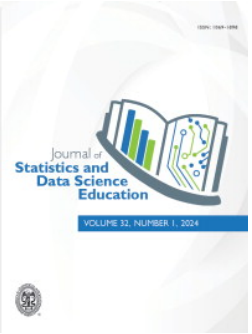


Taylor & Francis
Taylor & Francis Group

Journal of Statistics and Data Science Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ujse21

Neglected, Acknowledged, or Targeted: A Conceptual Framing of Variability, Data Analysis, and Domain Consequences

Zachary del Rosario

To cite this article: Zachary del Rosario (08 Mar 2024): Neglected, Acknowledged, or Targeted: A Conceptual Framing of Variability, Data Analysis, and Domain Consequences, Journal of Statistics and Data Science Education, DOI: 10.1080/26939169.2024.2308119

To link to this article: https://doi.org/10.1080/26939169.2024.2308119

Published online: 08 Mar 2024.

JOURNAL OF STATISTICS AND DATA SCIENCE EDUCATION
2024, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/26939169.2024.2308119

Taylor & Francis
Taylor & Francis Group

Neglected, Acknowledged, or Targeted: A Conceptual Framing of Variability, Data Analysis, and Domain Consequences

Zachary del Rosario

Franklin W Olin College of Engineering, Needham, MA

ABSTRACT
Variability is underemphasized in domains such as engineering. Statistics and data science education research offers a variety of frameworks for understanding variability, but new frameworks for domain applications are necessary. This study investigated the professional practices of working engineers to develop such a framework. The Neglected, Acknowledged, or Targeted (NAT) Taxonomy describes whether one's data analysis choices engage with variability, and whether those choices target the potential consequences of variability, within a given domain. A targeted analysis is the most beneficial rung for engineering applications and is therefore a useful concept for instruction. This study describes the qualitative methods used to develop the NAT Taxonomy and describes how the taxonomy can be used in statistics and data science education, particularly in support of other domain applications.

ARTICLE HISTORY
Received June 2023
Accepted January 2024

KEYWORDS
Engineering education;
Professional practice;
Variability

1. Introduction

Engineers are responsible for designing safe solutions to human problems; these problems are always subject to uncertainty. Variability is a key source of uncertainty in engineered systems. While statisticians have developed sophisticated tools to deal with variability in engineering (e.g., Shewhart 1931), productive responses to variability are not universally deployed. Studies in statistics education (Reading and Pegg 1996; Mathews, Pleasant, and Clark 2007; Zieffler et al. 2008) and behavioral economics (Kahneman and Tversky 1972; Konold 1989) provide ample empirical evidence that individuals are biased in their treatment of variability. Within engineering, neglecting the consequences of variability has the potential for disaster; for instance, design for "the average man" led to uncontrollable and dangerous aircraft, with the ultimate fix being to design explicitly for the variation in human dimensions (Daniels 1952; Rose 2015).

This work is part of a larger study investigating the statistical thinking of practicing engineers. Increasingly, statistics and data science educators are seeking to teach statistical thinking in other domains, such as finance (McCarthy and Kuhlemeyer 2023), biomedical science (Miller and Pyper 2023), and engineering (Huang, London, and Perry 2022). There is limited prior work on statistical thinking in an engineering context (e.g., Hjalmarson 2007; Huang, London, and Perry 2022), motivating the present study. Existing frameworks in statistics education define variability, but tend to focus on statistical inference (Makar and Rubin 2009, 2018). The framework presented here focuses on the domain application consequences of variability. This emphasis on consequence was not an a priori feature of the study, but rather emerged from qualitative analysis of interview data. This work presents a novel taxonomy for understanding data analysis choices in terms of whether an analyst's choices target the consequences of variability. Implications for teaching with the taxonomy are also discussed.

1.1. *Motivating a Consequence-Focused Framing of Variability*

The mishandling of variability has led to engineering disaster. Prior to the 1950s, U.S. aircraft cockpits were designed for the dimensions of the "average man" (Daniels 1952). As aircraft grew more performant from jet technology, the Air Force found that pilots were unable to control these new aircraft, resulting in as many as seventeen crashes per day (Rose 2015). The investigation by Gilbert Daniels (1952) eventually showed that, out of thousands of Air Force pilots, precisely zero were average (across 10 key dimensions simultaneously). In designing their aircraft for the "average man," the Air Force had designed their planes for no one. Modern ergonomic design now rejects the idea of design for average dimensions, instead designing for the variation across human bodies (Watkins 2015). Going beyond the average also promotes diversity and inclusion, as aircraft designed for an average *man* are inherently exclusive of women.

Similar mishandling of variability continues to this day. In automotive design, crash test dummies are presently based on a median male, with females crudely modeled as a scaled-down version of the median male dummy (GAO 2023; Schiebinger et al. n.d). The Government Accountability Office asserts that this poor treatment of variability leads to worse vehicle safety design; crash data from 1998 to 2008 shows that the odds of

CONTACT Zachary del Rosario zdelrosario@olin.edu Franklin W Olin College of Engineering, Needham, MA 02492-1200.

female passenger injury are 47% higher than male passengers (Bose, Segui-Gomez, and Crandall 2011).

While aircraft design now handles variation in people, the standard design practice since the 1960s has been to quantify certain material properties (such as the elasticity and Poisson's ratio) in terms of their sample mean ("Federal Register Vol. 29, No.250, December 24, 1964 - Content Details - FR-1964-12-24" n.d). This data analysis choice leads to a reduction in structural safety that exposes passengers to elevated risk (del Rosario, Fenrich, and Iaccarino 2021). This particular example presents a curious puzzle: In aerospace engineering, material properties such as strength are quantified with highly conservative values: tolerance intervals (Meeker, Hahn, and Escobar 2017) with a minimum sample of $n \geq 100$ (MMPDS-04: Metallic Materials Properties Development and Standardization (MMPDS) 2008). The use of tolerance intervals promotes system safety. However, other properties such as elasticity are quantified with the sample mean, which lead to elevated risk. Both conservative and mean values are considered allowable values for aerospace design, a puzzle that motivates the present study.

The examples above motivate a consequence-focused framing of variability: The issues detailed above stem from a lack of attention to the domain application consequences of variability. The importance of consequence was not an a priori focus of this study; it emerged from the qualitative analysis of professional practices. However, the existing literature on reasoning under uncertainty has limited alignment with this domain application consequence-focused lens. This gap is articulated in the following literature review.

1.2. Literature Review

Variability is broadly considered a challenging but important concept. Biology educators opine that a misunderstanding of variability contributes to student misunderstanding of the theory of evolution (Shtulman and Prassede 2012). Abelson (1995) opines that people tend to believe that deterministic factors are more consequential than chance factors. Statistics educators broadly agree that variability is a core component of statistical thinking (Wild and Pfannkuch 1999; Garfield and Ben-Zvi 2005; Wild, Utts, and Horton 2018; Wood et al. 2018).

While the examples above illustrate the importance of variability to engineering, variability is underemphasized in engineering education research and pedagogy. A recent study on engineering numeracy reviewed the education literature on mathematics within engineering-related disciplines and found only 2 out of 5466 articles that discussed "uncertainty" or "error" (Hadley and Oyetunji 2022). A review of engineering textbooks found that concepts related to uncertainty are considered peripheral to engineering education, as concepts such as "force" appeared 2.5x as frequently as "uncertainty" and 2.0x as frequently as "statistics" (Vo et al. 2023). Statistics is unique as a discipline in its focus on uncertainty (Abelson 1995); therefore, statisticians are uniquely positioned to help engineers reason about variability.

Statistics education researchers have defined several frameworks to articulate statistical thinking in a variety of settings. However, these frameworks tend to be silent on the domain application consequences of variability; they typically focus on inference. Statisticians certainly engage deeply with context and consider domain application consequences of variability in their own practice (delMas 2004; Pfannkuch, Ben-Zvi, and Budgett 2018; Wild, Utts, and Horton 2018). However, the consequences of variability may not be a strong focus when teaching statistics to students. Without this emphasis, students may not transfer their learnings about variability (Barnett and Ceci 2002) and may not view variability as consequential in other contexts. This transfer is particularly fraught, as students often take a single course in statistics (Wood et al. 2018).

An important component of the K-12 conception of statistics is the Advanced Placement (AP) Statistics course,[1] which is organized around four conceptual themes: exploring data, collection of data, anticipating patterns, and statistical inference (Haines 2015). Haines also reviews statistics guidelines from the National Council of Teachers of Mathematics (NCTM), the American Statistical Association (ASA), and the Common Core standards, and finds that these align with the four College Board-defined conceptual themes listed before.

Formal statistical inference concerns the use of sample statistics and hypothesis testing to make statements about a target population. Numerous frameworks from statistics education researchers seek to articulate statistical thinking with a broader conception than formal, probabilistic inference. However, these frameworks generally do not consider the consequences of variability to a domain application. Alacaci (2004) compared expert and novice knowledge of inferential statistics to articulate how statisticians choose inferential tests. The framework of informal inference is a useful generalization of formal statistical inference to K-12 settings, considering generalization beyond data, using data as evidence, and framing statements in probabilistic language (Makar and Rubin 2009). Peters (2011) developed a framework to articulate robust understanding of statistical variation grounded in expert reasoning. Her framework describes variability from three perspectives: design, data-centric, and modeling—but does not discuss consequence. Garfield and Ben-Zvi (2005) developed a framework to assess thinking about variability, which included developing an intuitive sense for variability. However, their framework focuses on "explaining" variability rather than addressing its consequences, "We can try to understand why things vary: By thinking about and examining the variables we can try to explain the different reasons and sources for variability." Arnold and Franklin (2021) developed a framework to help identify good statistical (inferential) questions. Other instruments (Garfield 1998; Watson et al. 2003; Jacobbe et al. 2014; Groth 2014; Harrell-Williams et al. 2015) are similarly silent on the domain application consequences of variability.

Some prior work in statistics education has close connections to the present work. Chance (2002) compared definitions of statistical thinking across authors and synthesized recommendations for instruction: She highlighted "constant relation of data to the context of the problem and interpretation of results in non-statistical terms." This gestures at the consequences of variability, but does not center consequences, nor recommend

[1] However, as an editor of this manuscript noted, the AP Statistics course has not been substantially revised since 1996.

how to make decisions in response to consequences. Similarly, Recommendation 3 from the revised GAISE College report—"Integrate real data with a context and purpose"—may naturally lead to a consideration of consequence (Wood et al. 2018). However, the consequences of variability are not an explicit emphasis of this recommendation.

Wild and Pfannkuch (1999) constructed a detailed model of statistical inquiry, including a rich discussion of the nature of variability. They introduced the dichotomy of real versus induced variability, which informed the design of the present study. Namely, the interview protocol was designed to present research participants with the possibility of real variability through datasets of material properties. Through qualitative analysis of participant responses, it was determined that, under certain circumstances, participants would carefully coordinate their analysis with the perceived consequences of that variability.

This literature review is not meant to suggest that inference is unimportant! The tools of statistical inference are critically important to modern life. However, "the average man" and crash test examples above illustrate cases where a failure to consider the domain application consequences of variability led to loss of human life. Existing statistics education frameworks do not emphasize consequence in this sense—an important emphasis that may be lost when teaching statistical thinking to engineers.

As Hicks and Irizarry (2018) opine, statistics education is often misaligned with the needs of teaching data science in that it rarely connects statistical ideas with solving real-world problems. This study was initiated to develop new theoretical ideas to help bridge statistics with engineering practice. This study engaged in a qualitative, empirical study of practicing engineers to identify beneficial data analysis choices from professional practices.

2. Materials and Methods

This section describes the study design and qualitative data analysis approach. This work was completed under an IRB exempt protocol approved by the Brandeis IRB under protocol number #22134 R-E.

2.1. Theoretical Framework: Knowledge-in-Pieces

This work adopts the theoretical framework *knowledge-in-pieces* (KiP) (diSessa 2019). KiP asserts that knowledge is best understood not as monolithic theories, but rather as smaller knowledge elements. KiP research cannot be supported through pre- and post-intervention studies, but instead employs detailed investigation of "short-term" changes in reasoning using qualitative methods, such as clinical interviewing (diSessa 2007).

KiP frames two key aspects of this study: The importance of varied context in interview design and the interpretation of data analysis choices as beneficial (or not). KiP asserts that context is core to the application of knowledge. Prior studies have documented that reasoning processes change dramatically depending on the context the reasoner considers, such as different key moments in a ball's trajectory (diSessa, Elby, and Hammer 2003) or whether a student considers their own learning or teaching a younger peer (Hammer and Elby 2003). Hence, the clinical interview design for this study varies the context presented to research participants, in order to elicit a variety of responses.

KiP also guides the interpretation of "correctness" in human reasoning. KiP asserts that different knowledge pieces activate in a person's mind as they recognize contextual features in their environment. These pieces are not thought to be correct or incorrect, but rather to have beneficial uses that vary by context (Hammer and Elby 2003). For instance, Elby and Hammer (2001) suggest that treating *knowledge as tentative* in the context of believing the earth to be round (vs. flat) would be an unproductive viewpoint, but that treating *knowledge as tentative* in the context of judging mass extinction theories is productive. This study does not assume a correct data analysis for any task, but does seek to articulate more and less beneficial uses of such choices, depending on the context.

2.2. Study Population: Practicing Engineers

The target population of this study was practicing engineers, with a focus on variability. Studies of practitioners have been used to develop useful educational frameworks, as practitioners have knowledge that is tailored for the industries that students will join. For instance, Peters (2011) studied teachers of AP Statistics to develop a holistic framework for understanding of variability. Peters selected this population "under the assumption that (AP Statistics instructors) were more likely to exhibit robust understandings of variation than secondary mathematics teachers in general." Similarly, Wild and Pfannkuch (1999) synthesized findings from interviews with both students and practicing professional statisticians to develop their model for statistical thinking.

The goal of the broader study is to investigate the statistical thinking of practicing engineers; this work specifically investigates their data analysis choices. Participants were drawn from multiple engineering subfields, in order to maximize variation in observed behavior. However, the interview tasks rely on domain-specific knowledge, which varies by subfield. Hence, the subfields of aerospace, civil, and mechanical engineering were chosen for their common emphasis on the strength of materials and structural design. The interview tasks were designed to present this population with tasks which were hypothesized to vary in difficulty and objective; this was to maximize the chance of seeing a mix of more and less beneficial uses of data analysis choices. The evidence below suggests that this design for variation in behavior was successful.

2.3. Sample: Recruitment and Description

Participants were contacted via snowball sampling, initiated through the professional network of the lead author and via discipline-specific LinkedIn groups. Informed consent was obtained at the recruitment stage and again at the beginning of each interview.

Selection criteria required that participants possess a college-level engineering degree, have at least two years of professional experience, and be at least somewhat familiar with the concepts of mechanical stress, strain, yield failure, and buckling. These criteria were self-reported at the recruitment stage, but the interview protocol included warmup tasks to assess participant

Table 1. Summary of participant demographics.

Experience	Two years: 3	Three years: 2	Four years: 8	Five+ 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		

understanding of relevant phenomena (see below). Participants were chosen based on self-reported subfields: aerospace, civil (including geotechnical and structural), and mechanical (including bioengineering and manufacturing). One participant (15) completed an engineering B.S., but now works in the financial sector (subfield other). Out of 273 individuals who completed the recruitment form, 26 participants were invited to emphasize a diverse representation of races, genders, alma maters, and engineering subdisciplines.

Ultimately n = 24 persons agreed to join the study and participate in interviews, a sample size in-line with recommendations for qualitative research (Creswell 2014) and comparable with those from other qualitative statistics education research studies (Peters 2011; Reinhart et al. 2022; Glantz et al. 2023). Table 1 describes the sample.

Compared with degrees awarded in 2020 (IPEDS 2020), the 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).

Participants were incentivized to join the study with an offer to join a professional development course on data science in engineering, taught by the author. Participants completed their interview before the professional development course began. Interviews were conducted by the author and trained research assistants following a structured interview protocol. Interviews were conducted via web conferencing software (Zoom), video recorded, then professionally transcribed. Interviews typically lasted one hour, though varied from 45 min to over 1.5 hr. The interview protocol is described next.

2.4. Clinical Interview Design and Analysis

Interviews followed a structured clinical interview protocol (diSessa 2007). This work is a subset of a larger study; only relevant portions of the full interview protocol are described here. All interviewers used an identical interviewer guide and shared a slide deck over Zoom to introduce each prompt; these materials are linked in Appendix C and described briefly Q1 here.

The interview began with a brief warmup that had participants describe (in their own words) structural engineering fundamentals: the concepts of stress and strain, material properties such as density, elasticity, Poisson's ratio, and strength. This warmup served to acclimate participants to narrating their thought process, which facilitated later stages of the interview (Reinhart et al. 2022). The warmup also included a schematic (Figure 1) that contextualized the data as having the possibility of real variability (Wild and Pfannkuch 1999)—but did not name this concept directly.

The structured interview included ten tasks, the majority of which asked participants to study a dataset (with 10 observations) and conduct an analysis (either description or design). Participants were provided with a visual aid via the slide deck, and with a text reminder of the prompt read aloud by the researcher. For instance, Figure 2 gives the slide that accompanied the first task (Q1) where participants were asked to describe an elastic alloy using a dataset.

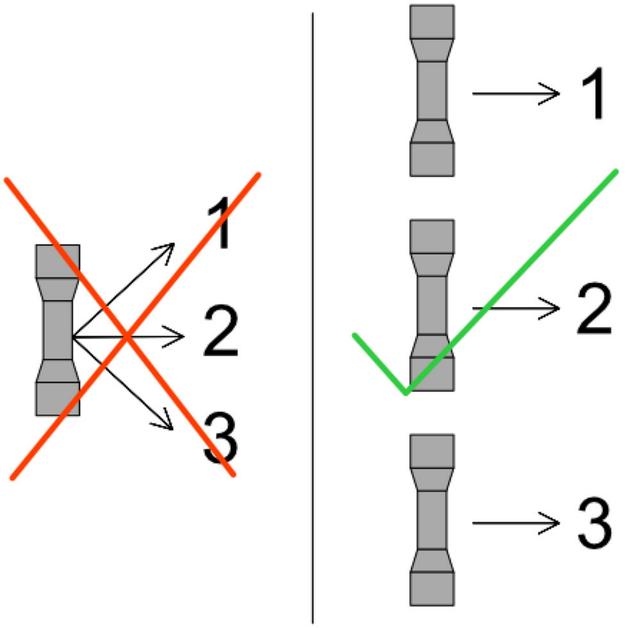


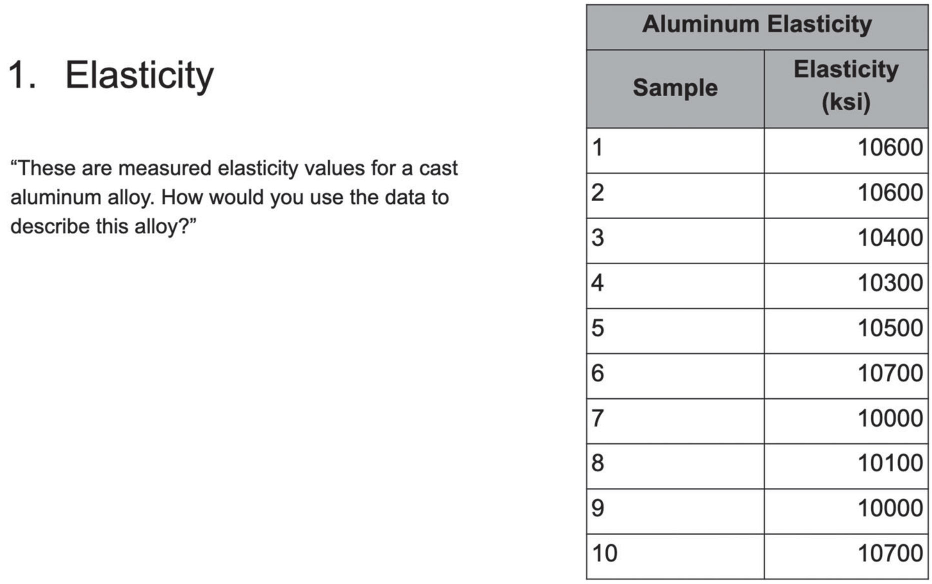
Figure 1. Schematic figure used to clarify the context of the data presented to participants. Under the scenario of measurements from independent specimens, the possibility of real variability is clarified (without explicitly describing the concept).

Interview questions were carefully designed not to promote a specific data analysis approach (e.g., Figure 2). In this way, participants were not guided toward using a specific number of numerical summaries. Ultimately 23/24 participants offered more than one summary in their interview. Follow-up questions were included in the interview protocol to contextualize participant responses. For instance, a follow-up was designed to have participants justify their analysis choices,

Researcher (R). (For each quantity [mean, standard deviation, etc.] a participant uses) "Why did you use [that quantity]?"

The interview tasks were designed to vary contextual features hypothesized to lead to different analysis choices: the task *direction*, the *property* considered, and provided *artifacts*. The *direction* of the task was either to describe the data or design with the data. Engineers tend to exercise caution when designing, but simply describing a dataset does not have the same risk-averse connotation. Tasks were designed to involve different material *properties*, inspired by the allowables puzzle. In some tasks, an engineering artifact was provided (an equation or diagram) to aid with reasoning, which was intended to modulate the difficulty of reasoning about variability in context. Table 2 summarizes all interview tasks.

The design tasks asked participants to design simple structures using the provided data. The strength design task had participants design a simple structural member to survive uniaxial tension (at risk of yield), the elasticity design task focused on designing a column to survive a compressive load (at risk of buckling), and the density design task asked participants to



1. Elasticity

“These are measured elasticity values for a cast aluminum alloy. How would you use the data to describe this alloy?”

Aluminum Elasticity	
Sample	Elasticity (ksi)
1	10600
2	10600
3	10400
4	10300
5	10500
6	10700
7	10000
8	10100
9	10000
10	10700

Figure 2. Example slide from interview materials (Q1, the describe elasticity task). Slides displayed the dataset to be studied and the planned interview prompts.

Table 2. Summary of data analysis tasks in the interview protocol.

Question	Direction	Property	Equation	Diagram
1	Describe	Elasticity	No	No
2	Describe	Strength	No	No
3	Design	Strength	No	No
4	Design	Density	No	Yes
5	Design	Elasticity	Yes	Yes
7	Design	Strength	Yes	Yes
9	Describe	Poisson’s Ratio	No	No

NOTE: Questions 6 and 8 were not data analysis tasks, and are hence excluded from this study.

design a neutrally buoyant hollow sphere (which could either float or sink). While the tension and buckling problems have straightforward, canonical ways to encourage safety in design (using lower values), the neutrally buoyant sphere problem is more challenging, as both higher and lower values of density will lead to a state of failure. As can be seen in the full set of coded episodes (Appendix B), this led to a more difficult reasoning task for participants.

Interview transcripts were initially coded using elemental methods: descriptive, process, and in vivo coding (Saldaña 2013). While the final analytic product of this study was a closed coding scheme, these open coding methods were used to maintain close agreement between the data and the developing analysis (Charmaz 2014). Initial coding was performed by all members of the research team, conducted individually on ∼4 interviews per analyst. This open coding identified data analysis choices (descriptive coding) and justifications (process and in vivo coding) in the transcripts. Peer review and full-team debriefings were used to revise and converge on shared meaning of codes, to highlight episodes that other researchers missed, and to revise coding choices as the codebook evolved. Once the full team reached consensus on common meaning in the codes, a complete coding of the full corpus was completed. This open coding served as the starting point for taxonomy development.

2.5. *Taxonomy Development*

Prior work on mathematical reasoning among structural engineers has found that mathematics is both essential to and inadequate for engineering practice (Gainsburg 2007). Specifically, mathematical analysis is required in engineering work, but engineering judgment is used to select appropriate yet unprovable assumptions on which that analysis is based. Thus, this study focuses on participants’ data analysis choices and their justifications for those choices. The analytic goal was to produce a taxonomy to categorize participant behaviors in response to each task. Note that the focus on domain application consequence was not an explicit a priori goal of analysis; as discussed below, this was a finding that emerged from the data that stabilized the taxonomy, and ultimately framed the writing of this manuscript. The author and one research assistant collaborated to develop a closed coding scheme to realize the taxonomy: The approach is described here.

Based on a pilot study (Aggarwal et al. 2021) and inspired by the allowables puzzle stated above, an initial two-level taxonomy was proposed corresponding to the dualistic approach in aerospace design: a central value or a conservative value. However, initial coding showed that participants frequently used a wide variety of mathematical choices, including different summary values (mean, median, standard deviation), extrema (minimum and maximum), distribution quantities (quantiles), and explicit probabilistic calculations (e.g., probability of floating). Often, participants would state multiple choices for a single task. The empirical reality exhibited by participants rejected the initial two-level taxonomy, suggesting that an intermediate level was necessary. The revised lowest rung still corresponded to use of a single value, while the intermediate level corresponded to the use of multiple analysis choices to *acknowledge* the variability.

The author and a research assistant conducted independent (closed) coding using the developing closed coding scheme on a random subset of the data. The initial version of the scheme focused exclusively on the analysis choices of participants,

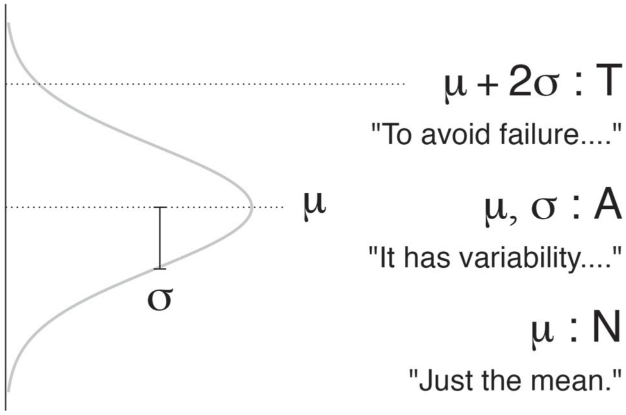


Figure 3. Schematic depiction of the Neglected, Acknowledged, or Targeted (NAT) Taxonomy in the case of an applied load. In this case, catastrophic failure may occur at larger values. Analysis indicators and justifications given here are examples only; the full NAT closed coding scheme is given in Appendix A.

which led to an unacceptably low level of agreement. Through debriefing and review of coded transcripts, it was determined that participants who justified their analysis choices in terms of the consequences of variability exhibited a distinct behavior, deemed *targeting*. Revising the taxonomic coding scheme to emphasize this finding and re-coding on a new subset of the data, a substantial level of interrater reliability (IRR) was achieved (Cohen's $\kappa = 0.726$, $n = 21$) (Landis and Koch 1977). Incorporating IRR assessment techniques in the qualitative approach surfaced and resolved holes in the taxonomy.

It is important to note that the taxonomy's aim is to categorize engineers' episodic behaviors, not engineers themselves. The interview protocol was designed to present different contextual features, based on the context-dependent nature of knowledge implied by the KiP Framework.

3. Results

3.1. The NAT Taxonomy

The finalized taxonomy is operationalized via a closed coding scheme (Appendix A). The Neglected, Acknowledged, or Targeted (NAT) Taxonomy seeks to describe a person's data analysis choices in an engineering context according to one of three rungs:

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.

Figure 3 provides an illustration of the taxonomy in a case where variability is consequential to the application: The design of a structural member subject to a variable applied load. Use of the mean alone is a neglected analysis. The mean ignores cases where the applied load is larger; a design based on the mean load may suffer catastrophic failure in these neglected cases. Computing the mean and standard deviation is an acknowledged analysis. This quantifies the variability in the load, but does not yet incorporate that variability into design. One way to produce a targeted analysis is to combine the mean and standard deviation to produce a conservative upper value.

Appendix B reports the NAT (closed) codes for all participants and data analysis tasks. Full investigation of this pattern is outside the scope of the present work. However, this full coding does provide evidence that the interview protocol design was successful at eliciting a variety of participant responses. For instance, participants largely targeted variability in design questions (such as Q5 and Q7), but had more difficulty targeting the variability in the sphere design question (Q4). Furthermore, participants generally targeted at a lower rate for describe questions (Q1, Q2, Q9) than design questions.

Note that participant justifications are necessary to apply the NAT coding scheme; these were elicited in the present study through clinical interviews. The remainder of this section illustrates the application of the NAT Taxonomy to interview excerpts.

3.2. Neglected Rung

An analysis neglects variability most commonly by reporting the average alone. For instance, Participant 1 used a neglected analysis for every interview task (Appendix B) by computing the average of the data alone,

P1, Q1: "I think normally in this kind of case I would average. I would average out a sample from 1 to 10 and find the average of the value of the elasticity of the aluminum."

The use of the mean as the sole summary of a dataset is an implicit neglect of variability. However, some participants explicitly described their neglect of variability, even in the context of design (Q3),

P18, Q3: "I'm just guessing the average would be something like 157 but I guess just looking at these numbers, if it was some material that was made up completely, but if I just am doing a simple uniaxial tension design, I will just pick the average and not worry too much about the deviations."

The mean is the quintessential example of a data analysis that neglects variability, as reported in numerous previous studies (Reading and Pegg 1996; Hjalmarson 2007; Mathews, Pleasant, and Clark 2007). All 24 participants mentioned the mean at least once in their interview. However, as seen below, use of the mean does *not* necessarily correspond to a neglected analysis. Moreover neglected analyses were rare: participants more frequently acknowledged or targeted variability (Appendix B).

3.3. Acknowledged Rung

An acknowledged analysis recognizes the existence of variability, but does not address consequences. Participant 3 automatically deployed multiple statistics in response to Q1:

R: "How would you use the data to describe this alloy?"

P3, Q1: "That there is variability in the alloy and then I would probably try to find the mean and how much it varied over this sample of 10."

In this instance, Participant 3 used an unspecified quantity for variability ("how much it varied"), which acknowledges the variability informally. Similar to the concept of informal inference (Makar and Rubin 2009), the coding scheme takes informal measures of variability as indicators of an acknowledged analysis. Interestingly, an indicator for an informal central value was not necessary, only for informal measures of variability. This suggests that measures of central tendency were much more available than measures of spread among participants.

Participant 4 took a different approach to acknowledge variability by using formal summaries (mean and standard deviation) and by seeking a distribution model,

P4, Q1: "We have 10 independent tests. For example, I can talk about the mean stiffness—elasticity—of the aluminum alloy. I can talk about the variation observed across the ten tests in terms of, for example, standard deviation. I can also put a distribution curve based on this 10 data, and that will give me information, let's say, about… for example, if we anticipate lots of variation across tests in the value recorded, that means we might need to consider larger specimen. We might want to do more testing."

Despite the productive ideas expressed here about statistical inference (fitting a distribution, gathering a larger sample), this analysis is still not targeted as it does not address the consequences of variability.

3.4. Targeted Rung

A targeted analysis accounts for the application-specific consequences of variability, which necessarily includes some acknowledgement of variability. In response to the steel description question (Q2), Participant 4 specialized her analysis to target the consequences of variability:

P4, Q2: "Yes. I will definitely look into the variance or the standard deviation, even in this case, the distribution I might be interested to look at, because as far as my experience goes, we don't use the mean value strengths for practical design. We might be interested in the fifth percentile strength from this dataset because we want to be on the safer side."

Here, Participant 4 selects a specific quantile of the population ("fifth percentile") to mitigate the consequences of variability ("be on the safer side"), and actively rejects the mean for the purposes of design. This is a design approach commonly used in Canadian Civil Engineering, reflecting the traditions of her training ("Design Values for Canadian Species Used in Canada" n.d; Madsen 1975; Fan, Wong, and Zidek 2023). Her analysis clearly targets the consequences of variability.

Once a design context was added to the tensile strength data (Q3), Participant 3 started to combine her statistical measures to target consequences.

R: "Here would you also compute the standard deviation like with the previous cases?"

P3, Q3: "Yes, since for a load-bearing design, I'd want to make sure that I designed it well below at least one standard deviation of that Tensile Yield Strength to make sure that it never deforms to the point where it won't return to its original value."

Combining the mean with the standard deviation to construct a lower value is one way to perform a targeted analysis. Here, Participant 3 constructs this conservative value to avoid a potentially adverse outcome ("make sure that it never deforms… where it won't return").

Constructing a targeted analysis in response to the sphere design question (Q4) was rare (Figure 5), but some participants noted the consequences of variability in density and used the variability to inform their design process. Participant 17 described using the standard deviation to assess the expected number of built objects that would achieve the design goal,

P17, Q4: "… if it's a very small standard deviation you're going to find that very few are outside of spec, or if it's a very large standard deviation, which I can't really do by eye, so I'm not going to even try."

Participant 17's reasoning here is an informally probabilistic analysis of variability, expressed in terms of a failure rate ("few are outside of spec"). While other analyses among participants were aimed at preventing the consequences of variation, this analysis describes a more detailed prediction of frequency-quantified consequences.

While use of the mean without justification is a neglected analysis, Participant 6 justified his use of the mean by deeming the variability small and inconsequential,

P6, Q9: "These numbers vary in very small quantities. … I'm not thinking that there's a lot of concern with quoting the mean value in this case. If we're working with some very precision equipment where the contraction for a given load is really important, then I might use the maximum value as a design, but in run-of-the-mill calculations that I would do on a daily basis, I almost think that using the mean value from this set of data is adequate."

Analysis of this episode turns on the interpretation of "correctness" according to the KiP Framework: Recall that knowledge elements in KiP are not thought to be correct or incorrect, but rather to have more or less beneficial uses, depending on the context. While both Participant 1 and Participant 6 ultimately choose to use the mean in task Q9, Participant 6 justifies his use of the mean in terms of a lack of consequences of variability. His justification rests on assessing the variability as small, which he operationalizes using an informal quantification of variability. Essentially, the mean is an analytic tool that can be used for more or less beneficial applications: to either target or neglect variability.

4. Discussion

The data analysis choices and justifications of practicing engineers were studied to produce the Neglected, Acknowledged, or Targeted (NAT) Taxonomy: Their analyses either neglect variability in data, acknowledge variability, or target the consequences of variability. A targeted data analysis choice is one that engages with variability and supports addressing the domain application consequences of that variability. By targeting the consequences of variability, data analysis choices are by-construction useful for an application's goals. Therefore, a targeted analysis is the most beneficial rung of the NAT Taxonomy.

For instance, "the average man" episode (documented in the Introduction) utilized a neglected analysis, as no study of variability was conducted prior to that by Gilbert Daniels (1952). The solution to "the average man" episode used a targeted analysis: defining ranges for adjustable seats based on the observed pilot variation. Due to a tighter correspondence of data analysis choices with engineering considerations, a targeted analysis will generally lead to safer engineering decisions.

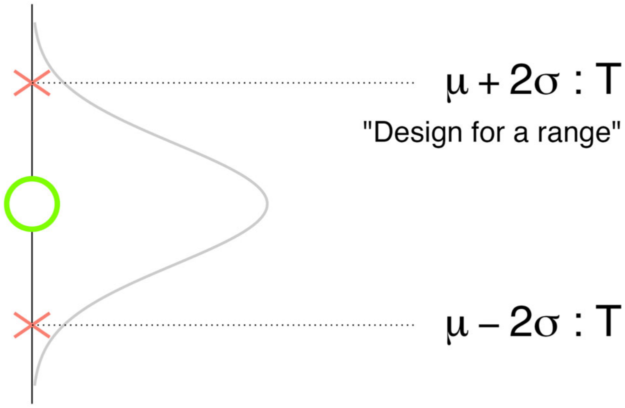


Figure 4. A targeted analysis of "the average man" episode. Failure occurs at both high and low values of a measured body dimension (consequences); therefore, a targeted analysis would produce a range of values—based on the observed variability—that subsequent design would need to accommodate (analysis).

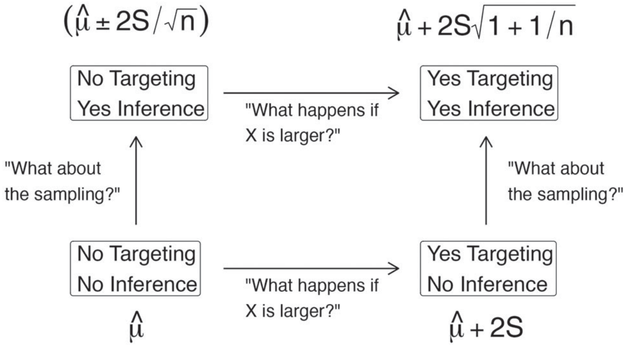

$(\hat{\mu} \pm 2S/\sqrt{n})$
$\hat{\mu} + 2S\sqrt{1+1/n}$
No Targeting
Yes Inference
Yes Targeting
Yes Inference
"What happens if X is larger?"
"What about the sampling?"
"What about the sampling?"
No Targeting
No Inference
Yes Targeting
No Inference
"What happens if X is larger?"
$\hat{\mu}$
$\hat{\mu} + 2S$

Figure 5. A hypothetical teaching episode integrating targeting and statistical inference. The scenario concerns a random quantity X where larger values lead to failure. Math expressions represent student work, while arrows and quotations represent teaching interventions. Two teaching paths are depicted, starting from engagement with either the sampling or the consequences of variability.

In a more modern context, the ongoing disparities in automotive crash injuries (∼50% higher odds for women) can be attributed in part to a neglected analysis (Bose, Segui-Gomez, and Crandall 2011). Recall that modern crash testing attempts to represent women by using a scaled version of the median male crash test dummy (GAO 2023). This post hoc adjustment of median male data neglects not only the variation among people, but also the systematic anatomical differences between male and female bodies; for instance, the Government Accountability Office notes that the described practice "may not adequately reflect females' greater risk of lower leg injuries in crashes than males." A targeted analysis in this context would use data on female occupants (rather than make post hoc adjustments to male data) to inform appropriate crash test dummy design.

The revised GAISE College report reflects the profession's consensus on teaching introductory statistics (Wood et al. 2018). Recommendation 3 suggests that instructors "Integrate real data with a context and purpose." For introductory statistics aimed at engineering students, the NAT Taxonomy is one way to teach context and purpose grounded in the professional practices of engineers. Teaching students to target the consequences of variability will help them connect their data analysis choices to engineering design; the following examples illustrate how this might be done.

4.1. Teaching with the NAT Taxonomy

This work introduces the NAT Taxonomy, which was developed to categorize the data analysis choices and justifications of practicing engineers. While formal development of teaching interventions is outside the scope of the present work, it is important and productive to sketch these ideas.

Since a targeted analysis represents an engagement of analysis with the consequences of variability, this suggests a pattern of teaching intervention: encourage students to consider the consequences of the observed variability, then choose an analysis that mitigates those consequences.

Figure 4 illustrates an application of this consequence-then-analysis proposal to "the average man" episode. The distribution (Figure 4) depicts the variability in a single bodily dimension (say, an arm length). Clearly, a pilot will have difficulty controlling an aircraft if their arm is either longer or shorter than some assumed value. Since adverse outcomes may occur for both the upper and lower sides of the distribution, this suggests that designing for a range of values is necessary. Additional engineering work would be required to address this range (e.g., designing an adjustable seat), but choosing a range that reflects the observed variation (e.g., using the standard deviation) would be a targeted analysis that sets the design requirements.

4.2. Integrating Targeting and Statistical Inference

The variability targeting concept presented in this study is intended to help bridge the gap between statistics/data science and engineering applications. While distinct from statistical inference, this work should not be interpreted as questioning the value of inference. Rather, the full value of targeting variability will only be realized by integrating these ideas with statistical inference.

To illustrate integrating variability targeting and statistical inference, Figure 5 depicts a hypothetical teaching episode. In this case a student is working on an engineering design problem with data on an applied load, represented by X. As with many structures, failure is more likely to occur when X is larger. Engineering design in this context typically proceeds by assessing and reducing the risk of failure of the design. An engineer conducts design by adjusting design features (e.g., the thickness of structural members) to reduce the risk of failure to an acceptably low level, based on analysis. This analysis requires a quantification of the loads, which here exhibit variability. The student initially chooses to take the sample mean and base their design on this quantity.

An instructor responding to the initial student work in Figure 5 should respond to both the inferential and targeting limitations of the mean alone. The instructor in this hypothetical episode asks questions to direct the student's attention to limitations in their work. While there are multiple possible pathways through this teaching episode (two are depicted in Figure 5), attending to both inference and targeting will result in inferential conclusions that appropriately target the consequences of variability.

4.3. Limitations and Future Work

The sample for this study was a snowball sample of n = 24 engineers with professional experience. This approach was appropriate for the qualitative methods used within this KiP-framed study; pre–post investigations cannot support a KiP-framed research project (diSessa 2019), and the sample size was appropriate for the qualitative analysis methods employed (Creswell 2014). While the sample design does not threaten the utility of the developed NAT Taxonomy, it does present important limitations to inferences that can be drawn. For instance, while it was observed that 21/24 participants produced at least one targeted analysis in their interview, this cannot be used to conclude that a majority of engineers will target variability in practice. Future work would be necessary to address such questions.

The NAT Taxonomy describes data analysis choices *only*, which limits the scope of the concept. For instance, safety factors are deliberately left out of this framework. This was important to analyze engagement with variability in the data; in professions such as civil/structural engineering (Galambos 1981) and aerospace engineering (del Rosario, Fenrich, and Iaccarino 2021), safety factors are required by regulations. Therefore, use of a safety factor may reflect engagement with variability, consideration of other factors (such as uncertainty in the loads or models), a "by-the-book" approach, or some combination thereof. A safety factor may obviate the danger of a neglected analysis; this means the NAT Taxonomy cannot be considered a comprehensive criterion for the soundness of an engineering analysis. Future work with careful interview protocol design would be necessary to tease apart what engineers consider when reasoning with safety factors.

The context of the interviews necessarily limits the actions participants could exhibit: For data analysis in a professional setting, an engineer could interact with the data using whatever tools they find useful, such as an interactive visualization. While some participants described how they would use such tools, it was not possible to observe participants using those tools directly. This setting may have biased participants toward easy-to-describe numerical summaries (e.g., the mean, standard deviation) rather than more elaborate approaches (e.g., a domain-specific visual). Given the central importance of mathematics in engineering practice (Gainsburg 2007), it is important to study how practitioners select summaries in data analysis. However, future work could develop the NAT Taxonomy to consider broader elements of data analysis. For instance, in situ studies with professionals would likely elicit behaviors that could not be observed in this study.

The sample included engineers with industry experience in Aerospace, Civil, and Mechanical subdisciplines. Furthermore, participants were presented with datasets on material properties only. This study design was important to ensure all practitioners had some domain-specific knowledge of the quantities that varied; however, this presents limitations to the generalizability of results. This obviously affects the operationalization of the taxonomy (closed coding scheme), but may present challenges to applying the NAT Taxonomy itself. Some participants used highly specific data analysis choices endemic to subdisciplines (e.g., the 5th percentile strength); qualitatively different analysis choices may exist in other areas of professional practice.

The NAT Taxonomy has potential but untested applications to other domains, such as health or social sciences. The consideration of consequence is likely to transport to other domains, but the details are likely to differ substantially. For instance, clinicians in medical work must think about consequences not only in terms of patient outcomes (analogous to engineers considering safety), but also patient comfort and relations. Future work could test and expand the NAT Taxonomy in other domains.

Table A1. Neglected-Acknowledged-Targeted (NAT) Taxonomy coding scheme.

Short	Long	Indicators
Neglected (N)	Analysis neglects variability by using a single value	- Uses a measure of central tendency only (e.g., mean or median), no spread or extreme values
Acknowledged (A)	Analysis acknowledges, but does not address the consequences of, variability	- Uses a measure of central tendency AND a measure of variability (e.g., standard deviation, interquartile range, or an unspecified measure of variability) - ALTERNATIVELY, uses multiple extreme values (e.g., min and max, multiple quantiles) - ALTERNATIVELY, uses a distribution to fully describe the variability
Targeted (T)	Analysis addresses the consequences of variability	- Integrates central tendency and variability to construct an appropriate extreme value - e.g., mean − 1 * sigma for strength, mean + 1 * sigma for load - Note that an unspecified quantity for variability would not qualify - ALTERNATIVELY, focuses on the appropriate quantile or extreme value - e.g., min for strength, max for load, lower quantile for strength - Note that paying equal attention to min and max would not qualify - Note that using a distribution without specifying a quantity derived from that distribution would not qualify - ALTERNATIVELY, describes finding reasonable bounds that the measured quantity could land within, and still achieve a desired outcome - e.g., the density of the metallic sphere can land within a specified range, and still be (effectively) neutral buoyant - ALTERNATIVELY, argues that the mean is appropriate, as the variability is sufficiently small - This approach makes use of a measure of variability, including informal variability - ALTERNATIVELY, uses a distribution to quantify the probability of an event that is relevant to the desired outcome

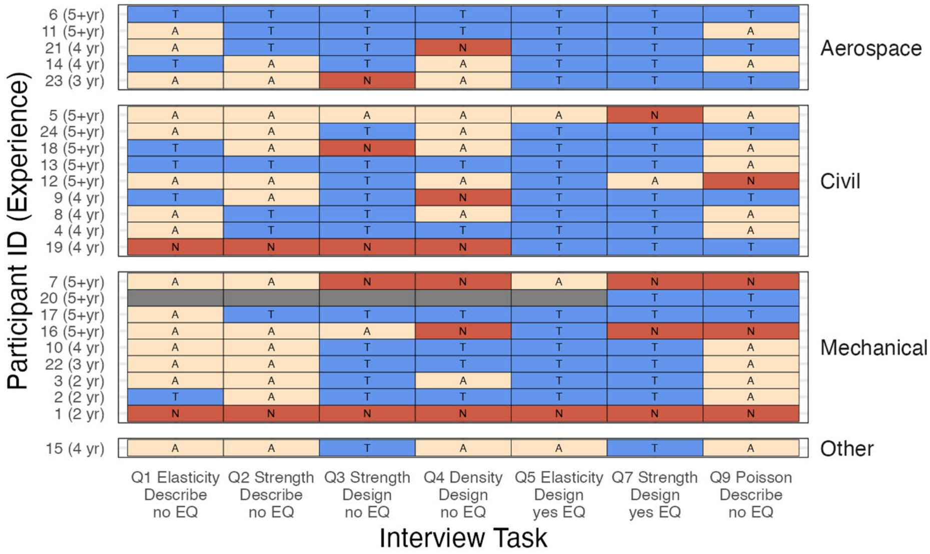


Figure B1. Closed NAT codes for all data analysis tasks and participants.

While the NAT Taxonomy may be used to design teaching interventions (e.g., to encourage targeting), it is not itself a teaching intervention. The examples illustrated above (Figures 4 and 5) demonstrate how instruction might be guided by the NAT Taxonomy; however, future work could use the ideas developed here to design and formally test teaching interventions.

Appendices

Appendix A: Neglected, Acknowledged, or Targeted Taxonomy: Coding Scheme

A Table A1 provides the closed coding scheme for the Neglected, Acknowledged, or Targeted (NAT) Taxonomy. This is a coding scheme about the data analysis choices a person makes when encountering variability. Participants may use other sophisticated engineering procedures (like safety factors or adjusting the manufacturing process) that are not considered indicators for the purpose of this coding scheme.

When applying this coding scheme, assign the "highest" rung that you can, based on participant behavior. For instance, if they describe both a Neglected and Acknowledged analysis, code it as Acknowledged. Note that: Neglected $\ll$ Acknowledged $\ll$ Targeted.

Appendix B: Full Coding Results

B Figure B1 displays all NAT (closed) codes for all codable instances and all interview participants. A full investigation of these results is outside the scope of the present work.

Appendix C: Interview Protocol Details

C This section provides further details on the interview protocol. The as-used materials are published in open-access form with the following DOIs:

- Interview slides: 10.6084/m9.figshare.23552808
- Interview guide: 10.6084/m9.figshare.23552844

Interviewers followed the interview guide to conduct the clinical interviews, and presented the interview slides to research participants (using screen-share on Zoom). The following material provides references for the datasets used in the interview and relevant engineering structural mechanics theory.

Table C1. Datasets used in interview protocol.

Tensile Yield Strength (ksi)	Elasticity (ksi)	Poisson's Ratio (−)
157.0	10600	0.321
159.6	10600	0.323
155.6	10400	0.329
165.8	10300	0.319
157.4	10500	0.323
158.4	10700	0.328
157.6	10000	0.315
156.4	10100	0.312
157.7	10000	0.311
155.7	10700	0.321

Table C1 presents the datasets shown to participants. Strength values are the tensile yield of a cast steel (Ruff 1984), while elasticity and Poisson's ratio are of a rolled aluminum alloy (Stang, Greenspan, and Newman 1946). All datasets were small (10 observations) to avoid overwhelming participants.

Questions 5 and 7 also provided equations to help participants reason about the physical behavior of the system. Question 5 presented the Euler's equation for the buckling load of a column with fixed ends (Salmon, Johnson, and Malhas 2009),

$$F_{cr} = \frac{\pi^2 I}{(\frac{1}{2}L)^2} E,$$

where I is the second moment of area of the beam cross-section, L is the length of the column, and E is the elasticity of the material. Question 7 presented the applied stress for a constant cross-section member in uniaxial tension,

$$\sigma_{app} = F/A$$

where F is the applied load and A is the cross-sectional area of the member.

Acknowledgments

The research assistants on this project were AJ Evans, Ellie Ramos, Emma Fox, KD Vo, and Maeve Stites. Emma Fox also assisted with taxonomy development and the interrater reliability study.

Data Availability Statement

Participants in this study agreed to be quoted in published research, but did not consent to sharing the full contents of their interviews. Therefore, the raw interview data is not available. The deidentified coded data and analysis script are available via a digital object identifier (DOI): 10.6084/m9.figshare.24298951.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

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

ORCID

Zachary del Rosario http://orcid.org/0000-0003-4676-1692

References

“Design Values for Canadian Species Used in Canada (n.d), ” Canadian Wood Council.

“Federal Register Vol. 29, No.250, December 24, 1964 - Content Details - FR-1964-12-24” (n.d), Available at *https://www.govinfo.gov/app/details/FR-1964-12-24*.

Abelson, R. P. (1995), *Statistics as Principled Argument*, Hillsdale, NJ: L. Erlbaum Associates.

Aggarwal, R., Flynn, M., Daitzman, S., Lam, D., and del Rosario, Z. (2021), “A Qualitative Study of Engineering Students’ Reasoning about Statistical Variability,” in *American Society for Engineering Education*, p. 17.

Alacaci, C. (2004), “Inferential Statistics: Understanding Expert Knowledge and Its Implications for Statistics Education,” *Journal of Statistics Education*, 12, 6. DOI:10.1080/10691898.2004.11910737.

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

Barnett, S. M., and Ceci, S. J. (2002), “When and Where Do we Apply What we Learn?: A Taxonomy for Far Transfer,” *Psychological Bulletin*, 128, 612–637. DOI:10.1037/0033-2909.128.4.612.

Bose, D., Segui-Gomez, M., and 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, 2368–2373. DOI:10.2105/AJPH.2011.300275.

Chance, B. L. (2002), “Components of Statistical Thinking and Implications for Instruction and Assessment,” *Journal of Statistics Education*, 10, 1–14. DOI:10.1080/10691898.2002.11910677.

Charmaz, K. (2014), Constructing Grounded Theory, London: SAGE Publications.

Creswell, J. W. (2014), *A Concise Introduction to Mixed Methods Research*, London: SAGE Publications.

Daniels, G. (1952), *The “Average Man”?*, Wright-Patterson AFB OH: Air Force Aerospace Medical Research Lab.

del Rosario, Z., Fenrich, R. W., and Iaccarino, G. (2021), “When Are Allowables Conservative?,” *AIAA Journal*, 59, 1760–1772. DOI:10.2514/1.J059578.

delMas, R. C. (2004), “A Comparison of Mathematical and Statistical Reasoning,” in *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, eds. D. Ben-Zvi and J. Garfield, pp. 79–95, Dordrecht Springer. DOI:10.1007/1-4020-2278-6_4.

diSessa, A. A. (2007), “An Interactional Analysis of Clinical Interviewing,” *Cognition and Instruction*, 25, 523–565.

diSessa, A. A. (2019), “A Friendly Introduction to ‘Knowledge in Pieces’: Modeling Types of Knowledge and Their Roles in Learning,” in *Compendium for Early Career Researchers in Mathematics Education, ICME-13 Monographs*, eds. G. Kaiser and N. Presmeg, pp. 245–264, Cham Springer. DOI:10.1007/978-3-030-15636-7_11.

diSessa, A., Elby, A., and Hammer, D. (2003), “J’s Epistemological Stance and Strategies,” in *Intentional Conceptual Change*, Mahwah, NJ: L. Erlbaum Associates.

Elby, A., and Hammer, D. (2001), “On the Substance of a Sophisticated Epistemology,” *Science Education*, 85, 554–567. DOI:10.1002/sce.1023.

Fan, S., Wong, S. W. K., and Zidek, J. V. (2023), “Knots and Their Effect on the Tensile Strength of Lumber: A Case Study,” arXiv.

Gainsburg, J. (2007), “The Mathematical Disposition of Structural Engineers,” *Journal for Research in Mathematics Education*, 38, 477–506.

Galambos, T. V. (1981), “Load and Resistance Factor Design,” *Engineering Journal AISC*, 9, 74–82.

GAO. (2023), Vehicle Safety: DOT Should Take Additional Actions to Improve the Information Obtained from Crash Test Dummies, U.S. Government Accountability Office.

Garfield, J. B. (1998), “The Statistical Reasoning Assessment: Development and Validation of a Research Tool,” in Proceedings of the Fifth International Conference on Teaching Statistics, Voorburg, The Netherlands.

Garfield, J., and Ben-Zvi, D. (2005), “A Framework for Teaching and Assessing Reasoning about Variability,” *Statistics Education Research Journal*, 4, 92–99. DOI:10.52041/serj.v4i1.527.

Glantz, M., Johnson, J., Macy, M., Nunez, J. J., Saidi, R., and Velez, C. (2023), “Students’ Experience and Perspective of a Data Science Program in a Two-Year College,” *Journal of Statistics and Data Science Education*, 31, 248–257. DOI:10.1080/26939169.2023.2208185.

Groth, R. E. (2014), “Using Work Samples from the National Assessment of Educational Progress (NAEP) to Design Tasks That Assess Statistical Knowledge for Teaching,” *Journal of Statistics Education*, 22, 1–28. DOI:10.1080/10691898.2014.11889712.

Hadley, K., and Oyetunji, W. (2022), “Extending the Theoretical Framework of Numeracy to Engineers,” *Journal of Engineering Education*, 111, 376–399. DOI:10.1002/jee.20453.

Haines, B. (2015), “Conceptualizing a Framework for Advanced Placement Statistics Teaching Knowledge,” *Journal of Statistics Education*, 23, 5. DOI:10.1080/10691898.2015.11889747.

Hammer, D., and Elby, A. (2003), “Tapping Epistemological Resources for Learning Physics,” *Journal of the Learning Sciences*, 12, 53–90. DOI:10.1207/S15327809JLS1201_3.

Harrell-Williams, L. M., Sorto, M. A., Pierce, R. L., Lesser, L. M., and Murphy, T. J. (2015), “Identifying Statistical Concepts Associated with High and Low Levels of Self-Efficacy to Teach Statistics in Middle Grades,” *Journal of Statistics Education*, 23, 4. DOI:10.1080/10691898.2015.11889724.

Hicks, S. C., and Irizarry, R. A. (2018), “A Guide to Teaching Data Science,” *The American Statistician*, 72, 382–391. DOI:10.1080/00031305.2017.1356747.

Hjalmarson, M. A. (2007), “Engineering Students Designing a Statistical Procedure for Quantifying Variability,” *The Journal of Mathematical Behavior*, 26, 178–188. DOI:10.1016/j.jmathb.2007.06.001.

Huang, W., London, J. S., and Perry, L. A. (2022), “Project-Based Learning Promotes Students’ Perceived Relevance in an Engineering Statistics Course: A Comparison of Learning in Synchronous and Online Learning Environments,” *Journal of Statistics and Data Science Education*, 31, 179–187. DOI:10.1080/26939169.2022.2128119.

IPEDS. (2020), IPEDS: Integrated Postsecondary Education Data System, National Center for Education Statistics.

Jacobbe, T., Case, C., Whitaker, D., and Foti, S. (2014), “Establishing the Validity of the Locus Assessments Through an Evidenced-Centered Design Approach,” 9th International Conference on Teaching Statistics, Flagstaff, AZ, July 13–18. *ICOTS9_7C2_JACOBBE.pdf*

Kahneman, D., and Tversky, A. (1972), “Subjective Probability: A Judgment of Representativeness,” *Cognitive Psychology*, 3, 430–454.

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

Landis, J. R., and Koch, G. G. (1977), “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, 33, 159. DOI:10.2307/2529310.

Madsen, B. (1975), "Strength Values for Wood and Limit States Design," *Canadian Journal of Civil Engineering*, 2, 270–279. DOI:10.1139/l75-025.

Makar, K., and Rubin, A. (2009), "A Framework for Thinking about Informal Statistical Inference," *Statistics Education Research Journal*, 8, 82–105.

Makar, K., and Rubin, A. (2018), "Learning about Statistical Inference," in *International Handbook of Research in Statistics Education, Springer International Handbooks of Education*, eds. D. Ben-Zvi, J. Garfield, and K. Makar, pp. 261–294, Cham Springer. DOI:10.1007/978-3-319-66195-7.

Mathews, D., Pleasant, M., and Clark, J. M. (2007), "Successful Students' Conceptions of Mean, Standard Deviation, and the Central Limit Theorem," 12.

McCarthy, K. A., and Kuhlemeyer, G. A. (2023), "Preparing Analytics-Enabled Professionals in Finance Using a Simultaneous Team-Teaching Approach: A Case Study," *Journal of Statistics and Data Science Education*, 32, 98–107. DOI:10.1080/26939169.2023.2197470.

Meeker, W. Q., Hahn, G. J., and Escobar, L. A. (2017), *Statistical Intervals: A Guide for Practitioners and Researchers, Wiley Series in Probability and Statistics*, Hoboken, NJ: Wiley. DOI:10.1002/9781118594841.

Miller, A., and Pyper, K. (2023), "Anxiety around Learning R in First Year Undergraduate Students: Mathematics versus Biomedical Sciences Students," *Journal of Statistics and Data Science Education*, 32, 47–53. DOI:10.1080/26939169.2023.2190010.

MMPDS-04: Metallic Materials Properties Development and Standardization (MMPDS). (2008), "Federal Aviation Administration."

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

Pfannkuch, M., Ben-Zvi, D., and Budgett, S. (2018), "Innovations in Statistical Modeling to Connect Data, Chance and Context," *ZDM*, 50, 1113–1123. DOI:10.1007/s11858-018-0989-2.

Reading, C., and Pegg, J. (1996), "Exploring Understanding of Data Reduction," in *Proceedings of the Conference of the International Group for the Psychology of Mathematics Education, International Group for the Psychology of Mathematics Education*, pp. 187–194.

Reinhart, A., Evans, C., Luby, A., Orellana, J., Meyer, M., Wieczorek, J., Elliott, P., Burckhardt, P., and Nugent, R. (2022), "Think-Aloud Interviews: A Tool for Exploring Student Statistical Reasoning," *Journal of Statistics and Data Science Education*, 30, 100–113. DOI:10.1080/26939169.2022.2063209.

Rose, T. (2015), *The End of Average: How We Succeed in a World That Values Sameness*, New York: HarperOne.

Ruff, P. E. (1984), "An Overview of the MIL-HDBK-5 Program," Battelle's Columbus Laboratories, p. 55.

Saldaña, J. (2013), *The Coding Manual for Qualitative Researchers*, Los Angeles: SAGE Publications.

Salmon, C. G., Johnson, J. E., and Malhas, F. A. (2009), *Steel Structures: Design and Behavior: Emphasizing Load and Resistance Factor Design*, Upper Saddle River, NJ: Pearson/Prentice Hall.

Schiebinger, L., Klinge, I., Paik, H. Y., de Madariaga, I. S., Nielsen, M. W., Oertelt-Prigione, S., Schraudner, M., and Stefanick, M. (n.d), "Inclusive Crash Test Dummies: Rethinking Standards and Reference Models," Gendered Innovations in Science, Health & Medicine, Engineering, and Environment, Stanford University.

Shewhart, W. A. (1931), *Economic Control of Quality of Manufactured Product*, New York: D. Van Nostrand Company, Inc.

Shtulman, A., and Prassede, C. (2012), "Cognitive Constraints on the Understanding and Acceptance of Evolution," in *Evolution Challenges: integrating Research and Practice in Teaching and Learning about Evolution*, pp. 47–65, Oxford; New York: Oxford University Press.

Stang, A. H., Greenspan, M., and Newman, S. B. (1946), "Poisson's Ratio of Some Structural Alloys for Large Strains," *Journal of Research of the National Bureau of Standards*, 37, 211. DOI:10.6028/jres.037.012.

Vo, K., Evans, A., Madan, S., and del Rosario, Z. (2023), "A Scoping Review of Engineering Textbooks to Quantify the Teaching of Uncertainty," in *ASEE Annual Conference and Exposition*.

Watkins, S. (2015), *Functional Clothing Design: From Sportswear to Spacesuits*, New York: Fairchild Books & Visuals.

Watson, J. M., Kelly, B. A., Callingham, R. A., and Shaughnessy, J. M. (2003), "The Measurement of School Students' Understanding of Statistical Variation," *International Journal of Mathematical Education in Science and Technology*, 34, 1–29. DOI:10.1080/0020739021000018791.

Wild, C. J., and Pfannkuch, M. (1999), "Statistical Thinking in Empirical Enquiry," *International Statistical Review*, 67, 223–248. DOI:10.1111/j.1751-5823.1999.tb00442.x.

Wild, C., Utts, J., and Horton, N. (2018), "What is Statistics?," in *International Handbook of Research in Statistics Education, Springer International Handbooks of Education*, eds. D. Ben-Zvi, J. Garfield, and K. Makar, pp. 5–36, Cham Springer. DOI:10.1007/978-3-319-66195-7.

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

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., and Chang, B. (2008), "What Does Research Suggest about the Teaching and Learning of Introductory Statistics at the College Level? A Review of the Literature," *Journal of Statistics Education*, 16, 8. DOI:10.1080/10691898.2008.11889566.