants could adequately interpret the information displayed in the dotplots; an important prerequisite for the Variability Targeting tasks later in the survey.

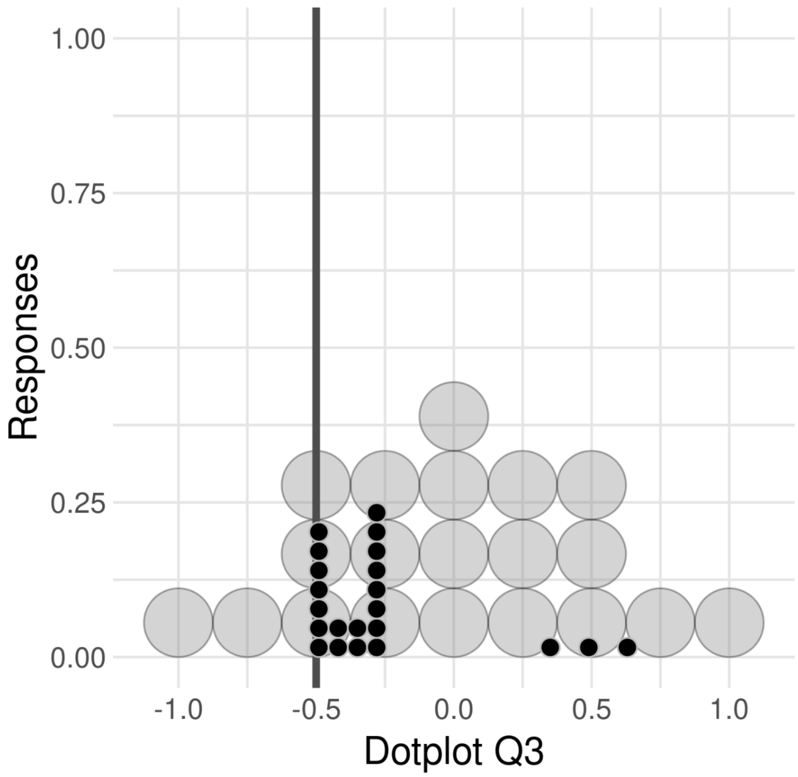

Figure 7. Set of participant responses to Dotplot Comprehension question 3 (25th percentile). Correct answer shown as dark vertical line. Original dotplot (Fig. 6) is provided as a transparent underlay for comparison. The majority of participants could identify the correct answer, following loose coding.

All participants (100%) answered Dotplot Comprehension questions 1 (mean) and 2 (max) unambiguously correctly. This is important, as it shows that all participants can compute both the mean and extreme values from a symmetric dotplot. However, inspection of the percentile-related responses revealed variation in participants' interpretation of the dotplot. Figure 7 shows all responses to Dotplot Comprehension question 3 (25th percentile). Note that a plurality of participants provide a numerical value at or near the correct answer (-0.5). However, some participants seem to have utilized the boundary of the dots (around -0.375) or the next-larger dot stack (-0.25) to answer this question. Therefore, we code participant responses to the Dotplot Comprehension quantile questions using both a *strict* and *loose* definition. Strict coding deems a response correct if it is 0.05 units within the correct value, while loose coding deems a response correct within 0.25 units.

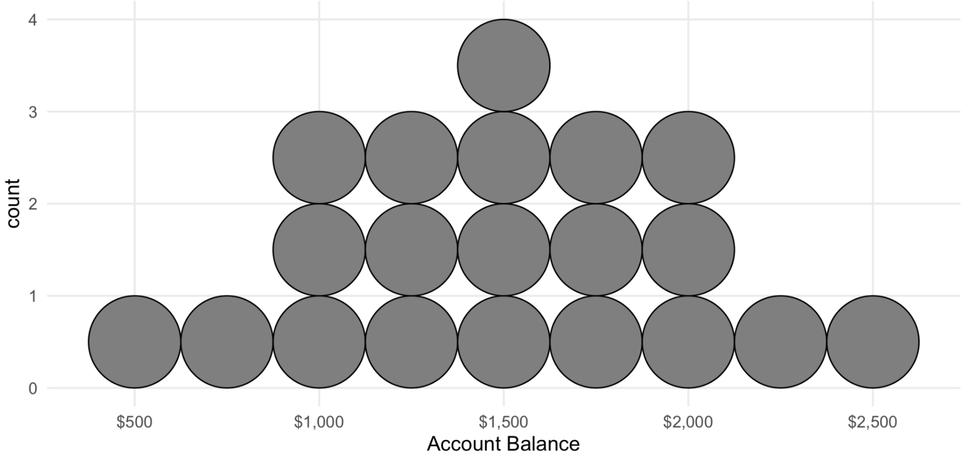

Figure 8. Image used for Overdraft question.

Variability targeting questions were designed to have clear context that would cue thinking about the consequences of variability, but with no obviously correct single value. For instance, the "Overdraft" question presented participants with Figure 8, and was accompanied by the following prompt (emphasis as in survey),

> This plot represents the balance of your bank account, observed at different months in a year. If you write a check for more than the current balance of your account, you will get an overdraft penalty.
>
> What is the **largest** check you can **safely** write without first checking the balance of your account?
>
> (Enter your answer in units of dollars; the same units shown above.)

While this question asks for the largest check, the variation in Figure 8 suggests that one should write a smaller check to avoid overdraft penalties. We coded responses to the Overdraft question as Targeting if the provided value was less than $1,250; a value based on the loose coding approach developed for the Dotplot Comprehension questions. We designed six Variability Targeting tasks that varied in units (dollars and time), directionality (high or low), and context. Table 2 summarizes these questions and their associated coding schemes.

Table 2. Table of Variability Targeting questions from Study 2 survey. *Means approximate due to lognormal distribution.

Question	**Distribution Mean**	**Coded Targeting if…**
Overdraft	$1,500	< $1,250
Groceries	$150	> $175
Auction	$1,500	> $1,750
Shuttle Departure	~0 min*	< -4 min (arrive early)
Driving Commute	~30 min*	> 45 min
Cup Stacking (excluded)	~15 sec*	< 10 sec

Study 2 results. In contrast with Study 1, in Study 2 we found that 100% of survey participants targeted variability at least once. To summarize results from the survey, we compute the *individual response fraction* for each individual, separated by question classes. For Dotplot Comprehension questions we compute the response fraction as the individual's number of correct responses over the total number of comprehension questions (four) according to *both* the strict and loose codings. For Variability Targeting questions, we compute the response fraction as the individual's number of targeting responses over the total number of questions. Due to persistent misunderstanding of the Cup Stacking question (described below), we exclude this question and compute the response fraction over the five remaining Variability Targeting questions. The distribution of response fractions is reported in Figure 9.

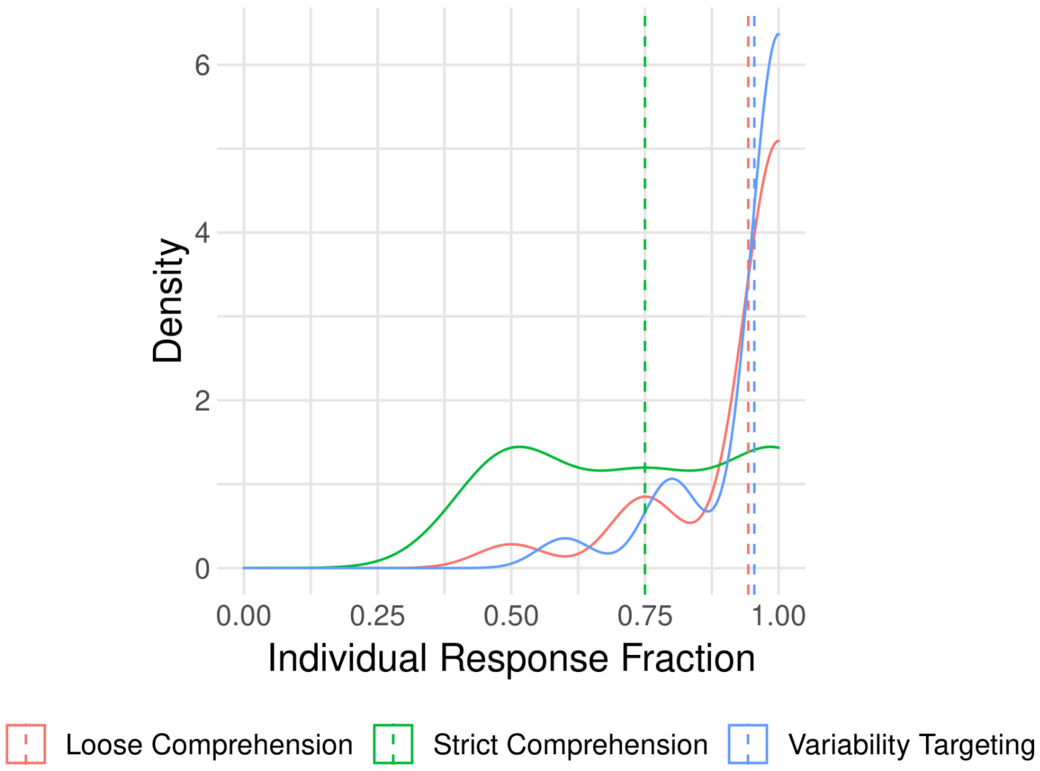

Figure 9. Distribution of individual response fractions for all n=22 individuals (solid curves) and sample means (dashed lines), according to correctness (loose or strict) for the Dotplot Comprehension questions and according to targeting for the Variability Targeting questions.

Figure 9 illustrates that the highest response fraction among participants tended to be Targeting, followed by loose Dotplot Comprehension, with strict Dotplot Comprehension the lowest. Differences in mean individual response fraction were significant at the 1% level between strict comprehension and targeting (paired t-test, t = 4.3, df = 21, p = 0.0003), but insignificant between loose targeting and targeting (paired t-test, t = 0.52, df = 21, p = 0.60).

In addition to the individual response fraction, we also report a *question response fraction* for each question, with coding based on correctness (Dotplot Comprehension questions) or targeting (Variability Targeting questions). These fractions are computed as the number of positive codings over the total number of participants (n=22). Table 3 reports only those tasks with a question response fraction less than 1.0. The Variability Targeting questions with a question response fraction < 1.0 are those we asked participants about in the m=7 follow-up interviews.

Table 3. Table of question response fractions across Study 2 tasks. Dotplot questions are coded for comprehension, while other tasks are coded for targeting variability. Unreported tasks have a question response fraction of 1.0.

Task	**Question Response Fraction**
Dotplot 3 (strict)	0.41
Cup Stacking	0.50
Dotplot 4 (strict)	0.59
Dotplot 3 (loose)	0.86
Groceries	0.91
Dotplot 4 (loose)	0.91
Overdraft	0.91
Driving Commute	0.95

Follow-up interviews with participants. While the quantitative results showed encouraging levels of targeting among participants, we were interested in cases where participants did *not* target. We revisited the Groceries question with Participant 2.2, who initially responded with a central value ($150). He attributed this approach to "rushing,"

> **Researcher:** Participant 2.2, could you take a moment to reread this prompt?
>
> **Participant 2.2:** Yes, [laughs] I'm laughing because it's like the same situation we just talked about, but I had a completely different answer.
>
> **Researcher:** That's actually exactly what I want to talk about. Could you try to tell me about what was your reasoning when you answered this question, and how's your reasoning different now, say within this interview?
>
> **Participant 2.2:** Yes. Well, I think analytically what I was thinking was that I wanted to take the average grocery bill and say that that is how much that the person might typically spend. I think also the context of when the situation was happening, me rushing to do this questionnaire to get pizza versus sitting down for an interview and just thinking about it.

This is aligned with behavioral economics research on judgment heuristics [25], particularly the distinction between a slow deliberative approach and a fast, intuitive approach. Here, use of the average value was intuitive for Participant 2.2. This suggests a higher cuing priority for p-prims that emphasize the mean, which would contribute to reification of the average.

Certain questions have a greater sense of consequence and can cue other interpretations. The same participant targeted variability in the Auction question, bidding $2,000—well over the mean. When asked about his different approach, he explained,

> **Participant 2.2:** I feel like part of it might've been the wordings of the questions. I feel like the "what if" is written into the question, or maybe me in the past felt that "what if" was written to the question of the auction, having the wording of almost guarantee had that scenario of like, there might be these extremes, but you still want this painting. I feel like me in the past was like reading the grocery question and was just thinking how much cash do you bring to the grocery, how much do you typically spend at the grocery.

Particular prompt features such as "almost guarantee" carry a connotation of "extremes" for this participant; this is beneficial to targeting as extremes cannot exist in a reified understanding of the world. Rather, extremes imply variability. Note that this text was bolded in the survey prompt (Appendix A2), which may explain why Participant 2.2 noticed this particular feature of the prompt.

While most of the Variability Targeting questions had a response fraction over 0.9, the Cup Stacking question had a strikingly low fraction of 0.5. Through interviews, we discovered this is in part due to a common misinterpretation of the prompt,

> **Researcher:** Okay. One reason I'm asking about this question is, this is the first time I've actually tried using this question in a survey and so I'm still working on some of the kinks. One of the things that-- actually I've seen multiple people miss in here is that this actually represents the stacking times for the other competitors.
>
> **Participant 2.7:** Ah, okay.
>
> **Researcher:** What this would mean is that your friend will actually need to beat these times over here in order to win the competition.
>
> **Participant 2.7:** Yes, I missed that entirely. I'm sorry.
>
> **Researcher:** No. Actually, you're not the only person. I actually did another interview today who missed exactly the same thing, which indicates to me that I need to change the way that I phrase this to make it clear.

While the distribution was labeled as the "stacking times for the other competitors," this interpretation as the times *of the individual competing* occurred multiple times. However, and more importantly, we found that a normative targeting approach for this question may not exist,

> **Participant 2.7:** How does telling her a number encourage her? Is that a number she's somehow planning to aim for?
>
> **Researcher:** You can think of it that way.
>
> **Participant 2.7:** Like, "Pace yourself. Go for it." 10 seconds. Sorry, this is just a hard question to figure out the boundaries around, how do you encourage someone with a number. Like if it's a track meet or something, it's similar.
>
> **Researcher:** Like that.
>
> **Participant 2.7:** I don't know. The danger is psyching the person out right before a competition. Being like, you need to be whatever at whatever point in time. They look up for half second, they see that they're not meeting the time, they give up. That would be bad. I think you're supposed to practice at 85% of your potential so that the competitions you always score a little bit better. You don't burn yourself out but maybe that's only for track. I don't know about this situation to do a recommendation. I think if I was just being like, "Wow, you can go do it. This is a reasonable time." I think I'd be 13 seconds.

Participant 2.7's analysis of the Cup Stacking question forced us to reflect on the task's value as a means to measure normative behavior: While the other questions have fairly unambiguous consequences since they affect the decision-maker individual alone, the Cup Stacking question asks one to consider how a different person will react. Since it is not reasonable to expect that all persons will react in the same way to an identical response, our inequality-based coding scheme is not appropriate for this question.

Discussion

This project seeks to understand how engineers reason under variability: cases where repeated observations of "the same" scenario exhibit different numerical outcomes. We adopted a theoretical framework based on the knowledge-in-pieces framework [9], [10], particularly the concept of cuing priority for knowledge elements. We also constructed a conceptual framework based on reification of numerical values [17], [18] and a consequence-focused understanding of variability as real or erroneous [11]. We reported results from two empirical studies: Study 1 was a qualitative study of practicing engineers working with materials data exhibiting variability. From the first study, we identified and articulated a behavior of *targeting* the consequences of variability through data analysis choices. Study 2 was a mixed-methods study of engineering students. From the second study we developed a survey instrument to measure the targeting behavior in general-knowledge settings.

The results of Study 2 suggest that targeting is a broadly-available behavior, as 100% of n=22 survey participants exhibited at least one response that targeted the consequences of variability.

However, this result does not assess the cuing reliability of *consequence recognition*; the tasks of Study 2 were designed to clearly-articulate the consequences of variability. In practice, reasoning agents must recognize these consequences on their own. We saw cases of participants who could target variability, but would not target for certain questions. For instance, Participant 2.2 answered the Grocery question in haste, did not recognize the consequence of trying to pay with too little cash on hand, and answered using the average value.

Use of the mean occurs more frequently than other data analysis approaches; this suggests a higher cuing priority of knowledge elements that utilize the mean. All n=24 participants in Study 1 took the mean at least once, and some participants used the mean as their only data analysis strategy. In contrast, targeting took the form of many different approaches: the minimum observed value, the mean adjusted by the standard deviation, and specific percentiles of the underlying population. Furthermore, use of the mean was often done in a reified fashion: an approach that ignored the possibility of other analyses. While reification (in data analysis) is a "forgetting" of other approaches, targeting represents a diversity of reasonable ways to address the potential consequences of variability.

Limitations and Future work. While Study 2 strongly suggests that targeting is a broadly-accessible approach to data analysis, due to the nature of our sample, we certainly cannot conclude that this behavior is *universally* accessible. Future work should study a larger and more diverse sample to establish the accessibility of the targeting behavior.

More importantly, the tasks of Study 2 were designed to be maximally intelligible: the data were presented using optimized practices in uncertainty visualization [24], the prompts were engineered to clearly articulate consequences, and the scenarios were based on general-knowledge. Engineering practice occurs in cases where the data are confusing, the consequences of variability are obscured, and scenarios require highly-specialized knowledge. The questions from Study 2 serve as a useful baseline; our future work will vary these problem features to see how they affect cuing of the targeting behavior.

Implications. Statistics as a discipline is unique in its focus on directly studying uncertainty [26]. Many engineers take courses in statistics, which provides them with valuable training in data analysis. However, statistics as a discipline is focused on identifying stable patterns: separating signal from noise [27]. While this orientation towards data analysis is useful to engineers, it does not address all engineering concerns. Engineers are responsible for dealing with the *consequences* of variability.

The behavior of *targeting* is an important concept for framing data analysis education for engineers: In order to make decisions that will ensure the safe operation of systems, engineers must be able to target the consequences of variability for those systems. The results of Study 2 suggest that the behavior of targeting is broadly-accessible, *so long as students recognize the consequences of variability*. Rather than presenting numerical values solely as averages (reifying these values), instructors can directly discuss variability and its potential consequences in their engineering courses. Useful examples abound in current industry practices, such as the use of basis values in Aerospace Engineering [5], or the use of 5th percentile strength in Civil Engineering [23]. Small changes to engineering pedagogy may help new engineers recognize the consequences of variability more readily, and lead to safer engineered designs.

Acknowledgements

Many thanks to Diane Lam, Andy Ziccarelli, Alexandra Coso-Strong, and Jonathan Stolk, who served on an advisory board for this project. Study 1 interviews were conducted by AJ Evans, Ellie Ramos, Emma Fox, KD Vo, and Maeve Stites. Thanks to Abe Levitan and Mike Baiocchi, who engaged in early conversations about the cause/source axes and provided very useful feedback. Thanks also to Jonathan Adler, who provided invaluable expertise on the qualitative methods used in these studies. This material is based upon work supported by the National Science Foundation under Grant No. 2138463.

References

[1] D. Kahneman and A. Tversky, “Subjective probability: A judgment of representativeness,” *Cognit. Psychol.*, vol. 3, no. 3, pp. 430–454, 1972.

[2] C. Konold, “Informal conceptions of probability,” *Cogn. Instr.*, vol. 6, no. 1, pp. 59–98, 1989.

[3] T. Rose, *The End of Average: How We Succeed in a World That Values Sameness*, First Edition. New York: HarperOne, 2015.

[4] G. Daniels, “The ‘Average Man’?,” AIR FORCE AEROSPACE MEDICAL RESEARCH LAB, WRIGHT-PATTERSON AFB OH, AD010203, 1952.

[5] Z. del Rosario, R. W. Fenrich, and G. Iaccarino, “When Are Allowables Conservative?,” *AIAA J.*, vol. 59, no. 5, pp. 1760–1772, May 2021, doi: 10.2514/1.J059578.

[6] “Federal Register Vol. 29, No.250, December 24, 1964 - Content Details - FR-1964-12-24.” https://www.govinfo.gov/app/details/FR-1964-12-24 (accessed May 11, 2021).

[7] D. J. Peery, *Aircraft structures*. Mineola, N.Y: Dover Publications, 2011.

[8] A. J. Magana, “The role of frameworks in engineering education research,” *J. Eng. Educ.*, vol. 111, no. 1, pp. 9–13, Jan. 2022, doi: 10.1002/jee.20443.

[9] A. A. diSessa, “Toward an Epistemology of Physics,” *Cogn. Instr.*, vol. 10, no. 2–3, pp. 105–225, 1993.

[10] A. diSessa, “A History of Conceptual Change Research: Threads and Fault Lines,” in *The Cambridge handbook of: The learning sciences*, Cambridge University Press, 2006, pp. 265–281.

[11] C. J. Wild and M. Pfannkuch, “Statistical Thinking in Empirical Enquiry,” *Int. Stat. Rev.*, vol. 67, no. 3, pp. 223–248, Dec. 1999, doi: 10.1111/j.1751-5823.1999.tb00442.x.

[12] W. A. Shewhart, *Economic control of quality of manufactured product*. D. Van Nostrand Company, Inc., 1931.

[13] E. Deming, “Quality, productivity, and competitive position,” 1991.

[14] A. D. Kiureghian and O. Ditlevsen, “Aleatory or epistemic? Does it matter?,” *Struct. Saf.*, vol. 31, no. 2, pp. 105–112, Mar. 2009, doi: 10.1016/j.strusafe.2008.06.020.

[15] J. Mullins, Y. Ling, S. Mahadevan, L. Sun, and A. Strachan, “Separation of aleatory and epistemic uncertainty in probabilistic model validation,” *Reliab. Eng. Syst. Saf.*, vol. 147, pp. 49–59, Mar. 2016, doi: 10.1016/j.ress.2015.10.003.

[16] R. C. Smith, *Uncertainty quantification: theory, implementation, and applications*. Philadelphia: Society for Industrial and Applied Mathematics, 2013.

[17] S. J. Gould, *Full house: the spread of excellence from Plato to Darwin*, 1st ed. New York: Harmony Books, 1996.

[18] A. Honneth, J. Butler, R. Geuss, J. Lear, and M. Jay, *Reification: a new look at an old idea*. Oxford ; New York: Oxford University Press, 2008.

[19] O. H. Pilkey and L. Pilkey-Jarvis, *Useless arithmetic: why environmental scientists can't predict the future*. New York: Columbia University Press, 2007.

[20] F. O. Hoffman and K. M. Thiessen, "The use of Chernobyl data to test model predictions for interindividual variability of 137Cs concentrations in humans," p. 6, 1996.

[21] E. Fox, A. Evans, K. Vo, E. Ramos, M. Stites, and Z. del Rosario, "Context-Engaged Engineering Data Analysis: A Grounded Theory of Engineers Targeting the Consequences of Variability," PsyArXiv, preprint, Oct. 2022. doi: 10.31234/osf.io/q6a3j.

[22] "IPEDS : Integrated Postsecondary Education Data System," National Center for Education Statistics, 2020.

[23] B. Madsen, "Strength Values for Wood and Limit States Design," *Can. J. Civ. Eng.*, vol. 2, no. 3, pp. 270–279, Sep. 1975, doi: 10.1139/l75-025.

[24] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson, "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*, San Jose California USA, May 2016, pp. 5092–5103. doi: 10.1145/2858036.2858558.

[25] D. Kahneman, *Thinking, fast and slow*, 1st pbk. ed. New York: Farrar, Straus and Giroux, 2013.

[26] R. P. Abelson, *Statistics as principled argument*. Hillsdale, N.J: L. Erlbaum Associates, 1995.

[27] C. Konold and A. Pollatsek, "Data Analysis as the Search for Signals in Noisy Processes," *J. Res. Math. Educ.*, vol. 33, no. 4, p. 259, Jul. 2002, doi: 10.2307/749741.

[28] P. E. Ruff, "AN OVERVIEW OF THE MIL-HDBK-5 PROGRAM," Battelle's Columbus Laboratories, AFWAL-TR-84-1423, 1984.

Appendices

A1. Study 1 Protocol Details

Before the analyzed tasks, participants were presented with Figure 10. This was to clarify the context of data that was presented in the interview—presented values arise from multiple independent specimens, rather than repeated measurements on a single specimen. This was to ensure the possibility of real variability in the data, without directly naming the concept.

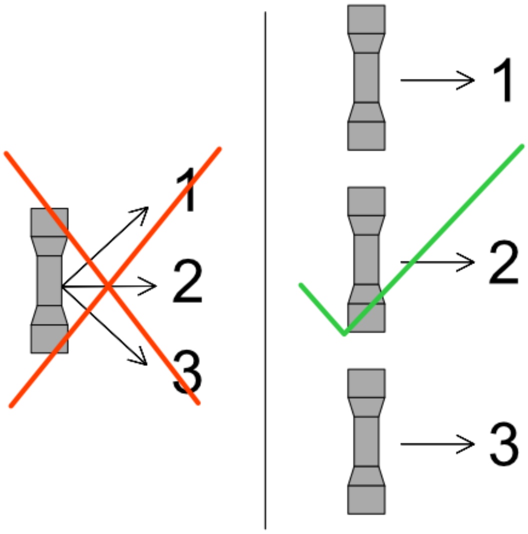

Figure 10. Image used to describe the presented data: independent specimens, rather than repeated measurements.

Early in the protocol, participants were presented with a table of strength values (Tab. 4) and simply asked to describe the data.

> **(Q1.2)** These are measured tensile yield strength values for a cast steel alloy. How would you use the data to describe this alloy?

The interview protocol also contains the following optional follow-up questions. The (parentheticals) describe conditions when the interviewer should ask the follow-up.

- (If the participant does not mention anything codable according to the Analysis Style rubric) "How would you describe the data numerically?"
- (If the participant gives a number, but does not describe how they got it) "So that I understand, how did you get that number?"
- (If the participant mentions a measure of spread) "How would that quantity relate to the alloy?"
- (For any differences with the previous question) "Why did you use [that different approach]?"
- (If the participant says they need more data) "Fair enough; suppose you had a large enough dataset. How would you analyze that dataset?"

Later, participants were asked to use a dataset to help design a rod by selecting its cross-sectional area. The design context of the interview task was a geometrically-simple member subject to uniaxial tension, pictured in Figure 1. The following prompt was accompanied by the same dataset as Q1.2 (Tab. 4).

(Q1.7) Imagine you were going to design a rod to withstand a tensile load, using the cast alloy described by this dataset. How would you use this dataset to help design the rod? Please just describe your process; you don't need to do any calculations."

Table 4. Dataset presented in interviews, values are the tensile yield strengths of a cast steel [28].

Steel Strength	
Sample	**Tensile Yield Strength (ksi)**
1	157.0
2	159.6
3	155.6
4	165.8
5	157.4
6	158.4
7	157.6
8	156.4
9	157.7
10	155.7

A2. Study 2 Instrument Details

This section provides further details on the survey instrument from Study 2.

Dotplot Comprehension questions

The Dotplot Comprehension questions and their correct answers are given below. These were paired with Figure 6.

- Based on the dotplot above, what is the mean of x? (Ans: 0.0)
- Based on the dotplot above, what is the largest value of x observed? (Ans: 1.0
- Based on the dotplot above, what is the value where 1/4 of the dataset (no more, no less) is less than (to the left of) that value? (Ans: -0.5)
- Based on the dotplot above, what is the value where 3/4 of the dataset (no more, no less) is less than (to the left of) that value? (Ans: +0.5)

Variability Targeting questions

We designed six Variability Targeting questions, listed in Table 2 (along with their targeting coding rules): Overdraft, Groceries, Auction, Shuttle Departure, Driving Commute, and Cup Stacking. The Overdraft question was reported above (Fig. 8), the remaining 5 questions are reported here.

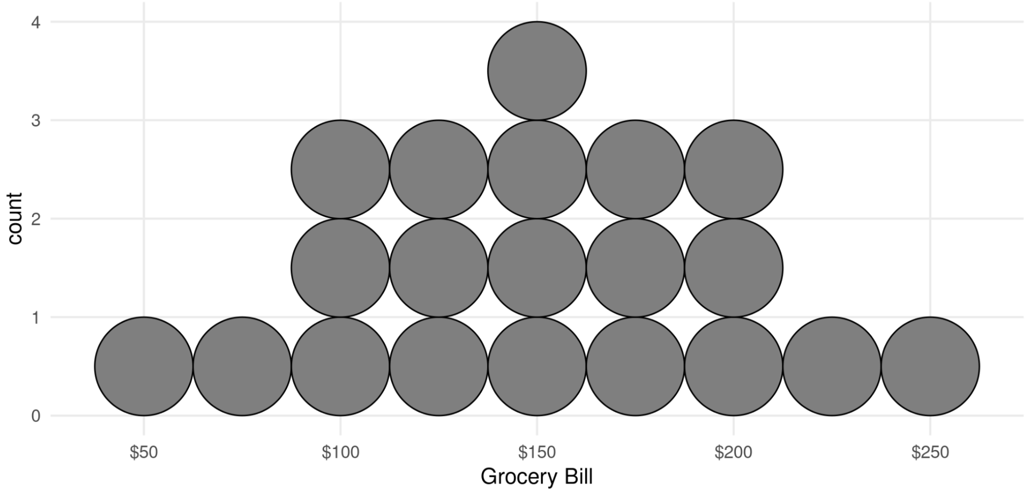

Figure 11. Quantile dotplot accompanying the Grocery question.

The Grocery question displayed Figure 11, and used the following prompt:

> This plot represents your weekly grocery bill, observed at multiple different weeks.
>
> Supposing you need to pay in cash, how much cash should you bring to the grocer?
>
> (Enter your answer in units of dollars; the same units shown above.)

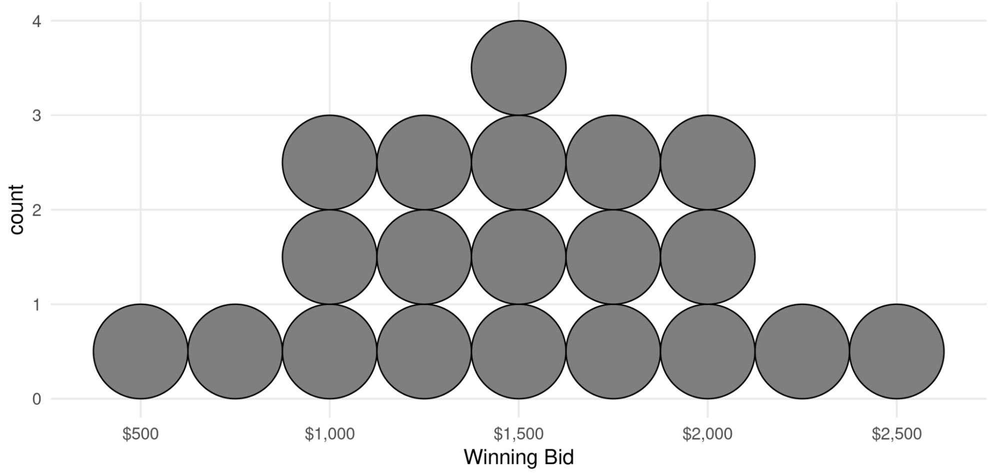

Figure 12. Quantile dotplot accompanying the Auction question.

The Auction question displayed Figure 12 and used the following prompt (emphasis as in survey):

> You are participating in an auction for paintings from local artists. This plot represents the winning bids for previous paintings. In this auction, you write down just one number and that is your final bid.
>
> Imagine that a painting you really want has come up in this auction. What dollar amount would you bid to **almost guarantee** that you will win the painting?
>
> (Enter your answer in units of dollars; the same units shown above.)

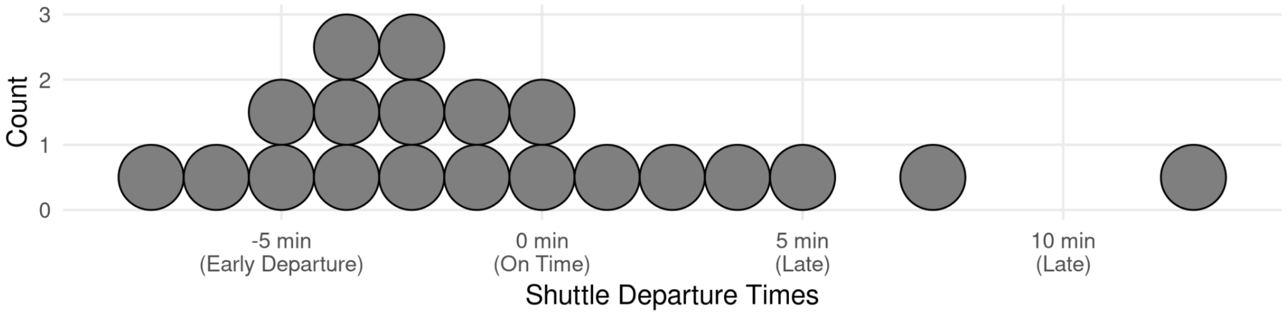

Figure 13. Figure accompanying the Shuttle Departure question.

The Shuttle Departure question displayed Figure 13 and used the following prompt:

> This plot represents the departure times for an airport shuttle, relative to published departure times.
>
> Assume you need to catch this shuttle, or else you will miss a connecting flight. What time should you arrive for the shuttle, relative to the published departure time?
>
> (Please enter your answer in terms of minutes after the published departure time, the same units shown above. Use “-5” for 5 minutes early, “+4” for 4 minutes late, etc.)

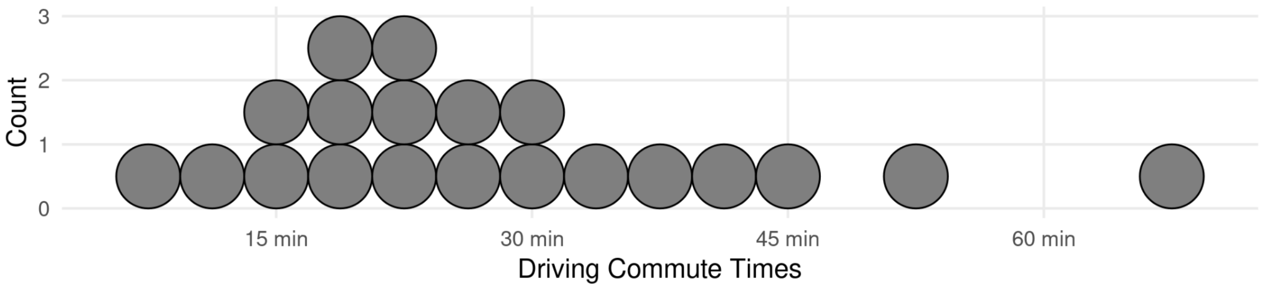

Figure 14. Figure accompanying the Driving Commute question.

The Driving Commute question displayed Figure 14 and used the following prompt:

> This plot represents your driving commute time to work over the past few weeks.
>
> Assume you need to arrive at work for a very high-stakes meeting. Your continued employment depends on this meeting going well, so you must be there on-time. How early should you leave to commute to work?
>
> (Please enter your answer in terms of minutes; the same units shown above.)

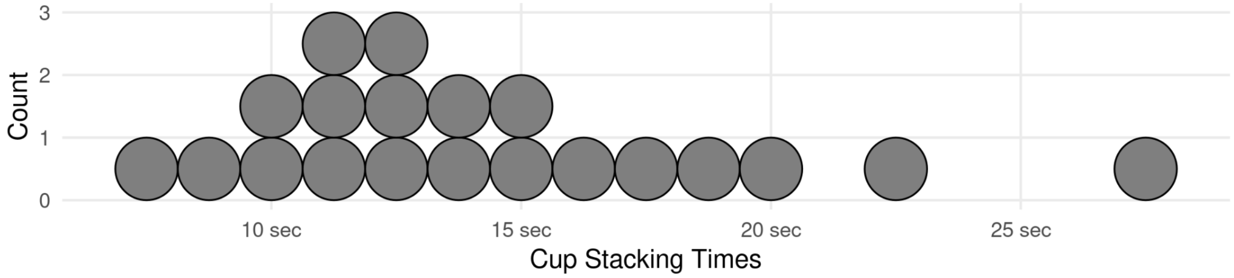

Figure 15. Figure accompanying the Cup Stacking question.

The Cup Stacking question displayed Figure 15 and used the following prompt:

> You are at a competitive cup stacking competition to watch a friend compete. This plot represents the stacking times for the other competitors.
>
> Imagine you are trying to encourage your friend before the competition. She is favored to win the competition, but is a little nervous. You're planning to remind her that she's gotten some really good stacking times before, but you can't remember them offhand, and all you have to go off of is the data above.
>
> What cup stacking time should you tell your friend to encourage her?
>
> (Please enter your answer in terms of seconds; the same units shown above.)