ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #35668

A Qualitative Study of Engineering Students' Reasoning About Statistical Variability

Riya Aggarwal, Franklin W. Olin College of Engineering

Riya is senior at the Olin College of Engineering studying Engineering with a concentration in User Experience Design.

Mira Flynn, Olin College of Engineering

Mira is a 3rd year undergraduate student at Olin College. They are majoring in Engineering with a concentration in Computing.

Sam Daitzman, Olin College of Engineering

Sam Daitzman is a senior studying Engineering with a concentration in Human-Centered Product Design/Computing at Olin College of Engineering.

Dr. Diane Lam

Prof. Zachary Riggins del Rosario, Franklin W. Olin College of Engineering

Zachary del Rosario is a visiting assistant professor 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 Engineering Students' Reasoning About Statistical Variability

Introduction

Every aircraft you have ever flown on has been designed using probabilistically-flawed, potentially dangerous criteria [1]. These criteria have been in-use since at least the 1960's [2], but their limitations were only formally recognized recently. While prior work has thoroughly articulated the technical issues in these flawed design criteria [1], [3], the present work aims to support formal study of how engineers recognize and treat variability, with an eye towards understanding how the aforementioned flaws evaded notice for over a half-century.

In this work, we present a novel theoretical framework and initial empirical results. We use the proposed cause-source framework to analyze aircraft design flaws and to design an interview protocol. Through interviews with engineering students, we find initial evidence of an *induced variability bias* among participants; more specifically, we find that participants choose analysis techniques that are inconsistent with their own attribution of variability to physical mechanisms in engineering systems.

It is important to note that this paper presents not only empirical data and findings related to how engineers think about variability, but also presents a theoretical framework that can support measurement and analysis of how engineers consider variability.

Background and Related Work

The Allowables Issue. Previous research in aerospace engineering describes the in-use "allowables" design criteria, while more recent works critique these criteria. Safety-critical material properties in aircraft design (such as the strength of material) are quantified using a single-value allowed in design, appropriately called design allowables [1], [4], [5]. Some statistical treatment of material properties is necessary as all manufactured components exhibit unavoidable variability in their properties [6]. Different material properties are treated with different classes of allowable value; for instance, strength values are treated with a conservative value called a basis value (operationalized as a lower tolerance interval [7]), while modulus properties such as elasticity are treated with a typical value (operationalized as a sample mean value). However, modern materials such as advanced composites exhibit considerable variability in their matrix properties. Treating these properties with typical values leads to a variance deficit that can potentially result in dangerously-undersized structural components [1].

Note that rigorous procedures that comprehensively solve these issues are available [8]; these techniques lie within a literature of analysis and design under uncertainty that spans multiple disciplines. However, it is not common to see these rigorous treatments of uncertainty in industrial use, and even within aerospace engineering the inconsistent treatment of variability is not widely recognized; for instance, a NASA reference on probability and statistics [9] mentions "basis values" but makes no mention of "typical values" whatsoever. To the best of our knowledge, there is very little work studying engineering knowledge and treatment of variability.

Related Statistics Education Literature. A great body of work studying variability belongs to statistics education, as variability is core to statistical thinking [10]. Literature in the early 2000s characterized the research on students' understanding of variability as limited [11]. Furthermore, research on how engineers apply statistical techniques to a dataset, and especially how they make design decisions, is still lacking. An exception, and of particular interest to the present work, is a study of Hjalmarson [12], who developed a data-analysis task to study how engineering students use statistics to make operations decisions. Hjalmarson noted "[a]lthough the students successfully computed statistics that would measure variability, the justification for the use of particular statistics was sometimes lacking or the students' strategy seemed to be to compute everything they knew for the sake of computation, without considering what might be applicable or the relationship between statistical measures." These results suggest a disconnect between the use of statistical measures (analysis procedure) and the underlying problem features (attribution), an observation that informs the design of the present study.

Engineering as a discipline is primarily interested in statistical thinking insofar as it supports design and control. However, past work in statistics education has used theoretical frameworks that give primacy to concepts that do not comprehensively treat engineering concerns; for instance the framework of Peters [13] considers statistical experimental design, data analysis, and statistical modeling. Missing from this extant framework are the means to determine when an engineer should or should not treat variability as random, and how to link observed variability to engineering outcomes of interest.

Thus, for our work we found it necessary to develop a theoretical framework that would better characterize variability and operationalize how one might study it in engineering contexts. We began by drawing on both statistics education literature and classical work on engineering quality. The work of quality experts such as Shewhart [6] and Deming [14] focused on recognizing variability as either under statistical control or due to a traceable cause. These ideas are termed *chance* and *assignable cause* [6]. Additionally, the work of statistics educators Wild and Pfannkuch [10] distinguishes between *real* and *induced* variability. We term this dichotomy *real* and *induced source*, and further develop the concept to support the analysis of engineering systems. We elaborate on these below.

Given this background, our research questions (RQ) for this study were:

- (RQ 1) What theoretical framework can support the analysis of variability in engineering, specifically the aircraft allowables criteria?
- (RQ 2) What factors related to variability can explain the design and use of faulty aerospace allowables, particularly their going undetected for a half-century?
- (RQ 3) With regards to engineering education, do any factors encourage engineers to use a more problem-relevant statistical analysis?

Theoretical Work

Proposed Theoretical Framework

The theoretical framework for this study builds upon classic ideas in the engineering quality and statistics education communities. Figure 1 illustrates the core of this framework in the *cause-source variability quadrants*. These quadrants are formed by the intersections of independent dichotomies: the cause and source dimensions.

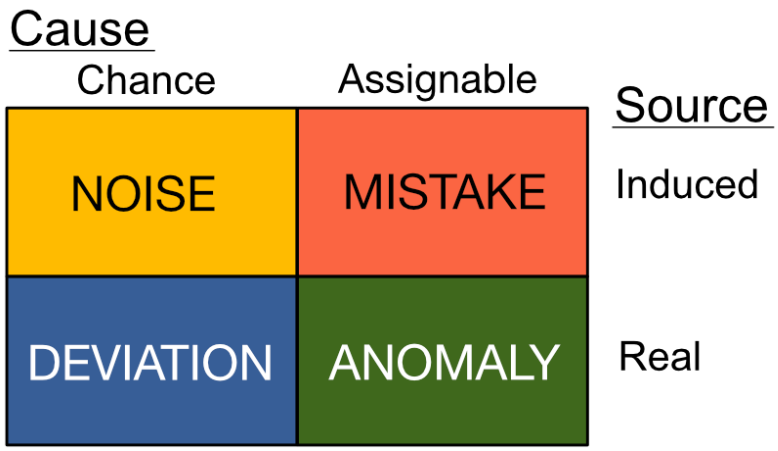

Figure 1. The Cause-Source variability quadrants.

The Cause Dimension. The concept of *chance* and *assignable cause* was formulated in the engineering quality community to support the study and reduction of variability in manufacturing [6], [14]. Using definitions from Reference [15], we define *assignable cause* as any form of variability that is practical to describe and control / represent deterministically, while *chance cause* is any form of variability that is impractical to describe deterministically and is best-described with a random variable. For example, in aerospace design it is impractical to completely eliminate cracks in structural components. Thus, aerospace engineers treat the size of cracks present in a component as a random process---a chance cause [16]. However, if consistent issues arise in a particular manufacturing line, then manufacturing engineers will investigate to identify the issue---perhaps they will find a failure to follow experimental procedure [17]---and

work to eliminate the assignable cause. The concept of cause provides practical guidance on when to treat variability as random, and when to invest resources to investigate further.

The Source Dimension. The concept of *real* and *induced source* appears in the statistics education work of Wild and Pfannkuch [10], [18], but it is presented as self-evident and without a precise definition. To help clarify our use of these terms, we adopt the definitions of Reference [15]: The *scopus* is the real value that one aims to study, while the *measurement* is the value one manages to record. Thus, variability is said to be *real* if it affects the scopus, and *induced* if it affects the measurement only. Figure 2 illustrates this definition schematically, where a source of real chance variability (called *deviation*) first generates the scopus, and an additional source of induced chance variability (called *noise*) corrupts the scopus to form the measurement.

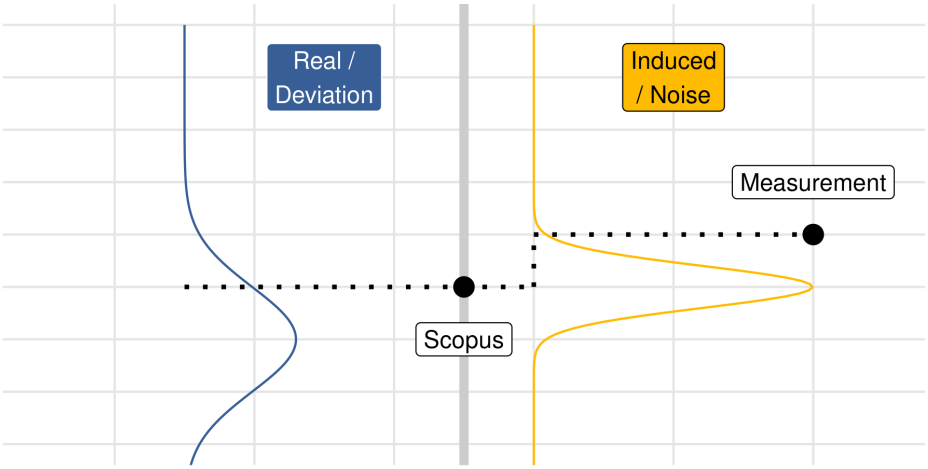

Figure 2. Schematic depiction of the scopus and measurement, relating real and induced variability that generate and corrupt the scopus (respectively).

For example: All materials exhibit unavoidable variability in their properties. The appropriate scopus is not the "mean" property, but rather the as-built properties that will determine as-built performance. Material variability can arise due to small fluctuations in the manufacturing process, generating real variability in their mechanical properties: a source of deviation. However, measurement techniques themselves are imperfect, generating induced variability in the characterization of those properties: a source of noise. Statisticians characterize multiple sources of variability by preparing multiple specimens and taking multiple measurements of each specimen: a nested experimental design [17].

Aspects of Reasoning. In addition to theorizing how variability might be operationalized in the engineering context, we also study two aspects of how participants consider variability: analysis and attribution. We define a participant's *analysis* as the mathematical technique they use to study a scenario, and their *attribution* as the physical mechanisms they use to explain the existence of variability. Analysis and attribution can be real or induced, and need not be in agreement, as we will see below.

Study of Allowables by Theoretical Framework

At this point, we revisit the allowables issue [1] in light of the theoretical framework, both to illustrate the utility of the theoretical framework and to address RQ 1: finding a theoretical framework to understand the allowables issue. First, note that materials characterization and manufacturing follows a highly-controlled process that seeks and eliminates assignable causes [4], [19]; thus, in design it is assumed that variability is largely chance cause (Fig. 1, Noise and Deviation quadrants).

A defensible scopus for aerospace manufacturing is the realized[1] material property of each manufactured component: While typical or nominal material properties may have utility in describing or selecting materials, the realized properties are what dictate probabilities of failure and hence safety. Fundamentally, there exists real, unavoidable variability in realized material properties, as evidenced by prior studies on manufacturing [6], [14], [20]. Aerospace engineers use basis values (conservative values) to quantify this variability when treating a material's strength. However, aerospace engineers use typical values (the sample mean) when treating properties such as the elastic modulus [21]. This typical statistic is understood to be a low-reliability quantity, as it is referred to as having "a statistical connotation of 50% reliability with a 50% confidence" [19]. Put differently, the sample mean does not acknowledge variability. Given the dependence of the buckling failure mode on the elastic modulus [22], it is obvious that the use of such a low-reliability allowable would translate to a low-reliability structure.

[1] Note that *realization* is a statistical term that connotes an observed value of a random variable; a realized material property is the as-manufactured performance of a part, rather than the nominal performance. Allowables are used in part because realized strength values can be less than their nominal values, which has important ramifications for structural safety.

However, the use of the sample mean would be justified if the underlying material variability were induced only. In this (counterfactual) case, there would exist a "true" material property, and standard statistical assumptions (zero-mean additive noise) would endorse the sample mean as an efficient estimate of this posited true value [23], [24]. Thus, we associate a conservative statistic (such as a tolerance interval [7], or even a minimum value) with an analysis assuming real variability, and a central statistic (such as the mean or median) with an analysis assuming induced variability only. Table 1 maps the two forms of aircraft allowable to the cause-source variability quadrants.

Table 1. Aircraft allowables mapped to the cause-source quadrants. Note that a typical value is inappropriate for treating a source of deviation, while a basis value would be overly-conservative for a source of noise.

Chance Cause	**Assignable Cause**	
Noise (Typical value appropriate)	Mistake	**Induced Source**
Deviation (Basis value appropriate)	Anomaly	**Real Source**

Thus, we can summarize the contradiction in aerospace engineers' behavior with allowables in terms of the theoretical framework: Aerospace engineers use quality control techniques to eliminate assignable causes, endorsing a statistical approach to analyzing the remaining chance causes. However, aerospace engineers select specific analysis techniques that suggest the source of variability is real for strength properties, but induced for engineering moduli. The assumption of induced variability is counterfactual to the manufacturing reality, thus the use of typical values is potentially dangerous.

Note that through studying published literature we have limited means to study engineers' attribution of variability. Thus, we have designed and conducted interviews with engineering students to study both their analysis and attribution of variability. Further, since only the source axis is operative in the allowables scenario, we designed our interview protocol to emphasize chance causes while probing differences between real and induced sources.

Hypotheses

In aircraft design, only particular material properties are treated with basis values; it is the default to treat material properties as a typical value [1]. Given the analysis above (Tab. 1), this suggests that aerospace engineers treat variability as induced by default. Further, some features of the basis value-treated properties (strength values) seem to encourage engineers to treat the variability as a real source. To support RQ 2, we formalized these observations as hypotheses and set out to test the following:

1. *Induced variability bias*: In the absence of provoking problem features, engineering students will tend to reason (use analyses and give attributions) more aligned with induced than real variability.
2. *Real provocations*: Problem features such as safety-critical applications or obviously questionable assumptions will tend to encourage engineering students to reason in terms of real variability at a higher rate (compared to the base-rate implied by hypothesis 1).

Empirical Methods

In this section we describe the sample of data collected, the interview protocol, and the qualitative data coding scheme. This work was determined to be IRB Exempt by the Brandeis IRB under protocol number #21164R-E.

Sample

Seven participants were recruited from an Engineering college in the Northeast United States for hour-long interviews on Zoom. They included both students and recent graduates of the college and were required to have coursework in at least one of material science, analysis of structures, or college-level statistics. These selection criteria helped ensure participants could interpret the interview questions.

Interview Protocol

The interview was structured in three primary parts: a brief warmup to acclimate the participant to the interview format, a structured part focused on analyzing materials data, and an unstructured part where participants defined and analyzed their own scenario. The present work considers results from the structured part only.

The interview's structured portion included four questions, summarized in the bullets below. To assess the *real provocations* hypothesis, these questions were designed to start with little context, and iteratively introduce more "provoking" problem features. The bullets below summarize the four questions, but further detail is provided in the Appendix.

1. **Aluminum:** No special problem features, considers material elasticity
2. **Steel:** Emphasizes safety-critical design task, considers material strength
3. **Compare:** Re-casts Aluminum question as a safety-critical buckling design task
4. **Critique:** Introduces questionable "true" value assumption to the Steel task

Since we were interested in the analysis procedure that participants would select, interview questions do not specify statistical methods, cf. [12]. Interviews were transcribed, providing the corpus for data coding, described next.

Coding Scheme

Transcripts were analyzed and coded at the grain-size of individual questions: Participants' responses were coded in terms of their complete response to each of the four questions. For each question, coding consisted of determining the absence or presence of indicators of real/induced analysis or attribution. This resulted in a set of 4 (=2x2) boolean values for each question, corresponding to the four codes: real analysis, induced analysis, real attribution, induced attribution. Across four questions, this yields 16 codes per participant. Note that, since participants could provide multiple answers in responding to a single question, many transcripts were found to exhibit indicators for multiple codes; for instance, both real and induced analyses.

In line with the Study of Allowables (Tab. 1), we took the use of a statistic measuring central tendency (e.g. mean, median) as an indicator of an induced analysis, while approaches that acknowledged spread (e.g. standard deviation) or targeted extreme values (e.g. a quantile) as an indicator of a real analysis. Reasons for variability that corresponded to the data collection mechanism indicated induced attribution, while reasons that were inherent to the object of study indicated real attribution. Table 2 below summarizes the indicators for the coding scheme.

Table 2. Indicators used for coding scheme.

		Indicators
Analysis	**Real**	Use of non-central quantity (e.g. quantile, minimum), or use of measure of spread (e.g. variance)
	Induced	Use of measure of central tendency (e.g. the mean)
Attribution	**Real**	Variability attributed to phenomena occurring inside the part / material
	Induced	Variability attributed to phenomena occurring during material characterization

Example coding. To illustrate the coding scheme, we provide excerpts from a transcript and the resulting codes. For instance, in response to the Aluminum question, one participant responded:

> "I would take a…I would take an average first, of the different data points you've got there because it looks like they all fall within the same range of about 10,00 to 10,700 ksi. I'd start by taking a statistical mean and….for my sake, I would, I would compare it to the elasticity of other materials, just so that I get some sense of how it compares to other known materials."

Since this participant took an average but did not mention any other statistics, their transcript was coded as {induced analysis: true; real analysis: false}. However, for the Steel question, the same participant stated,

> "If, for some reason the material had to be used anyway, like if this were some experimental material that has a wide variability in its yield strength, but is super light or something like that, then I would probably look at the, at the lowest yield strength that was available, and probably build in some margin there to go "well this is on the low end, but also I don't know how low the range could fall, but I do not have enough engineering experience to know how to, how to draw that lower, that lower baseline. So I'd have to stop there."

For the Steel question, this participant did not mention the mean, and instead chose the smallest observed value. Thus, their transcript was coded as {induced analysis: false; real analysis: true}.

Returning to the Aluminum question for the same participant, when asked in a follow-up what might be reasons that the table values were not all the same, the participant stated,

> "Okay – so – if – so that machine, well, you could have – well one obvious source of error would be if the machine is not calibrated correctly. … But I wonder how they were stored? I wonder if there were – if the storage conditions – like if some were very cold, some were very warm, what effect that would have. I don't know how big of a difference the environment itself would make. I imagine if the environment weren't at a constant temperature that... that's something that could affect the modulus of elasticity."

For the Aluminum question, this participant noted reasons for variability including measurement ("if the machine is not calibrated correctly") and intrinsic properties of the aluminum specimen ("that's something that could affect the modulus of elasticity"). Thus, their transcript was coded as {induced attribution: true, real attribution: true}.

To assess the reliability of the coding scheme, two analysts independently coded three transcripts and their results were compared. The resulting codes had a Cohen's kappa of 0.71, indicating substantial agreement across the subset [25]. The analysts came together to reconcile differences on this subset of the data, and one analyst coded the remaining interviews: These are the codes used in the Results section below.

Results

The overall results, summarized in Table 3, are in agreement with our hypotheses: Participants start with a bias towards induced variability (*induced variability bias*), as participants gave analyses and attributions at a higher count for induced (12) than real (11) for the Aluminum

question. This result suggests a very modest *induced variability bias*; however, we will see below that disaggregating analysis and attribution provides a more nuanced view. Participants tend to reason more in terms of real variability when exposed to problem features such as design criticality (as with the Steel question, 7 induced to 12 real), or when faced with questionable assumptions (as with the Critique question, 5 induced to 12 real).

Interestingly, the Compare question does not provoke an increased real response: The Compare question has 7 real responses, as compared with 12 real responses for the Steel question. This suggests that participants do not recognize the design-critical issues in the buckling design problem. It is possible that the causal pathway from "variability in elasticity" to "variability in buckling strength" to "variability in safety" was not so obvious as "variability in tensile strength" to "variability in safety." This has ramifications for the design of engineering pedagogy, which we return to in the Discussion.

Table 3. Counts of induced and real responses for all n=7 participants, out of 14 possible counts.

Question	**Induced**	**Real**
1. Aluminum	12	11
2. Steel	7	12
3. Compare	8	7
4. Critique	5	12

Disaggregating by analysis and attribution (Table 4) provides a richer view of participant reasoning in the Aluminum question. Note that all participants (7/7) provide at least one real attribution for variability in the first question, and yet only a small majority (4/7) deploy a real analysis technique on the data. This suggests a disconnect between analysis and attribution among participants. The striking decrease in induced analysis from the Aluminum (6/7) to the Steel question (3/7) indicates that participants are able to choose different data analysis approaches when subject to *real provocations*. Thus, we see evidence that an *induced variability bias* may indeed exist among engineering students; however, this bias seems to manifest in terms of analysis technique, and may be overturned by *real provocations*.

Table 4. Counts of induced and real responses for all n=7 participants, disaggregated by analysis and attribution. Out of 7 total possible counts.

Question	**Induced Analysis**	**Real Analysis**	**Induced Attribution**	**Real Attribution**
1. Aluminum	6	4	6	7
2. Steel	3	5	4	7
3. Compare	6	5	2	2
4. Critique	2	7	3	5

Discussion

Our results offer a potential explanation for how the allowables design error related to variability went undetected in aerospace practice for so many decades (RQ 2). The *induced variability bias* and *real provocations* concepts can explain the design decisions: If engineers tend to analyze variability as induced, this can explain why most material properties are analyzed using a sample mean in aerospace design. Further, the obvious failure connotations of material strength serve as a *real provocation*, encouraging engineers to use statistical procedures that treat the variability as real, providing an answer to RQ 3. Finally, the disconnect between analysis and attribution suggests a mechanism for undetected errors: Without the means to critique self-inconsistent statistical procedures, the errors in the design allowables approach could easily go undetected.

Surprisingly, we found that participants were adept at attributing variability to real physical mechanisms, but did tend to exhibit an *induced variability bias* in their analysis procedures. Given the sample limitations (below), it is difficult to tease apart the factors that may lead to this disconnect. Regardless of the underlying cognitive mechanism, this is an important discrepancy to understand and address in future engineering pedagogical work.

Limitations

Given the novelty of this work and the small sample size (n=7), the results above should not be taken as accurate estimates for population inference: However, the results above are useful for comparisons across questions / code types.

Attributing sources of variability for the interview questions relies on a background in structures/materials, while selecting an analysis procedure relies on statistical reasoning. Requiring participants to have relevant background in both areas would have severely restricted participant recruitment for this study. However, since participants were only required to have background in one of the required subject areas, we cannot exclude the possibility that some of the patterns above are due to participants' unfamiliarity in one of the topic areas. This is

especially salient for the Analysis category, as the mean is known to be a more accessible procedure than e.g. the standard deviation [26]. This particular limitation complicates the interpretation of the observed *induced variability bias*; this could be due to unfamiliarity with other possible statistics, a failure to integrate statistical reasoning with knowledge of materials, or other factors not related to participant educational background. However, given the generally-acknowledged poor state of statistics education [27], we believe this is a limitation of interpretation, not in the results themselves. The existence of an *induced variability bias* is problematic, regardless of its ultimate cause.

Finally, we studied engineering students in order to explain the behavior of practicing engineers. While the students do have training and experience relevant to aerospace structural design, none are full-time aerospace engineers. This extrapolation clearly limits the degree of confidence we can have in the external validity of our findings. While practicing engineers begin their careers as engineering students, practicing engineers also experience further skills development and professional enculturation that will tend to modify their behavior and cognition. Thus, we must regard our explanation of the historical allowables record as tentative.

Future work

The *real provocations* concept is a useful, potentially-generalizable mechanism to help engineers deploy analysis procedures appropriate for real variability. A potential educational intervention could aim to help engineering students see the potential safety hazards in more subtle cases of variability propagation; for instance, the lack of a *real provocation* response to the Compare question described above suggests that students need help reasoning through how variability in elasticity propagates to variability in buckling safety. Explicit instruction in variability (and uncertainty) propagation may help engineers leverage the *real provocations* mechanism, and expand their engineering design toolkit.

The present study was inspired by questions of structural design in aerospace engineering. However, variability is omnipresent in engineering. We hope that this study inspires investigators from other disciplines to consider: What sources of variability are important in their own discipline? Are those sources real or induced? What sorts of attributions and analyses do engineers in their discipline deploy? Does an *induced variability bias* exist in their own field? And what kind of *real provocations* can educators highlight to encourage a more statistically-grounded treatment of variability?

References

[1] 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.

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

[3] Z. del Rosario, “Precision Margin: First-Principles Margins for Aircraft Design Under Uncertainty,” Stanford University, 2020.
[4] *MMPDS-04: Metallic Materials Properties Development and Standardization (MMPDS)*, vol. 4. Federal Aviation Administration, 2008.
[5] *Composite Materials Handbook*, vol. 3: Polymer Matrix Composites Materials Usage, Design, and Analysis. Warrendale, PA: SAE International, 2012.
[6] W. A. Shewhart, *Economic control of quality of manufactured product*. D. Van Nostrand Company, Inc., 1931.
[7] W. Q. Meeker, G. J. Hahn, and L. A. Escobar, *Statistical Intervals: A Guide for Practitioners and Researchers*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2017. doi: 10.1002/9781118594841.
[8] Z. del Rosario, G. Iaccarino, and R. W. Fenrich, “Fast Precision Margin with the First-Order Reliability Method,” *AIAA J.*, vol. 57, no. 11, pp. 5042–5053, Nov. 2019, doi: 10.2514/1.J058345.
[9] M. H. Rheinfurth and L. W. Howell, “Probability and statistics in aerospace engineering.” NASA Marshall Space Flight Center, 1998.
[10] 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.
[11] J. M. Watson, B. A. Kelly, R. A. Callingham, and J. M. Shaughnessy, “The measurement of school students’ understanding of statistical variation,” *Int. J. Math. Educ. Sci. Technol.*, vol. 34, no. 1, pp. 1–29, Jan. 2003, doi: 10.1080/0020739021000018791.
[12] M. A. Hjalmarson, “Engineering students designing a statistical procedure for quantifying variability,” *J. Math. Behav.*, vol. 26, no. 2, pp. 178–188, Jan. 2007, doi: 10.1016/j.jmathb.2007.06.001.
[13] S. A. Peters, “Robust Understanding Of Statistical Variation,” *Stat. Educ. Res. J.*, vol. 10, no. 1, pp. 52–88, 2011.
[14] E. Deming, “Quality, productivity, and competitive position,” 1991.
[15] Z. del Rosario and G. Iaccarino, *All Models are Uncertain: Case Studies with a Python Grammar of Model Analysis*. Cambridge Scholar Press, Forthcoming. [Online]. Available: https://zdelrosario.github.io/uq-book-preview
[16] T. L. Anderson, *Fracture mechanics: fundamentals and applications*, Fourth edition. Boca Raton: CRC Press/Taylor & Francis, 2017.
[17] G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for experimenters: an introduction to design, data analysis, and model building*. New York: Wiley, 1978.
[18] C. Wild, “THE CONCEPT OF DISTRIBUTION,” *Stat. Educ. Res. J.*, p. 17, 2006.
[19] *Composite Materials Handbook*, vol. 1: Guidelines for Characterization of Structural Materials. Warrendale, PA: SAE International, 2012.
[20] W. A. Shewhart and W. E. Deming, *Statistical method from the viewpoint of quality control*. New York: Dover, 1986.
[21] S. D. Sheppard and B. H. Tongue, *Statics: analysis and design of systems in equilibrium*, Rev. printing. Hoboken, N.J.: Wiley, 2007.
[22] D. J. Peery, *Aircraft structures*. Mineola, N.Y: Dover Publications, 2011.
[23] S. Weisberg, *Applied linear regression*, Fourth edition. Hoboken, NJ: Wiley, 2014.
[24] R. Kenett and S. Zacks, *Modern industrial statistics: design and control of quality and reliability*. Pacific Grove: Duxbury Press, 1998.
[25] 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.
[26] D. Mathews, M. Pleasant, and J. M. Clark, "Successful Students' Conceptions of Mean, Standard Deviation, and The Central Limit Theorem," p. 12, 2007.
[27] R. V. Hogg, "Statistical Education: Improvements are Badly Needed," *Am. Stat.*, vol. 45, no. 4, pp. 342–343, Nov. 1991, doi: 10.1080/00031305.1991.10475832.
[28] A. H. Stang, M. Greenspan, and S. B. Newman, "Poisson's ratio of some structural alloys for large strains," *J. Res. Natl. Bur. Stand.*, vol. 37, no. 4, p. 211, Oct. 1946, doi: 10.6028/jres.037.012.
[29] P. E. Ruff, "AN OVERVIEW OF THE MIL-HDBK-5 PROGRAM," Battelle's Columbus Laboratories, AFWAL-TR-84-1423, 1984.

Appendix

Interview Questions

The following is a simplified listing of the interview questions. Data tables used in the study are given below.

(1. Aluminum) "Look at this table of material property data. These are the measured elasticity values for a rolled aluminum alloy of the same composition and processing method. Reminder, elasticity is a property of a material that determines how stiff a part is, so a material with higher value of elasticity is more stiff. How would you use this data to describe the elasticity of this alloy?"

"The measured elasticity values are not all the same; what are some reasons why that might be?"

(2. Steel) "Look at this table of material property data. These are the measured yield strength values for a cast stainless steel of the same composition and processing method. Imagine you were going to design a safety-critical structural component, loaded in tension, using this cast steel. How would you use this data to help design that component?"
"The measured elasticity values are not all the same; what are some reasons why that might be?"

(3. Compare) "Think back on your approach to the aluminum elasticity table: How did your approach there compare with your approach to the steel strength scenario?"

"Now imagine you were going to design a safety-critical structural member, loaded in compression, using the aluminum alloy from before. Would you use a different approach to process the aluminum data?"

(4. Critique) "Suppose a colleague of yours analyzes the Steel strength data, and plans to use the data to design a safety-critical part. He tells you 'The smallest value we saw was 155.6 ksi, so the true strength is probably around 155 ksi.' What do you think about his analysis?"

Data Tables

Data used in the Aluminum question [28].

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

Data used in the Steel question [29].

Steel Strength	
Observation	**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