A Qualitative Study of Engineers' Perception of Variability as 'Error'

Emma Fox, Franklin W. Olin College of Engineering
Dr. Zachary del Rosario, Olin College of Engineering

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

A Qualitative Study of Engineers' Perception of Variability as "Error"

Abstract

Variability is an unavoidable reality: People have different heights, built parts have different dimensions, and manufactured components have different material properties. It is common in statistics to refer to certain variations as "error;" however, the term "error" has very different meanings across disciplines. This work was motivated by a concerning observation of some statisticians: a refusal to accept other, non-statistical perceptions of the term "error." As part of a larger study of practicing engineers (n=24), we used qualitative methods to investigate their interpretation of the term "error" and their ensuing approach to analyzing data. We find that the term "error' tends to erode trust in the data (11/24 participants) and can lead to a more dangerous interpretation of variability (2/24 participants). These results have important implications for communication on interdisciplinary teams and teaching statistics to engineering students.

Introduction

Variability is ubiquitous in engineering but its impact is often ignored, sometimes to dangerous effect. For example, in the 1940s the U.S. Air Force had serious issues with uncontrollable aircraft: At the height of this calamity 17 pilots crashed in a single day [1]. The standard at the time was to design aircraft for "the average man," with non-adjustable controls assuming fixed human dimensions. Gilbert Daniels [2] studied the measurements of 4063 pilots, and found that precisely zero were average. The Air Force fixed this problem by designing adjustable seats to account for the observed pilot variation [1].

While variability in human dimensions is now considered obvious and easily handled, other sources of variability are still neglected or mishandled. In aerospace engineering, enormous resources are dedicated to quantifying the variability in material strength, but other properties such as elasticity are designed using average values [3]. This treatment of variability leads to a *variance deficit* that undermines structural safety.

Statistics is considered unique as a discipline that focuses on understanding variability [4]. For instance, Makar and Rubin assert that mathematical convention inherently emphasizes certainty [5]. In contrast, variability is core to statistical thinking [6]. Thus, Statistics has useful tools and ideas to help others (including engineers) reason about variability. However, the language of statistics is different from the languages of mathematics and engineering—terms such as "error" are highly overloaded, hence interpreted differently across disciplines. Some statisticians even refuse to adapt their terminology when communicating with other specialists, an orientation we call a "non-communicative stance." This study focuses on engineers' interpretation of the term "error" in the context of data variability.

Thus, we set out to study how engineers interpret “error” in a data analysis context exhibiting variability, with a focus on the impact on design decisions. The following sections review requisite Background, Frameworks, and Methods and summarize the key Results. We conclude with a Discussion including implications for collaborators on multidisciplinary teams, and for training engineering students to interpret statistical ideas.

Background & Frameworks

In this section we review relevant definitions of the term “error” and detail our theoretical and conceptual frameworks.

Definitions of error

In mathematics, error is often defined as the accuracy of an approximation against a well-defined true value [7]. However, error and other sources of uncertainty are not a strong focus in mathematics. For instance, a recent review of mathematics in engineering-related work found only 2 out of 5466 articles that discussed “uncertainty” or “error” [8]. This view of error as “unimportant” has deep roots; Salsburg [9] describes a common practice in the 1800’s,

> One way was to keep the precise mathematical formulas and treat the deviations between the observed values and the predicted values as small, unimportant error. [12, p. 15]

Thus, it is common in mathematics to view error as negligible and unimportant. In contrast, statistics as a field of study takes variability as the core object of study [6]. Wild and Pfannkuch articulate the orientation of statisticians towards understanding variability,

> Statisticians look for sources of variability by looking for patterns and relationships between variables ("regularities"). If none are found, the best one can do is estimate the extent of variability and work around it. [6]

Frequently, unexplained variability is modeled using an *error* term. This endemic terminology is far enough from common parlance that introductory texts carefully note the meaning, for instance, this except from *Online Statistics Education*,

> It is traditional to call unexplained variance error even though there is no implication that an error was made. [10]

Other statistical texts similarly attempt to extract the “human error” from “statistical error.” For instance, the following excerpt comes from a prominent text on *Biometry*,

> It is not an error in the sense of someone having made a mistake, but in the sense of providing you with a measure of the variation you have to contend with when trying to estimate significant differences among the groups. [11]

The term error in statistics has such a specialized meaning that there are documented cases of miscommunication. In a footnote, Salsburg [12, p. 239] recounts a case where a senior executive of the U.S. Food and Drug Administration refused to allow the term “error” to appear in an official report. Instead, Salsburg consulted other colleagues for an alternative term and selected “residual” instead.

Corroborating the motivation behind this study, Salsburg admits his distaste for adjusting technical language when communicating to a non-statistical audience,

> 'How can we admit to having error in our data?' he asked, referring to the extensive efforts that had been made to be sure the clinical data were correct. I pointed out that this was the traditional name for that line. He insisted that I find some other way to describe it. He would not send a report admitting error to the FDA. … It seems that no one, in the United States at least, will admit to having error. [12, p. 239]

Note the divergence in meaning: The FDA executive is clearly highlighting the fact that “error” will connote “human error” to the intended audience, while Salsburg insists that “error” has a different statistical meaning. While the statistician in this case acquiesced to an authority, he expresses frustration that others will not acknowledge variability *in the precise statistical language of error*. This is a documented example of the “non-communicative stance” that motivates this work.

In engineering, the term “error” has yet other meanings, depending on subdiscipline. For instance, Thunnissen [12] reviews uncertainty classification systems across engineering disciplines. Civil Engineering uses multiple frameworks to categorize uncertainties; the leading framework associates the term “error” with “human error” [13].

In this work, our goal is <u>not</u> to settle a normative definition of error <u>nor</u> to assess the degree to which participants’ beliefs align with a particular definition. Rather, we are interested in the variation in participant interpretations and responses to observed “errors”—their practical response to “error.”

Theoretical framework: Knowledge-in-Pieces

To guide this work we adopt the *knowledge-in-pieces* (KiP) theoretical framework [14], [15]. KiP asserts that knowledge is not monolithic; rather, it is composed of smaller—sometimes contradictory—elements that an individual uses contextually to reason about scenarios.

These knowledge elements act by being recognized in a scenario. Thus, a person may be aware of a particular phenomenon, but a *lack* of recognition or *closer* alignment with a different knowledge element may prevent activation. For instance, a person might be aware that variability arising from manufacturing can lead to a lower-than-nominal strength, but viewing variability as unimportant "error" may convince them to neglect that variability in decision-making. Further, knowledge elements differ in their sensitivity to activation, called *cuing priority.* KiP thus guides the design of our interview protocol (we present multiple tasks with varied contexts) and our analysis of participant responses (we expect elements of the context to cue different responses).

Conceptual framework: Consequences and Reification

Here we detail the ideas that constitute our conceptual framework [16]. In short, we are interested in the *consequences* of variability [17] and the *reification* of summaries [18].

While mathematics frequently treats variability as unimportant [9] and statistics treats variability as the central object of study [6], an often neglected aspect is the *consequence* of variability. Wild and Pfannkuch [6] present a unique taxonomy of variability, decomposing variability into *real* and *induced* sources. To avoid confusion with phenomena that an engineer would consider real (e.g., induced drag or induced current), we use the terms *real* and *erroneous* variability [19].

A source of real variability is any phenomenon that can affect the quantity under study. A source of erroneous variability leads to mischaracterization of that quantity, often through measurement imperfections. Figure 1 illustrates an example in material property characterization: Imperfections in the material are a source of real variability. However, slipping of the testing apparatus leads to a mischaracterization of strength, causing erroneous variability. The real/erroneous dichotomy enables an articulation of the consequences of variation.

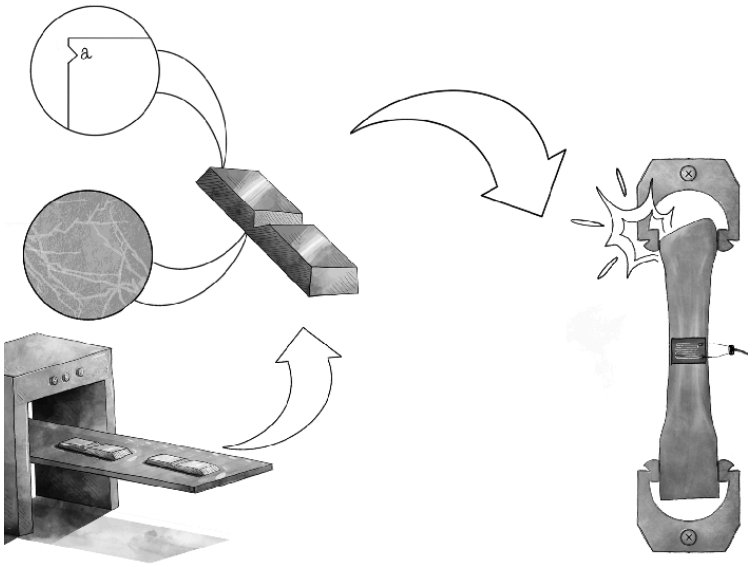

Figure 1. Examples of real and erroneous sources of variability. Imperfections in a material lead to real variability, while slippage during mechanical testing leads to erroneous variability. Image drawn by Alana Huitric.

Further complicating the interpretation of variation is the behavior of *reification*. Gould [18] defines reification as "the mental conversion of a person or abstract concept into a thing." Originally introduced in Marxist theory by Georg Lukács [20], reification describes a kind of "forgetting" where the reified interpretation precludes other interpretations. Reification is therefore considered problematic: Gould treats it as the central problem of his book *Full House,*

> This book treats the even more fundamental taxonomic issue of what we designate as a thing or an object in the first place. I will argue that we are still suffering from a legacy as old as Plato, a tendency to abstract a single ideal or average as the "essence" of a system, and to devalue or ignore variation among the individuals that constitute the full population.

Based on the definitions cited above, reification of the mean seems to be associated with treating variability as erroneous. While we have seen that "error" in statistics does not only connote erroneous variability, it is common to interpret "error" as deviations from a true value for the purposes of statistical inference. For instance, consider this passage from an introductory statistics textbook,

> Theoretically, the true score is the mean that would be approached as the number of trials increases indefinitely. An individual response time can be thought of as being composed of two parts: the true score and the error of measurement. [10]

In this example, the mean is reified as a "true value;" the goal of inference is then to reject variations from the mean. This particular reification of the mean will enter into our data analysis, detailed below.

Given this background, our research questions are:

1. How do engineers interpret variability if it is described as "error"? Do they associate "error" with real/erroneous sources, or some other meaning?
2. What—if any—effect does interpreting variability as "error" have on engineers' decision making when using data for design?

Methods

Recruitment and Data Collection

This work was completed under an exempt protocol approved by the Brandeis IRB (protocol number #22134 R-E). This investigation was part of a larger study of practicing engineers' understanding of variability [17]. Potential participants were recruited via the last author's professional network. Participants were then selected to have an engineering background, at least 2 years of professional experience, and to balance representation across race, gender, and subfield. Compared with degrees awarded in 2020 [21], our sample is relatively diverse in gender (sample Female 29% vs 2020 degree share 24%), race (sample white 33% vs 2020 degree share 56%), and nationality (including participants residing in Canada, Turkey, and the Philippines). Aligned with the goals of the larger study, participants were drawn from Aerospace, Civil, and Mechanical engineering disciplines. Demographics are summarized in Table 1.

Our sample size of n=24 is in line with recommendations for qualitative research [22], and is comparable with other peer-reviewed qualitative research projects [23], [24], [25].

Table 1. Summary of participant demographics.

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

Interviews were conducted on Zoom by the authors and four additional research assistants following a common protocol (described next). Interviews lasted between 45 to 90 minutes and

were recorded then professionally transcribed. These interviews followed a semi-structured protocol.

Interview protocol

All interviews followed a semi-structured protocol: All interviewers started from a common set of prompts, but followed-up on participant responses following ideas from *intensive interviewing* [26]. We used these follow-up questions to understand the participant's perspective, meaning, and experience. The structured portions of the interview had participants study small datasets of measured material properties and answer questions about how they would make decisions with the data. The protocol began with a short review of relevant concepts, including stress-strain curves and material property definitions. We also presented participants with an image clarifying the nature of the observed data: that values come from independent specimens rather than repeated measurements on the same specimen (see Appendix A1). This was to highlight the possibility of real variability, without directly naming that concept. The protocol then moved on to semi-structured questions about specific datasets.

The full interview considered a variety of material properties and scenarios [17]; the data analyzed in this project concerns participant responses in reaction to using a dataset of material strength values to design a simple structure. In this analyzed section, participants were first asked how they would use the data to help design a simple structure: a rod hung vertically from a fixed support. In this setting, a "normative" approach is to select a lower strength value to use in sizing the rod; in particular, federal regulations for aerospace design would *require* such an approach in order to ensure structural safety [3]. The use of a central summary (such as the mean or median) for design would result in a less-safe structure.

Participants were then asked,

> **Researcher:** In the written documentation for the data, the original collector of the data describes the observed variability as 'error'. What do you think this means?

This prompt was deliberately vague, aligned with the "non-communicative" stance described in the Introduction. Further details on the interview protocol, including earlier prompts, images & data presented to the participants, and suggested follow-up questions, are given in Appendix A1.

Open Coding

To analyze the interview data, we used two cycles of coding [27]. These included initial and focused coding to develop our initial understanding, and analytic memo writing to further develop our understanding and to produce a closed coding scheme. While we entered this study with preconceived ideas of real/erroneous variability and reification of summaries, we balanced

this prior orientation by grounding ourselves in the particulars of participant responses [26]. Through these methods, unexpected aspects of participant reasoning emerged from the data, including an interpretation of "error" as "all variability."

We chose to study two aspects of participants' reaction to variability as "error": their *Interpretation* of the meaning of the term, and their *Approach* to design. The two authors collaboratively assigned Approach codes through iterative rounds of focused coding, and developed a closed coding scheme to categorize Interpretations.

To illustrate our interview and open coding process, we present a short excerpt. The following starts from the beginning of the analyzed portion of the data. The researcher reads the formal prompt and gently guides the participant towards the intended interpretation of the question,

> **Researcher:** Let's think about the following. In the written documentation for this data, the original collector of the data describes the observed variability as error. What do you think that means?
>
> **Participant 4:** In the written documentation for that? As the original collector of the data, describes the observed variability as error. We have different tensile strengths in the difference. He described it as errors, right?
>
> **R:** Yes. Whoever collected this data, they describe this variability as being due to error.

The participant goes on to describe her interpretation of what "error" means in this context. To illustrate our open coding approach, we interleave codes in [*square brackets*] in the transcript.

> **P4:** I might think, for example, it's because he might be implying that if everything is done perfectly, that means this thing might have been smaller or vanished [*variability could be eliminated*]. It might be related to the procedures of testing [*considering the accuracy of the testing procedure*], it might be related to the variability in the specimen [*recognizing imperfect manufacturing*]. I believe that's what he's trying to hit. He's indicating maybe there should be a true value [*referencing a "true value"*], but because of certain factors, error is introduced into the collected data.

Through multiple rounds of initial and focused coding, we identified recurring and incisive codes across all 24 transcripts. These served as useful data to develop the closed coding scheme for participant Interpretations (described below).

Skipping ahead to the next scripted follow-up, we see a prompt designed to connect the participant's understanding of "error" to engineering design decisions.

> **R:** Knowing that the observed variability is error, would this change how you would use the data set to help design the rod?
>
> **P4:** Yes. Still, it's uncertainty. I have to account for it. Whatever could be a source for those variabilities or error, I still need to account them in my design.

Since the participant does not change her approach, we code this as [trust in their understanding of the data]. While the participant answers "yes" to the question above, it is evident that she plans to account for the observed variability in her analysis. Her previous approach was a conservative accounting for the observed variability: to use the 5th percentile strength of the data following Canadian Civil Engineering practice [28]. While her interpretation of the data has now been updated in response to the "error" prompt, her approach to designing using the data as input has not changed. In this sense, she trusts her understanding of the data.

Closed Coding

Above, we introduced the concepts of real and erroneous sources of variability as a means to design the interview protocol and anticipate potential responses. However, we also sought to operationalize the real/erroneous concept as a reusable closed coding scheme to describe participant Interpretations of "error." We used analytic memo writing [26] and discussion among the author team to develop the codes, indicators, and interrelations of the scheme. The first author conducted multiple readings of the data, wrote memos to develop intermediate forms of the coding scheme, and met with the second author to discuss. Once we arrived at a prototype form of the coding scheme, we split the data into randomized halves, applied the scheme independently, and returned to compare responses. We used disagreements to clarify codes and adjust the scheme, then coded the remaining data independently. We assessed interrater reliability as being substantial (Cohen's kappa=0.77, n=48) [29]. Satisfied with the reliability of the coding scheme, we resolved all remaining differences in the codes for the corpus—the finalized codes and closed coding scheme are described in the Results below.

Results

Interpretation (Closed coding results)

The closed coding scheme describes a participant's Interpretation of "error" using four independent boolean codes: **real variability**, **erroneous variability**, **human error**, and **all variability**. Figure 2 reports Interpretation (and Approach) codes for all participants. A full description of the indicators in the scheme is given in Appendix A2; as an example, we return to the excerpt from Participant 4.

> **P4:** I might think, for example, it's because he might be implying that if everything is done perfectly, that means this thing might have been smaller or vanished. It might be

> related to the procedures of testing [*mentions a protocol or procedure*], it might be related to the variability in the specimen [*variability inherent in a material*]. I believe that's what he's trying to hit. He's indicating maybe there should be a true value, but because of certain factors, error is introduced into the collected data [*machines or from measurement*].

Here we see indicators for real variability [*variability inherent in a material*], erroneous variability [*machines or from measurement*], and human error [*mentions a protocol or procedure*]. This response illustrates some of the complexities of practicing engineers' interpretation of "error" without clarification—"error" to some engineers can connote a mixture of real and erroneous sources.

Through open coding of the data, we found that the concepts of real and erroneous sources, while relevant, did not adequately capture the variation in participant Interpretations. For some participants, the concepts of real and erroneous variability did not adequately describe their Interpretation. Far more common was an interpretation of "error" involving a human.

> **P17.** I think this means that this person might need to take a statistics course or something. I'm thinking back to some real-world situations where it's not too dissimilar. They observed variability as error, error in what? I actually don't know what it means. I would assume that it means that they messed up and they don't want to redo it [*statement synonymous with "someone messed up"*]. I don't know.

While our framing of real/erroneous is focused on the consequences of variability, an Interpretation of human error is more focused on its cause, irrespective of consequences. This indicates further complexity to the interpretation of "error:" a complex mixture of consequences and causes.

Six participants interpreted "error" as encompassing all variability. For instance,

> **P13.** Let's say you have a steel rebar, you have 430 megapascals quality steel but when you do testing, … they will have dispersion and the error will mean the difference between any of the tests expected result. [*explicitly defines error as synonymous with variability*]

Participant 13 is a Civil Engineer with 5+ years of experience. Thus, he is accustomed to working with materials that are designated by a nominal strength value—here "430 megapascals quality steel." This reified value serves as a natural reference by which to judge observed values; compared with the target value of the steel designation, all other values are "error" to this

participant. Crucially, Participant 13 lists no physical reasons in his interpretation "error"—to him, all variations away from the nominal are "error."

While the "all variability" Interpretation often coincided with a lack of stated physical reasons (Fig. 2), Participant 11 offers a contrast. He listed a variety of physical mechanisms, but when asked if the "error" information would change his design process, he responded,

> **P11:** [silence] I would say no because I don't really distinguish-- maybe I'm still not really distinguishing how that error is different than what, up to this point, I had been thinking of variability.

From Figure 2, we see large variation in participant interpretations of "error." The most-frequently occurring Interpretation was human error (18/24), followed by erroneous variability (13/24). Interpretations of real variability were very rare; only 4/24 participants' responses indicated such an interpretation.

These results are aligned with the analysis of statistical interpretations of "error" discussed in the Background: An interpretation of erroneous variability occurs far more frequently than real. However, these results also underscore the difficulties of insisting on a statistical definition of "error" devoid of human influences; for our participants, "error" has a much stronger connotation of human error than any other interpretation.

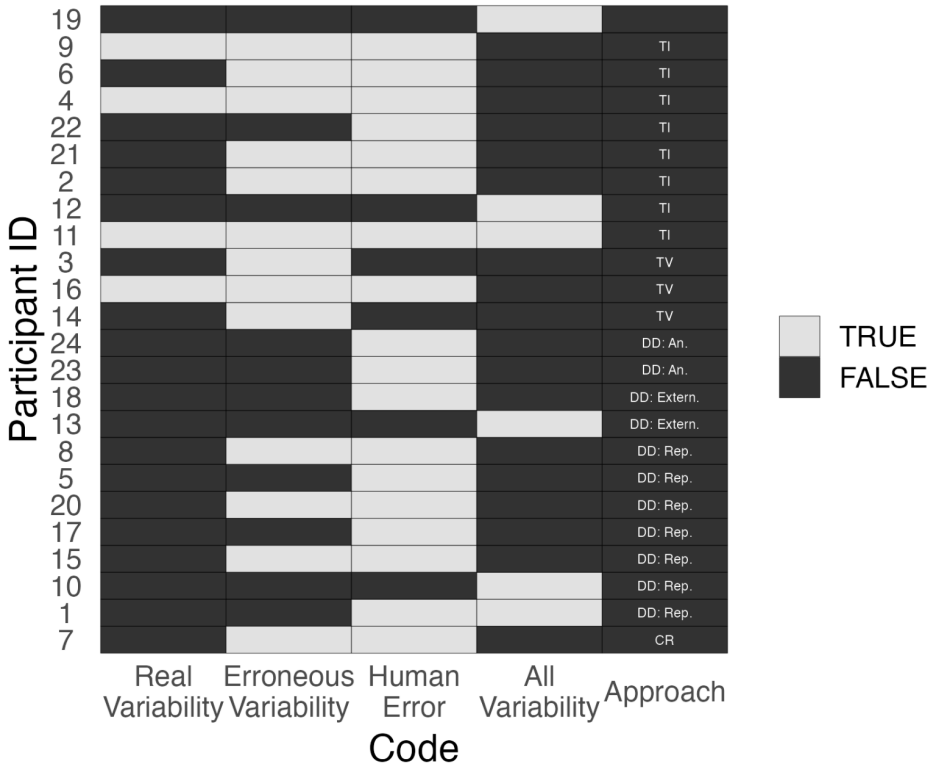

Figure 2. Finalized "error" Interpretation and Approach codes for all participants. Interpretation codes are boolean, while Approach codes are categorical. Participant 19's Approach was uncodable due to an operational issue (the relevant follow-up was not asked).

Approach (Open coding results)

The Approach codes describe the actions participants would take in response to "error." Broadly, when participants were asked if they would change their analysis in response to the "error" prompt, they tended to trust or distrust the data. The following section describes the trust-related Approach codes reported in Figure 2; the remaining distrust-related codes are described in Appendix A3.

Trust in Their Understanding of Data (TI)

The participant did not change their approach in response to the "error" prompt, representing trust in their understanding of the data. For instance, Participant 4 interprets "error" as including both real and erroneous variability. Participant 4 previously used the conservative value of 5th percentile strength to summarize the data (a standard in Canadian Civil Engineering) [28]. She describes a complementary, conservative understanding *of the data*,

> **P4:** Still, it's uncertainty. I have to account for it. Whatever could be a source for those variabilities or error, I still need to account for them in my design.

Clearly, her understanding of the data comes from a defensive stance; while she does not trust the data to provide perfect information, she trusts in her conservative *understanding* of the data. Note that participants whose Approach was coded TI varied widely in their Interpretation (and approach to analyzing the data); however, they shared a similarity in trusting their particular understanding of the data.

Only Trust a "True" Value (Internal to the Data) (TV)
The participant revised their approach to use a central value derived from the data, and treated it as a "true" value. All participants associated with this code interpreted "error" as including erroneous variability, two of whom (Participants 3 and 14) interpreted "error" as *exclusively* erroneous. For Participant 3, this justified a change from using a lower value for design to using the average.

> **P3:** Well, if the variability is considered to be error, I guess that would mean that the mean value would be seen as the true value that you could then use to design the rod and not have to worry about standard deviation since if the variability was errored, then there wouldn't really supposed to be a deviation, I suppose.

This interpretation of "error" bears a striking similarity to the "true score" interpretation of the population mean common in inferential statistics (e.g. [10]). In this episode, the term "error" cues an Interpretation of variability as erroneous and encourages reification of the mean. For Participant 3, this also reduces the cuing priority of other knowledge elements, as evidenced by her switch from a conservative analysis to using the average.

Participant 14 also initially used a conservative value. She switched to a central value (the median), but contextualizes her response to the "error" prompt,

> **P14:** It might make me slightly more likely to pick the more median value than a minimum but not hugely. … If I know that the measurement variability is error, then I know that it's unlikely that the actual yield strength of the material is on the lower end of what has been reported. Therefore the median value might be sufficient. Now, if I were actually designing this, I would never choose a factor safety so low that this mattered, and then it would become largely irrelevant.

Again we see a reified value, this time as the "actual yield strength" of the material. Participant 14 gives a statistically sophisticated differentiation between this target of inference and an estimate ("the median value might be sufficient"). However, this reified "actual" value de-prioritizes a conservative approach.

Germane to this episode (but outside the scope of this study) is the use of safety factors. Here, Participant 14 admits that the particulars of data analysis (use of the median or a conservative value) would be irrelevant when a safety factor is employed. Safety factors seem to be another phenomena that reduce the cuing priority of data analysis knowledge.

The TV code highlights a dangerous interpretation of "error:" Participants 3 and 14 initially used a conservative lower-value approach to analyze the data, but were inclined to use a central value upon interpreting the variability as "error." In the presence of a "non-communicative stance" about "error," this elevates the possibility of risk in engineering design.

Distrust of the data (DD Codes).
While 11 of 24 participants trusted the data (whether TI or TV), an equal number (11/24) distrusted the data upon hearing the "error" prompt. We describe these codes in detail in Appendix A3; in short, participants who distrusted the data either requested a repeat of the experiment (DD: Rep., 7/24), would consult a "true" value external to the provided data (DD: Extern., 2/24), or would re-analyze the data themselves (DD: An., 2/24). Among those participants who distrusted the data, all but two had an Interpretation of "error" that included human error.

While their specific approaches varied, for all 11 of these participants, hearing the variability in the data was "error" eroded their trust in the data. This presents a challenge for interdisciplinary teams that include both statisticians and engineers: While statisticians may accept "error" as a normal feature of data and analysis, insisting on using the term "error" in collaborations may introduce unnecessary communication challenges when communicating results to engineering colleagues. While a statistician may feel comfortable describing variation from an experiment as "error," eleven of our participants rejected otherwise trustworthy data when it was thought to have variability due to "error."

Discussion

This project sought to understand how practicing engineers understand the term "error" in the context of variability in data. We were particularly interested in their interpretation of the term, and their approach to data analysis and design in the face of "error." This project was inspired by observed and documented technical communication challenges: A wide diversity of meanings to the term "error," and a bias among some practitioners to *not* define or adapt their terminology in an interdisciplinary setting—what we called a "non-communicative stance".

We conducted an empirical, qualitative study of practicing engineers (n=24) using a combination of open and closed-coding methods. Our work was framed using the *knowledge-in-pieces* framework [14], [15], and was informed by concepts of reification [18], [20] and real/erroneous variability [6], [19]. We set out to answer two research questions, restated and addressed here.

RQ 1. How do engineers interpret variability if it is described as “error”? Do they associate “error” with real/erroneous sources, or some other meaning?

As is evident from the Background section, there are many different, sometimes contradictory, interpretations of “error.” Among our participants, interpretations of “error” most frequently included human error (18/24 participants). Insisting that definitions of statistical error be removed from human error, as some statisticians opine, is fraught.

Participants also much more strongly associated “error” with erroneous variability (13/24) than real variability (4/24). This is aligned with common uses of the term “error” in both mathematics and statistical inference. While “error” in statistical parlance technically includes both real and erroneous variability, it is clear that for our participants that the term has a biased interpretation. While our sample cannot support the inference that such a bias extends to the full population of engineers, it does suggest that such a bias may exist, and certainly indicates that there is the potential for a diversity of interpretations.

RQ 2. What—if any—effect does interpreting variability as “error” have on engineers’ decision making when using data for design?

Some participants, upon learning the observed variability was “error,” elected to change their approach from a conservative analysis to a more dangerous approach based on a central value (mean or median). While rare (2/24 participants), this phenomenon demonstrates the fraught nature of a “non-communicative” stance to interdisciplinary communication. Statisticians who insist on using an endemic meaning of “error” in engineering collaborations *without* clarifying terms open the door to miscommunication and an increase in potentially fatal risks.

In a less fraught outcome, many participants (11/24) after the “error” prompt began to distrust the data. Their responses ranged from requesting a repeat of the experiment, consulting external resources, or formulating a plan to re-analyze the data themselves. This highlights a practical issue communicating between statistical and engineering audiences: Statisticians accept and expect that variation will enter into data analysis, and normatively refer to certain variations as “error.” However, the term “error” may erode an engineers’ trust in a dataset.

Implications

These different interpretations of “error” encourage drastically different approaches to engineering design decisions. As shown above, linguistic differences between engineering and statistics can have potentially deadly consequences. Ideally, practitioners on interdisciplinary teams would work openly and clarify all terminology to minimize miscommunication.

As engineering educators, we can encourage a more "open stance" by exposing our students to different interpretations of terms. This can seed a more open view of terminology by showing that terms are used differently across an increasingly interdisciplinary workplace. Additionally, we can model a productive set of behaviors where collaborators ask "This is what error means to me, how do you interpret this term?" In this way, we can (hopefully) train engineers to have such discussions in their professional careers.

Limitations & Future Work

This was a qualitative (n=24) study on how engineers interpret and react to variability. While our work clearly demonstrates the variety of potential interpretations and reactions to "error," our methods are not aligned with making inferential statements about the population of engineers. For instance, we cannot conclude with certainty that engineers writ-large associate "error" with human error. While we found the closed coding scheme for interpretations of "error" to have substantial reliability, future work should test the generalizability of this scheme. Additional empirical work may also surface additional approaches and responses to "error" not seen in our sample.

Furthermore, our focus on variability necessarily limits the conclusions we can draw about engineers' perception of "error." Future work could investigate perceptions of "error" in the context of other aspects of engineering analysis. Perceptions of "error" likely interact with other elements of engineering practice; for instance, we saw that the use of safety factors reduce the cuing priority of other knowledge elements related to variability. These results suggest that studying the perception of "error" could be incorporated in a wide variety of studies.

Acknowledgements

The research assistants for this project were AJ Evans, Ellie Ramos, KD Vo, and Maeve Stites. This material is based upon work supported by the National Science Foundation under grant No. 2138463.

References

[1] T. Rose, *The End of Average: How We Succeed in a World That Values Sameness*, First Edition. New York: HarperOne, 2015.
[2] G. Daniels, "The 'Average Man'?," Air Force Aerospace Medical Research Lab, Wright-Patterson AFB OH, AD010203, 1952.
[3] Z. del Rosario, R. W. Fenrich, and G. Iaccarino, "When Are Allowables Conservative?," *AIAA J.*, vol. 59, no. 5, pp. 1760–1772, May 2021, doi: 10.2514/1.J059578.
[4] R. P. Abelson, *Statistics as Principled Argument*. Hillsdale, N.J: L. Erlbaum Associates, 1995.
[5] K. Makar and A. Rubin, "A Framework for Thinking About Informal Statistical Inference," *Stat. Educ. Res. J.*, vol. 8, no. 1, 2009.

[6] C. J. Wild and M. Pfannkuch, “Statistical Thinking in Empirical Enquiry,” *Int. Stat. Rev.*, vol. 67, no. 3, pp. 223–248, Dec. 1999, doi: 10.1111/j.1751-5823.1999.tb00442.x.
[7] G. H. Golub, J. M. Ortega, and J. M. Ortega, *Scientific computing and differential equations: an introduction to numerical methods*. Boston: Academic Press, 1992.
[8] K. Hadley and W. Oyetunji, “Extending the Theoretical Framework of Numeracy to Engineers,” *J. Eng. Educ.*, vol. 111, no. 2, pp. 376–399, Apr. 2022, doi: 10.1002/jee.20453.
[9] D. Salsburg, *The lady tasting tea: how statistics revolutionized science in the twentieth century*, First Holt paperbacks edition. New York: Henry Holt and Company, 2002.
[10] D. Lane, M. Hebl, R. Guerra, D. Osherson, and H. Zimmer, *Online Statistics Education: An Interactive Multimedia Course of Study*. Accessed: Jan. 18, 2023. [Online]. Available: https://onlinestatbook.com/
[11] R. R. Sokal and F. J. Rohlf, *Biometry: the principles and practice of statistics in biological research*, 3rd ed. New York: W.H. Freeman, 1995.
[12] D. Thunnissen, “Uncertainty Classification for the Design and Development of Complex Systems,” 2003.
[13] B. M. Ayyub, Ed., *Uncertainty modeling and analysis in civil engineering*. Boca Raton: CRC Press, 1998.
[14] A. A. diSessa, “Toward an Epistemology of Physics,” *Cogn. Instr.*, vol. 10, no. 2–3, pp. 105–225, 1993.
[15] A. diSessa, “A History of Conceptual Change Research: Threads and Fault Lines,” in *The Cambridge handbook of: The learning sciences*, Cambridge University Press, 2006, pp. 265–281.
[16] A. J. Magana, “The role of frameworks in engineering education research,” *J. Eng. Educ.*, vol. 111, no. 1, pp. 9–13, Jan. 2022, doi: 10.1002/jee.20443.
[17] Z. del Rosario, “Neglected, Acknowledged, or Targeted: A Conceptual Framing of Variability, Data Analysis, and Domain Consequences,” *J. Stat. Data Sci. Educ.*, 2024, doi: 10.1080/26939169.2024.2308119.
[18] S. J. Gould, *Full house: the spread of excellence from Plato to Darwin*, 1st ed. New York: Harmony Books, 1996.
[19] Z. del Rosario and G. Iaccarino, *Computational Modeling by Case Study: All Models Are Uncertain*. Cambridge Scholars Publishing, 2024. [Online]. Available: https://zdelrosario.github.io/uq-book-preview
[20] A. Honneth, J. Butler, R. Geuss, J. Lear, and M. Jay, *Reification: a new look at an old idea*. in The Berkeley Tanner lectures. Oxford ; New York: Oxford University Press, 2008.
[21] IPEDS, “IPEDS: Integrated Postsecondary Education Data System,” National Center for Education Statistics, 2020.
[22] J. W. Creswell, *A Concise Introduction to Mixed Methods Research*. SAGE Publications, 2014.
[23] S. A. Peters, “Robust Understanding Of Statistical Variation,” *Stat. Educ. Res. J.*, vol. 10, no. 1, pp. 52–88, 2011.
[24] A. Reinhart *et al.*, “Think-Aloud Interviews: A Tool for Exploring Student Statistical Reasoning,” *J. Stat. Data Sci. Educ.*, vol. 30, no. 2, pp. 100–113, May 2022, doi: 10.1080/26939169.2022.2063209.
[25] M. Glantz, J. Johnson, M. Macy, J. J. Nunez, R. Saidi, and C. Velez, “Students’ Experience and Perspective of a Data Science Program in a Two-Year College,” *J. Stat. Data Sci. Educ.*, pp. 1–10, Jun. 2023, doi: 10.1080/26939169.2023.2208185.
[26] K. Charmaz, *Constructing Grounded Theory*. 2014. Accessed: Jul. 20, 2022. [Online]. Available: https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5439903
[27] J. Saldaña, *The Coding Manual for Qualitative Researchers*, 2nd ed. Los Angeles: SAGE, 2013.
[28] B. Madsen, “Strength Values for Wood and Limit States Design,” *Can. J. Civ. Eng.*, vol. 2,

no. 3, pp. 270–279, Sep. 1975, doi: 10.1139/l75-025.
[29] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.
[30] P. E. Ruff, “An Overview of the MIL-HDBK-5 Program,” Battelle’s Columbus Laboratories, AFWAL-TR-84-1423, 1984.

Appendices

A1. Interview Protocol Details

Towards the beginning of the protocol, participants were presented with Figure 3. This was to clarify the context of data that was presented in the interview—presented values arise from multiple independent specimens, rather than repeated measurements on a single specimen. This was to ensure the possibility of real variability in the data, without directly naming the concept.

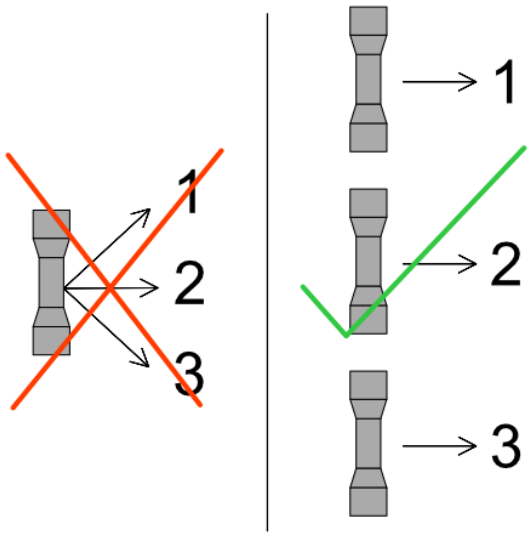

Figure 3. Image used to describe the presented data: independent specimens, rather than repeated measurements.

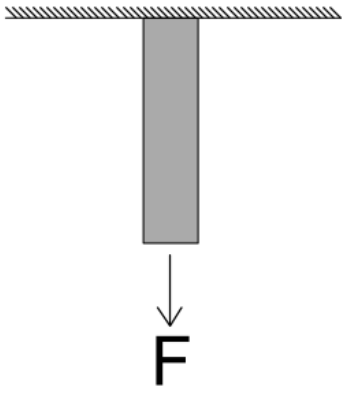

Figure 4. Image used to illustrate the design scenario. This structure was described as being in uniaxial tension.

Immediately prior to the "error" question, participants were asked to use a dataset to help design a rod. The design context of the interview task was a geometrically-simple member subject to uniaxial tension, pictured in Figure 4. The following prompt was accompanied by a dataset (Tab. 2).

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

Table 2. Dataset presented in interviews, values are the tensile yield strengths of a cast steel [30].

Steel Strength	
Sample	**Tensile Yield Strength (ksi)**
1	157.0
2	159.6
3	155.6
4	165.8

5	157.4
6	158.4
7	157.6
8	156.4
9	157.7
10	155.7

The following is the "error" prompt. Note that we deliberately withhold any definition. This question encourages participants to re-interpret the data in Table 2 in light of the "error" description. This prompt consists of an initial question about interpretation, and a follow-up about any possible changes in approach. Note that the (parenthetical) gives conditions for an optional follow-up; this is designed to help guide the participant towards the intended interpretation of the question.

"In the written documentation for the data, the original collector of the data describes the observed variability as 'error'. What do you think this means?"

- (If participant is confused by the question) "We're interested in how <u>you</u> interpret the word 'error'."

"Knowing that the observed variability is 'error', would this change how you would use the dataset to help design the rod?"

A2. Closed coding scheme: Participant Definition of "Error"

Table 3 reports the closed coding scheme developed to describe participants' Interpretation of "error."

Table 3. Closed coding scheme for Interpretation of "error."

Definition of "Error"	**Short Description**	**Long Description (with Examples)**
Real Variability	Participant's definition includes, but is not necessarily limited to, what the researchers would consider real variability.	- Definition includes variability in material properties and/or variability inherent in a material
Erroneous Variability	Participant's definition includes, but is not necessarily limited to, what the researchers would consider erroneous	- Definition includes error in machines or from measurement

	variability.	
Human Error	Participant's definition includes, but is not necessarily limited to, human errors. Human error can contribute to both real and erroneous variability.	- Explicitly defines error as "human error" - OR, mentions human operator error and/or training - OR, makes a statement synonymous with "Someone messed up" - OR, mentions a protocol or procedure - OR, uses the term "mistake"
All Variability	Participant's definition encompasses all variability.	- Explicitly defines error as synonymous with variability - OR, states that they do not understand the difference between error, variability, and/or uncertainty

A3. Open Approach codes

This section describes the Distrust of Data (DD) codes in greater detail, and details one additional open code.

Distrust Data: Re-process the data (DD: An.)

The participant distrusts the data, and would re-analyze it themselves (2/24 participants). All participants associated with this code interpreted "error" as human error only. Re-analyzing the data would allow the participant to overcome that human error; for instance, Participant 24 described this as "We neglect and avoid the error in your calculation."

Distrust Data: Consult an External "True" Value (DD: Extern.)

The participant distrusts the data and consults an external "true" value instead (2/24 participants). For instance, Participant 18 stated

> **P18:** I guess I would look at other established datasets and the typical values and not just depend on looking at the 10 values and I guess in my case I would look for more data [laughs].

The participant devalues the available data ("just depend on") and specifically refers to "established" data containing typical values. Note that in engineering, typical values are often

operationalized as the mean [3]. While this participant was previously willing to use the data for design, knowing the variability corresponds to "error" sows distrust.

Distrust Data: Repeat Experiment (DD: Rep.)

The participant distrusts the data and requests that the experiment be repeated (7/24 participants).

> **P10:** [I]f I see error, I probably want to find out why and what caused the error in the first place. It could be a lot of stuff. If this error is repeatable, for example, my error, what I'm thinking is a number that's so out of value. If that is repeatable, then yes, I probably would not use this dataset to help me design a rod.

"Repeatability" of a result requires repeated experiments; this participant would require more data collection in order to assess the "error" that appears in the original dataset.

For other participants, the term "error" suggested a paucity of data. For instance,

> **P8:** [I]f I had those ten samples, definitely affects how I used-- I might decide not to make any solid decisions on that. If I have a wider range of exact values, I can probably-- I want to assume, on average, the error will probably cancel out….

Participant 8 here uses the phrase "wider range" to refer to a larger sample of data. He outlines conditions that would allow him to trust the data: a larger sample that would reduce the "error." His Interpretation is one of human error and erroneous variability; combined with the sample size dependency noted before, this reasoning bears a strong resemblance to statistical estimation theory.

Contradictory Reasoning (CR)

Participant 7 provided a reasoning that we could not resolve and have decided to label "contradictory." He provided his interpretation of "error" as "Error for me is the variation that I can accept." Taking this statement at face value, we would not expect him to adjust his approach. However, when asked if he would change his approach in response to the "error" prompt, he stated

> **P7:** I think they consider the error before but now I think, consider about the error. We should keep more margin for this design.

Despite regarding "error" as acceptable variation, Participant 7 elected to add more design margin in response to the "error" variability. This participant's Interpretation was coded as including both erroneous variability and human error. It is possible that these two interpretations manifest as two separate knowledge elements that cue from the same term of "error." In this

case, additional contextual factors may alternatively cue the two interpretations. However, we lack the data in the present study to investigate this hypothesis.