TARGETING CONSEQUENCES OF VARIABILITY AS A COGNITIVE RESOURCE IN DATA LITERACY

Zachary del Rosario, Jin Ryu, and Erika Saur
Olin College
Wellesley College
zdelrosario@olin.edu

Variability is core to statistical thinking but is often neglected in other disciplines such as engineering. Our previous study developed the concept of targeting: responding to the consequences of variability. That study found that practicing engineers targeted variability at a low rate (~51% of tasks). It was unclear, however, whether lack of targeting is a prevalent misconception, or if targeting is a cognitive resource that some have trouble deploying. The present study investigated the rate at which college engineering students targeted variability in everyday scenarios and piloted a survey instrument for the targeting behavior. We found that students in our sample targeted at a very high rate (~90% of tasks), suggesting targeting should be considered a cognitive resource. Practically, statistics and data science educators can use targeting as a bridge between statistical thinking and making decisions under variability in other domains.

INTRODUCTION

Variability is core to statistical thinking and is a unique focus of statistics education (Wild & Pfannkuch, 1999). As statistics educators work to define interdisciplinary approaches of data science, it is important to contrast with other disciplines. For instance, variability is drastically neglected in engineering: Only 2 of 5466 articles identified in a systematic review of engineering mathematics articles discussed "uncertainty" or "error" (Hadley & Oyetunji, 2022). Only 11% of textbooks in a scoping review of engineering course reserve lists considered variability (Vo et al., 2023).

This neglect of variability has several negative consequences: The neglect of variability in the analog signals underlying digital circuits can lead to designs with fatal flaws (Ginosar, 2003). Neglect of variability in physical dimensions (height, arm length, etc.) across humans led to uncontrollable aircraft in the 1950s (Rose, 2015). In the present day, neglect of variability between males and females in crash testing contributes to 47% higher odds of female passenger injury in the U.S. (Bose et al., 2011; GAO, 2023).

In a prior study, we conducted a qualitative investigation of engineers designing under variability. In that study we formulated a conceptual framing of how engineers respond to variability—the *NAT Taxonomy* (del Rosario, 2024). Core to NAT is the idea of *targeting variability*, making intentional choices to mitigate the potential negative consequences of variability: for instance, designing for a range of heights rather than for the average person. This work represents a synthesis of the focus on variability in statistics education compared with the focus on design in engineering.

The original study presented engineers with 7 different tasks in a structured interview. The practicing engineers targeted variability in only ~51% of all tasks (del Rosario, 2024). Given the potential for neglected variability to lead to injury or death, this is surprisingly low. There are various factors, however, that could influence engineers to target variability (or not):

- The presentation of data as a table (rather than as a graph) could make thinking about variability more difficult.
- Reasoning about variability in engineering tasks (often with complex equations) may be more difficult than reasoning in everyday situations.
- Engineering practice may expect—or even demand (del Rosario et al., 2021)—certain analytic choices.

Therefore, this study was initiated to determine whether a higher rate of targeting would result from modifying the aforementioned factors. We expected to see a higher rate of targeting in the present study, the question being how much higher?

Furthermore, the concept of targeting has potential applications in data science—particularly data literacy—in undergraduate education. Ideally, students of all disciplines should learn not only to make inferentially sound conclusions, but also to make decisions that target the consequences of variability. The concept of targeting can serve as a bridge to help students engage with data in an

In: Kaplan, J. & Luebke, K. (Ed.) (2024). *Connecting data and people for inclusive statistics and data science education. Proceedings of the Roundtable conference of the International Association for Statistics Education (IASE), July 2024, Auckland, New Zealand.*

interdisciplinary approach—to simultaneously consider statistical variability and the relevant consequences from another discipline. We initiated the present study to extend the concept of targeting to everyday situations and begin development of a survey-based instrument to measure the behavior of targeting. This is part of a longer-term effort to study the behavior of targeting more broadly.

Our research questions were:

RQ 1. To what extent do college-age students target variability in everyday situations?

RQ 2. Can a survey-based instrument recording numerical responses and self-reported negative consequences ("negcon") measure the behavior of targeting?

Based on the results (presented below), we suggest that targeting of variability is a *cognitive resource*—a beneficial pattern of reasoning that our participants have ready access to when reasoning with data (diSessa, 2014). This resource can be built upon in data science literacy efforts.

BACKGROUND, STUDY DESIGN, AND QUALITY PLAN

Background

The Neglected, Acknowledged, Targeted (NAT) Taxonomy was developed to describe the data analysis practices of engineers in response to variability (del Rosario, 2024). The taxonomy rungs are:

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

This taxonomy was used to design the survey in this study: The tasks are designed to study whether participants target variability in several everyday scenarios.

Study Design

Here we provide a brief overview of the study design. The Quality Plan in the following section provides more details on the interdisciplinary, quality-promoting methods embedded in this research design. This work was conducted following a human subjects protection protocol approved by the Brandeis University IRB, protocol number #23053R-E.

This study was a mixed-methods design (Creswell, 2014) combining a Qualtrics survey response with think-aloud interviewing simultaneously (Reinhart et al., 2022).[1] The survey consisted of nine tasks: one "warmup" to introduce the quantile dot plot visuals (Kay et al., 2016) used throughout, and eight tasks grounded in everyday scenarios. The task scenarios were selected based on situations typical U.S. young adults are likely to encounter in their lives, such as buying groceries or commuting via car or train. To describe each scenario, we provided the participants with a dot plot visual depicting variability, a description of the scenario, and their goal for the task.

A quantile dot plot (e.g., Fig .1) is a discrete representation of a distribution (Kay et al., 2016).

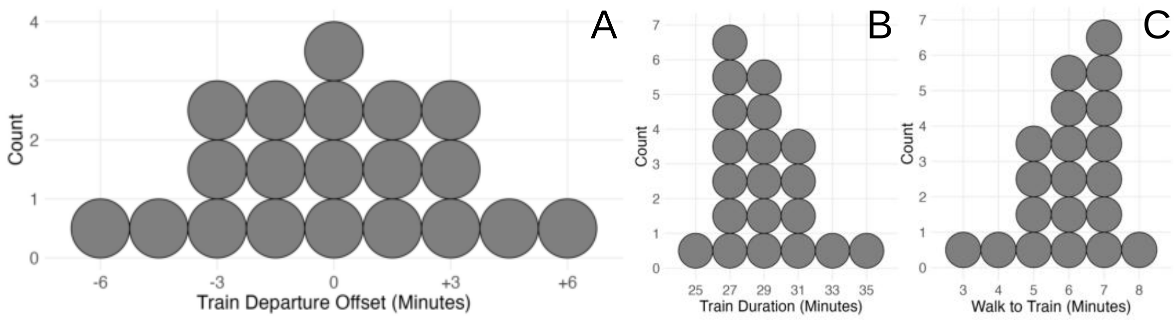

Figure 1. Example quantile dot plot for the Train task. Each panel shows a component of variation: (A) offset from stated departure time, (B) train ride duration, and (C) walking time to train. While the train question had three sources of variation, all other tasks had just one source (hence, one dot plot).

[1] The full survey is freely available on Figshare with DOI: 10.6084/m9.figshare.25551687

In this visual, equispaced probability points have been mapped to quantiles via the inverse CDF and depicted as stacked dots. Quantile dot plots have been shown to make uncertainty information more readily available for reasoning, compared to continuous distributions. Although quantile dot plots depict distributions, for simplicity, we described each dot plot as a sample within each task.

For each of these scenarios, the tasks were broken into two pages each so two questions could be displayed separately. Both contained a copy of the quantile dot plot as well as a description of the scenario. The first page contained the prompt for a numeric entry based on variability and task goal. The second page contained a negative consequence elicitation form ("negcon" for short). As seen in Figure 2, the values shown on the dot plot are divided into five ranges for negcon elicitation.

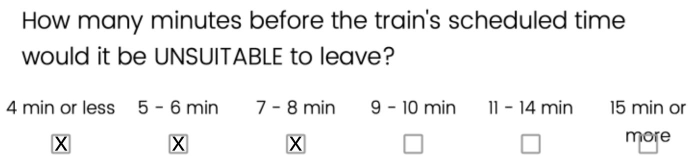

Figure 2. This example "negcon" question is affected by the train departure time (Fig. 1A) and walking time (Fig. 1C). An example response is shown; here, a corresponding numeric response greater than 8min would be a Targeted response according to the survey data.

Figure 2 illustrates the negcon entry form—the primary innovation of the survey. Targeting is a coordination of one's choices with one's perception of consequence. If a participant's numeric response lies in their self-reported zone of negative consequences, this suggests they are not targeting. Conversely, if a respondent's numeric response does not lie in their negcon zone, it suggests they targeted. The participants were selected for the think-aloud interview if their initial survey answers suggested non-targeting. This strategy of purposeful sampling served our goals (Charmaz, 2014), as our pairing of survey with think-aloud was meant to assess the alignment of the survey with the existing closed coding scheme—a ruleset for identifying a passage as N, A, or T (Saldaña, 2013).[2]

We gathered a sample of survey responses from 21 students at Olin College, a small liberal arts flavored engineering college in Needham, MA, USA. From those responders, we invited a subset (n = 8) with suspected neglected cases to retake the survey while conducting a think-aloud interview over Zoom to understand what process led them to a non-targeted answer. This design was intended to gather response process validity evidence (Reinhart et al., 2022), assessing RQ 2.

Table 1. Quality plan following the Q3 framework (Walther et al., 2013).

Facet	Data Collection	Data Analysis
Theoretical Validation	Designed several tasks to elicit thinking about variability across everyday life.	Investigated cases of coder-coder and survey-coding disagreement to understand coherence and complexity (Walther et al., 2013).
Procedural Validation	Designed tasks iteratively, first with internal team feedback, then with participant feedback.	Two analysts independently coded and iteratively debriefed to produce the NAT codes.
Communicative Validation	Used think-aloud techniques to elicit participant thoughts (Beatty & Willis, 2007; Reinhart et al., 2022).	Peer debriefing of closed codes during rounds of interrater reliability (Walther et al., 2013).
Pragmatic Validation	Used theory previously validated in similar context (del Rosario, 2024).	Tested the theory in a new context (students in life, rather than engineers designing).
Process Reliability	Recorded interviews on Zoom. Recorded intermediate & independent NAT codes.	Maintained a discussion log to track analytic developments.

[2] The closed coding scheme used in this study is reported in the open access article del Rosario (2024)

Quality Plan

To promote research quality actively, we employed the quality in qualitative research (Q3) framework to organize our data collection and analysis (Walther et al., 2013). Q3 is organized around "facets" of quality synthesized from the methodological literature; for instance, communicative validation attends to convergence of meaning between interviewer and participant (data collection) and adhering to the meaning conventions of a knowledge community (data analysis). While this framework was developed in the engineering education community, it draws on ideas from industrial statistics: namely, the process-oriented approach of Deming (2018), which focuses on quality throughout the entire research process, not just at the end to assess results. In line with total quality management, we selected quality-promoting methods to use throughout our study, in addition to quality checks near the end of the project (see Table 1).

RESULTS

Across all think-aloud participants and codable episodes (one per task), the coders agreed on "targeted" for ~90% of all tasks, and at least one analyst coded as "targeted" ~94% of tasks. In the original study (del Rosario, 2024) only ~51% of tasks were targeted. This provides a clear answer to RQ 1, but considerably limits our ability to answer RQ 2. Table 2 and Figure 3 summarize the coding and survey results from the study: For brevity we report results from only the n=8 think-aloud participants.

Table 2. Closed codes for the Train question components (think-aloud data only). N, A, T letters follow the NAT Taxonomy, with letter pairs corresponding to the two coders. Black highlighting denotes a code disagreement. Note: no survey responses for this question were in the participants' negcon zone.

	Participant							
Task	A	B	C	D	E	F	G	H
Train (leave offset)	TT	TT	TT	TT	NN	TT	TT	TT
Train (walking time)	TT	TT	TT	TT	TT	TT	TT	TT
Train (transit time)	TT	AA	NN	TT	TT	NN	TT	TT
Train (overall)	TT	TT	TT	TT	AT	TT	TT	TT

Table 2 reports closed coding of the Train question: each response would be defined as neglected (N), acknowledged (A), or targeted (T) following the existing coding scheme (del Rosario, 2024). The Train question (Figs. 1, 2) included multiple sources of variability: how early or late the train leaves compared to its scheduled time (leave offset, Fig. 1A), how long it takes the participant to walk to the train (walking time, Fig. 1B), and how long the duration of the train ride is (transit time, Fig. 1C). Closed coding of the think-alouds reveals that half of participants failed to target at least one source of variability. However, the survey is unable to resolve this detail: No participants gave a numerical choice in their self-reported negcon zone. Overall, we found that the survey did not reliably identify cases of non-targeting for the Train task.

Agreement between survey and coding

Studying Figure 3, nearly all episodes coded as targeted (think-aloud data) agree with a survey numerical response that does not land in the negcon zone. This supports the operating principle of the survey design: Targeting variability tends to be captured by the survey. Combined with the low detection rate of non-targeted cases noted with the Train question results, the instrument (as presently designed) will tend to overestimate a rate of targeting.

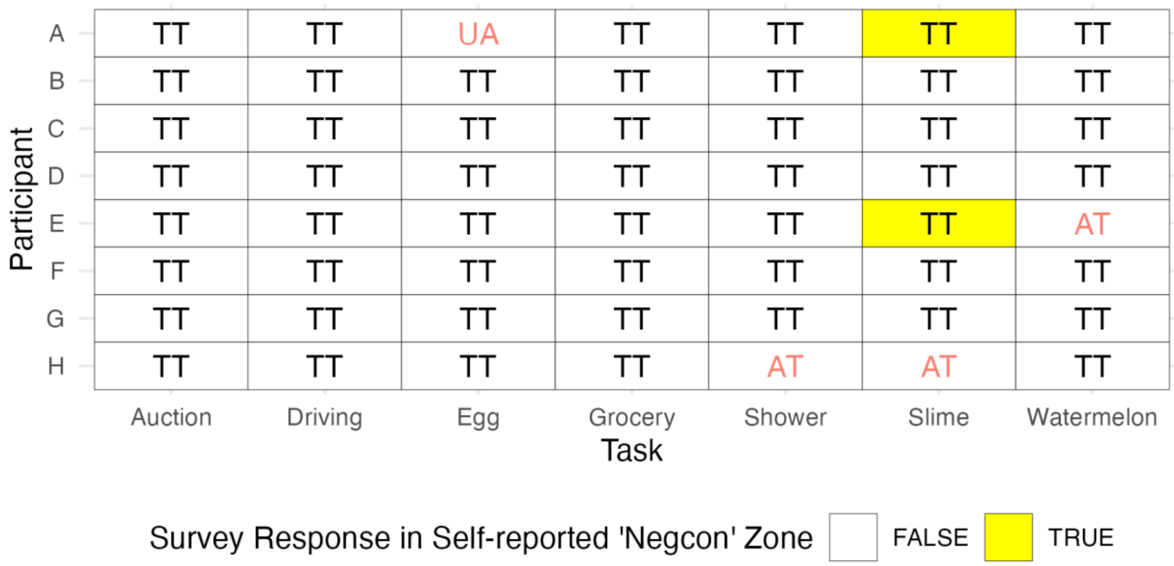

Figure 3. Survey responses and closed codes (from think-aloud data) by two coders: N, A, T correspond to the NAT Taxonomy, while U denotes "uncodable." Coding disagreements (between raters) are shown in red text, while survey responses in the negcon zone are highlighted yellow.

We illustrate a correctly identified targeting via survey response using the Auction task.[3] The Auction task presented a single a dot plot of values for previous winning bids, and asked participants to make a bid for artwork they wanted. Participant E described negative consequences associated with this task,

Participant E. *I think you would want to bid anywhere above maybe the halfway mark because it will be more than any of the leftmost previous bids. It might not guarantee that you'll win, but you can always bid up from there.*

Participant E selected ($999 or less, $1249] as their negcon range, which agrees with their statement "leftmost previous bids". Their numerical choice of $3500 is therefore targeted—they have chosen a value much larger than the upper limit of their negcon zone ($1249), and larger than the greatest previous bid ($2500). Participant E clearly saw the consequences of variation in bids ("might not guarantee that you'll win") and has adjusted their choices to target those consequences.

Disagreement between survey and coding

In two episodes (Slime task, Participants A & E) the survey suggested a non-targeted response, while the analysts coded them as targeted. Both episodes revealed important design and analysis considerations for the survey. First, Participant E's interpretation of the scenario involved stretching to a minimum length, and then continuing,

I think I'll do it at least 100 centimeters. I can't say for sure how long I'll stretch it because slime is variable.

This unintended interpretation of the survey was not a reflection of the participant's data literacy but rather the scenario description, which can be fixed by rewording the question. Meanwhile, Participant A's negcon zone was [110 cm, 130 cm], while their numerical response was 110 cm. Although Participant A's numerical response was technically in their negcon zone, their response lies on the boundary. In this sense, the survey provided a "strict" assessment of targeting; in coding an analyst may observe a participant *attempt* to connect variability to consequence, but make a numerical mistake (e.g., choosing a value that is slightly too large) along the way.

[3] Consult the full survey (available on Figshare) for task details. DOI: 10.6084/m9.figshare.25551687

DISCUSSION, LIMITATIONS, AND FUTURE WORK

Results from both coding and survey responses (Fig. 3 & Table 2) provide a clear answer to RQ 1: The students readily targeted variability in nearly every task (>90%). For this reason, we submit that targeting is a cognitive resource students naturally develop in everyday life (diSessa, 2014). This is in contrast with a misconception—that neglecting variability is a widespread cognitive pathology in need of correction. Compared with our prior study (del Rosario, 2024) of practitioners with engineering-specific tasks (~51% targeted), this suggests factors (such as cognitive load) may interfere with a person's natural resource to target variability. Framing a behavior as a cognitive resource has implications for teaching: A resource-oriented approach recognizes the skills that students bring and seeks to build constructively upon those resources (diSessa, 2014). Practically, an educator can draw comparisons between a student's positive inclination to target variability in an everyday situation (e.g., bidding high in an auction) with targeting in a discipline-specific way (e.g., designing an engineered structure assuming a larger-than-average load).

Our task and survey design limit the interpretation of these results. First, note the choice of everyday scenarios likely to be both understandable and relatable to all participants. Additionally, we used best-practices from the visualization literature (quantile dot plots; Kay et al., 2016) to make information about variability maximally available to participants regardless of their proficiency in data literacy. Thus, the high rate of targeting observed in this study is under "ideal" conditions. Having sketched the range of possible rates of targeting between this (>90% targeted) and a prior study (~51% targeted; del Rosario, 2024), our future goals are to assess the impact of factors on targeting.

Our choice of sample also limits interpretation of the results. Although we believe targeting is appropriately considered a cognitive resource, ours was a purposeful sample of traditional college aged engineering students. It is unclear whether all persons develop the resource to target, and when this cognitive resource for data literacy tends to develop. Future work could investigate different populations: both younger participants (K-12) and those in other disciplines.

Our evidence for the validity of the survey (RQ 2) is, admittedly, mixed. Although a large majority of interview codings agreed with their corresponding survey results, there was little variation in the observed behavior—the vast majority of episodes were targeted. The few cases of neglected variability identified in closed coding (5/88) were not identified by the survey, but by the additional information collected through the think-aloud interviews. Although there are too few such episodes in our data to assess the instrument's accuracy, as presently designed, it seems our instrument tends to overestimate the rate of targeting. The present negcon elicitation approach also gives false positives of non-targeted responses (see *Disagreement between coding and survey*). Results from the Train question (Table 2) suggest that increasing the variability of each component to make the negcon zone more sensitive to neglect of each component may help with this sensitivity problem. Instrument redesign should focus on increasing variation in behavior (task design) and revision of the negative consequence elicitation approach.

ACKNOWLEDGEMENTS

We would like to thank Adrienne Saur for help in clarifying the text of this manuscript. Two reviewers provided detailed comments that improved the clarity of this work. This material is based upon work supported by the National Science Foundation under Grant No. 2138463.

REFERENCES

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287–311. https://doi.org/10.1093/poq/nfm006

Bose, D., Segui-Gomez, ScD, M., & Crandall, J. R. (2011). Vulnerability of female drivers involved in motor vehicle crashes: An analysis of US population at risk. *American Journal of Public Health*, *101*(12), 2368–2373. https://doi.org/10.2105/AJPH.2011.300275

Charmaz, K. (2014). *Constructing grounded theory*. Sage Publications.

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. SAGE Publications.

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*, *32*(4), 432–443. https://doi.org/10.1080/26939169.2024.2308119

del Rosario, Z., Fenrich, R. W., & Iaccarino, G. (2021). When are allowables conservative? *AIAA Journal*, *59*(5), 1760–1772. https://doi.org/10.2514/1.J059578

Deming, W. E. (2018). *Out of the crisis* (Reissue). MIT Press.

diSessa, A. A. (2014). A history of conceptual change research: Threads and fault lines. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 88–108). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.007

GAO. (2023). *Vehicle safety: DOT should take additional actions to improve the information obtained from crash test dummies* (GAO-23-105595). U.S. Government Accountability Office. https://www.gao.gov/products/gao-23-105595

Ginosar, R. (2003). Fourteen ways to fool your synchronizer. *Ninth International Symposium on Asynchronous Circuits and Systems, 2003. Proceedings.*, 89–96. IEEE Computer Society https://doi.org/10.1109/ASYNC.2003.1199169

Hadley, K., & Oyetunji, W. (2022). Extending the theoretical framework of numeracy to engineers. *Journal of Engineering Education*, *111*(2), 376–399. https://doi.org/10.1002/jee.20453

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5092–5103. Association for Computing Machinery (ACM). https://doi.org/10.1145/2858036.2858558

Reinhart, A., Evans, C., Luby, A., Orellana, J., Meyer, M., Wieczorek, J., Elliott, P., Burckhardt, P., & Nugent, R. (2022). Think-aloud interviews: A tool for exploring student statistical reasoning. *Journal of Statistics and Data Science Education*, *30*(2), 100–113. https://doi.org/10.1080/26939169.2022.2063209

Rose, T. (2015). *The end of average: How we succeed in a world that values sameness* (First Edition). HarperOne.

Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed). SAGE.

Vo, K., Evans, A., Madan, S., & del Rosario, Z. (2023). A scoping review of engineering textbooks to quantify the teaching of uncertainty. *ASEE Annual Conference and Exposition*.

Walther, J., Sochacka, N. W., & Kellam, N. N. (2013). Quality in interpretive engineering education research: Reflections on an example study. *Journal of Engineering Education*, *102*(4), Article 4. https://doi.org/10.1002/jee.20029

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