SANDIA REPORT

SAND2017-1263
Unlimited Release
Printed February 2017

Analyst-to-Analyst Variability in Simulation-Based Prediction

Matthew R. Glickman
Vicente J. Romero

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: http://www.ntis.gov/search

SAND2017-1263
Unlimited Release
Printed February 2017

Analyst-to-Analyst Variability in Simulation-Based Prediction

Matthew R. Glickman and Vicente J. Romero
Cognitive Science and Systems and
V&V, UQ, and Credibility Processes
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-MS1327

Abstract

This report describes findings from the culminating experiment of the LDRD project entitled, "Analyst-to-Analyst Variability in Simulation-Based Prediction". For this experiment, volunteer participants solving a given test problem in engineering and statistics were interviewed at different points in their solution process. These interviews are used to trace differing solutions to differing solution processes, and differing processes to differences in reasoning, assumptions, and judgments.

The issue that the experiment was designed to illuminate—our paucity of understanding of the ways in which humans themselves have an impact on predictions derived from complex computational simulations—is a challenging and open one. Although solution of the test problem by analyst participants in this experiment has taken much more time than originally anticipated, and is continuing past the end of this LDRD, this project has provided a rare opportunity to explore analyst-to-analyst variability in significant depth, from which we derive evidence-based insights to guide further explorations in this important area.

ACKNOWLEDGMENTS

Principal thanks to our analyst participants for their patience, openness, and insightful reflections.

Errors, omissions, and all other shortcomings are entirely due to the PI (Glickman).

CONTENTS

FIGURES

NOMENCLATURE

ASC	Advanced Simulation and Computing
HSB	Human Subjects Board
LDRD	Laboratory Directed Research & Development
MCMC	Markov-Chain Monte Carlo
NNSA	National Nuclear Security Administration
PCMM	Predictive Capability Maturity Model
SNL	Sandia National Laboratories
UQ	Uncertainty quantification
V&V	Verification and validation

1. INTRODUCTION

This report focuses on the culminating activity of the LDRD project entitled "Analyst-to-Analyst Variability in Simulation-Based Prediction": A case study of multiple analysts working on solutions to a challenging test problem that involves making predictions and quantifying uncertainty with a computational model in an engineering domain.

1.1 Motivation

By definition, all models are abstractions of some sort, and hence differ in some way from the actual phenomena they are taken to reflect. As the statistician G. E. P. Box famously said, "All models are wrong. Some are useful."[1]

Accordingly, when high consequence decisions are to be made on the basis of predictions made using models, it is critical to understand precisely *how* the models in question are wrong and *to what degree*. Unfortunately, doing so is far from straightforward, particularly given models of great complexity such as the computational simulations of physical phenomena developed via the NNSA's Advanced Simulation and Computing (ASC) program. In recognition of this challenge, the ASCI program (ASC's predecessor) established a Verification and Validation (V&V) program in 1999, and Sandia National Laboratories (SNL) has at least one department (as well as many personnel in other departments) devoted to developing and disseminating expertise in this area.

An inventory of the dimensions one must consider when evaluating a complex, predictive model is specified as part of the Predictive Capability Maturity Model (PCMM) [2], and includes: representation and geometric fidelity, physics and material model fidelity, code verification, solution verification, model validation, and uncertainty quantification and sensitivity analysis. Nonetheless, in contrast to the rich body of knowledge that's been developed within each of these territories, there remains a critical link in the pipeline leading to simulation-based predictions that has barely been subject to formal study: that of human analytic judgment.

Although the potential significance of human judgment for simulation-based prediction is largely recognized among computational modelers, the most concrete evidence of this phenomenon comes from community responses to various V&V "challenge problems", some key instances of which ([3], [4]) have been sponsored by Sandia itself. In general, experts proposing solutions to these challenge problems tend to arrive at significantly different predictions, and even more alarmingly, often provide estimates of certainty that exclude the estimates derived by other analysts.

The scant research investigating the impact of human judgment on analyses based upon computational simulation includes a journal paper from 1993 [5] and a presentation given at the 2014 American Society of Mechanical Engineers' Verification and Validation Symposium [6]. The objective of the study described here has been to permit us to trace any analyst-to-analyst variability that results from independent solution of a common challenge problem to differences in methods and processes, and to in turn trace differences in methods and processes to

differences in reasoning, beliefs, and assumptions. Ideally, this characterization can then serve as the basis for theory to guide further research in this area.

1.2 Organization

Section 2 describes the experiment conducted, including the essential concept of the experiment, constraints that affected how it was ultimately realized, the test problem that was devised and used for the experiment, and the individual interviews conducted with the analyst participants. Section 3 presents the results of the experiment, focusing primarily on summarizing the information from the interviews that bears upon the individual differences that led to different choices in analysis, but also discusses analyses of the problem conducted to date and presents some of the proposed solutions, all or partial. Section 4 discusses the results and suggests some initial conclusions.

2. EXPERIMENTAL DESIGN

2.1 Concept

Discussions and background research conducted in preparation for and during this project consistently supported that:

- Analyst-to-analyst variability (in one form or another) is a broad, pervasive phenomenon.
- This phenomenon is a source of concern for many stakeholders concerned with making high-consequence decisions based on analyses from/on computational simulation.
- Although there are many ideas and opinions concerning the origin of this phenomenon, little if any research has been done that directly addresses it.

As an initial step toward building an understanding of this phenomenon, the basic concept for this experiment was straightforward: Arrange for multiple analysts to independently solve a chosen problem involving computational simulation, and conduct interviews with them before, during, and after their work on the problem to uncover the specific reasoning, beliefs, and judgments that lead to any observed differences in how they choose to conduct their analysis. While many hypotheses concerning sources analyst-to-analyst variability have been suggested, the goal here was to surface such hypotheses based upon carefully collected evidence.

2.2 Constraints

The initial plan was to recruit up to six qualified analyst participants for the study. Participants would work the test problem independently and would each be interviewed before, during, and after completing work on the problem. Unfortunately, we were unable to recruit enough volunteers (presumably because of scheduling problems) before funding to support analyst participants ran out.

Another opportunity presented itself when the test problem designed for the original experiment was selected for use in another project, the End-to-End UQ Frameworks project. This project was similar to the original experiment we had proposed in that multiple analysts were to work the problem, and their methods and solutions would be subject to some comparative analysis. Unlike the originally planned experiment, however, the focus of the End-to-End UQ Frameworks project was specifically on methods, so there was an opportunity to recruit volunteers for a parallel experiment from among the participants in the End-to-End UQ Frameworks project. Participants in this parallel experiment would then be interviewed, as planned for the original experiment, and the interviews would be analyzed with a focus on the human factors in choice of analytic methods, i.e. beliefs, judgments, and assumptions.

Another difference from the originally planned experiment was that there was no formal isolation between analysts working the problem; participants would meet together at regular intervals to present and discuss their approaches and results. One consequence of discussions and potential collaboration between analyst participants seemed to be a potential increase in complexity of analysis, in that such interactions could not be excluded as potential factors when tracing analyst-to-analyst variability that was observed. However, this possibility had to be weighed against the fact that formal, "complete" isolation—where individuals would work the problem entirely independently—was an artificial constraint that seemed to have the potential to magnify analyst-to-analyst variability beyond real-life practice. After consultation with Sandia's Human Subjects Board (HSB) to determine that there was no substantial increase in potential risks to participants, we decided to move forward with this moderately modified form of the original experiment.

A further constraint that emerged in time was that work on the problem ultimately took much longer than originally anticipated. Ultimately, participants just managed to complete an initial pass through the problem by the end of the fiscal year, and even then, the analysis that was conducted on the latter stages of the problem was necessarily less specific and complete than that performed earlier. Moreover, the End-to-End UQ Frameworks project is continuing, with some participants continuing to refine their analysis and elaborate on their answers in the current year.

As a result, the focus in this report is on the analysis conducted for part A of the test problem, with only limited discussion of work on parts B and C.

2.3 The Test Problem

2.3.1 Overall Problem Structure

The test problem (full details in Appendix A) originally devised for the LDRD experiment and subsequently used for the End-to-End UQ Frameworks project is focused on a rectangular cantilever beam:

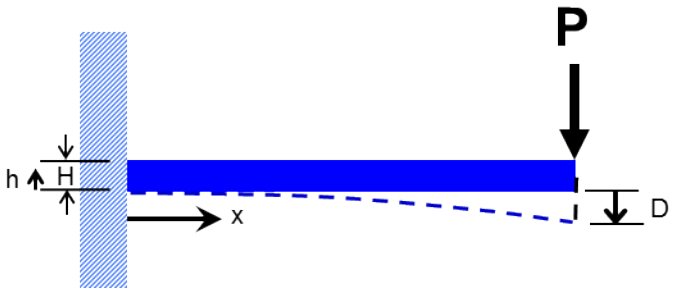

Figure 1: A representation of the cantilever beam on which the test problem is focused.

The ultimate objective is to assess the risk of structural failure in such beams on behalf of their hypothetical manufacturer. The dimensions of the beams are L (length), H (height) and W (width). When the beam is subject to a vertically applied downward load, P, the magnitude of its downward deflection is D.

E is a material property of the beam which varies with the temperature, T, and is a factor in the algebraic model that is given for D:

$$D = 4PL^3/(EWH^3)$$

Three parts are specified for the test problem: part A, characterization of data uncertainty; part B, model calibration; and part C, model validation and use for prediction.

2.3.2 Data Uncertainty, Without Measurement Error (Part A.1)

Part A is further divided into four stages. In part A.1, analysts are asked to quantify the aleatory variability in and/or epistemic uncertainty over the deflection of such beams in general based upon four sample observations provided, and are further and more specifically asked to estimate, with uncertainty, the probability of the deflection of such beams exceeding a critical response level of 0.1813 displacement units. Such a probability is referred to as an "exceedance probability", and represents a factor that might typically be considered in engineering design.

To provide further insight into the methods applied, three sets of four observations are provided and analysts are asked to provide estimates based upon *separate* consideration of each set of observations.

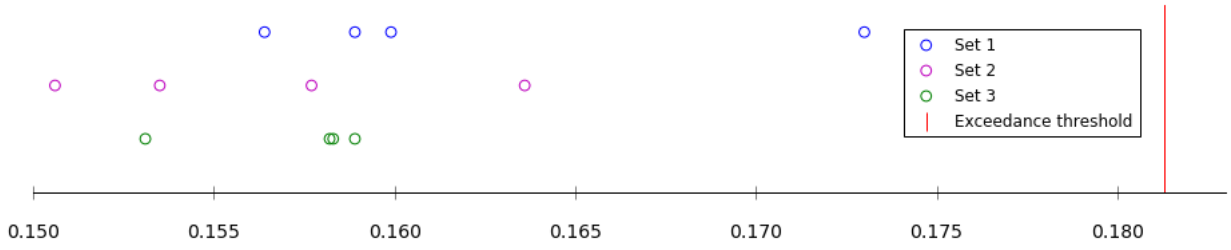

Figure 2: Beam deflection samples provided for part A.1.
Analysts were asked derive estimates from considering each sample separately.

2.3.3 Considering Measurement Error (Parts A.2 and A.3)

In parts A.2 and A.3, analysts are asked to consider the effects of measurement error in their estimates. Part A.2 involves systematic and random errors in the measurements of experimental outputs (beam deflections) in the multiple experiments. Part A.3 adds systematic and random errors in the measurements of beam experimental conditions (beam length, width, and height, and applied loads) in the multiple experiments.

For example, in A.2.a, analysts are provided with another table with three sets of four observations, where each value is reduced by some % magnitude relative to the values given in part A.1. Analysts are asked to use the reduced values to provide revised deflection estimates and exceedance probabilities, but considering this perturbation as a *systematic* error in measurement of uncertain magnitude that is expected to lie with uniform probability somewhere between 0% and -2.0%.

In A.2.b, analysts are provided with a third table, where each value corresponds to one from the data table in part A.2.a, but now with an additional random error term added to it. Unlike the error term added in part A.2.a, this error term is not systematic; the precise magnitude is presumed to vary from one observation to the next. Analysts are again asked to use these values to provide revised deflection estimates and exceedance probabilities, but now considering this additional source of measurement error with a magnitude of variation expected to be distributed normally, with a mean of zero and a standard deviation equal to 0.5% of the measured value.

2.4 Interviews

Three interviews were conducted with each participant: One just after they had begun work on the problem, and then two more after they had completed their work over the first year of the End-to-End UQ Frameworks projects. The first of the latter two interviews was focused on their own work on the problem, while the broader scope of the second included explicit consideration of other solutions that had been proposed.

3. RESULTS

3.1 Summary Comparison of Proposed Estimates

At least four distinct analyses were presented for part A, three of which were conducted by analysts who volunteered to participate in interviews for the parallel study.

Participant T's precise numerical results for part A.2.a and A.2.b were unavailable at the time of this report.

3.1.1 Part A.1 - No Measurement Error

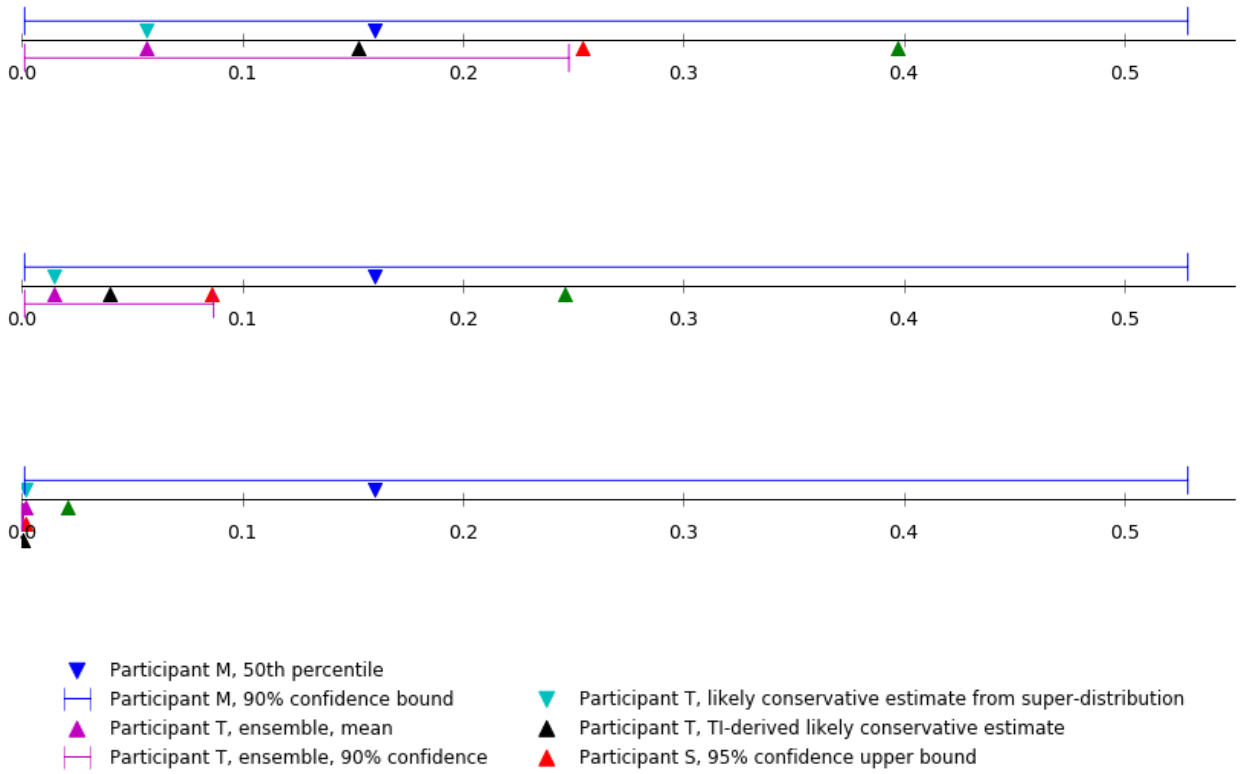

Figure 3: Proposed estimates of the probability of exceedance, part A.1 for supplied sample sets 1 (top), 2 (middle), and 3 (bottom)

3.1.2 *Part A.2.a – Including Uncertain Systematic Bias*

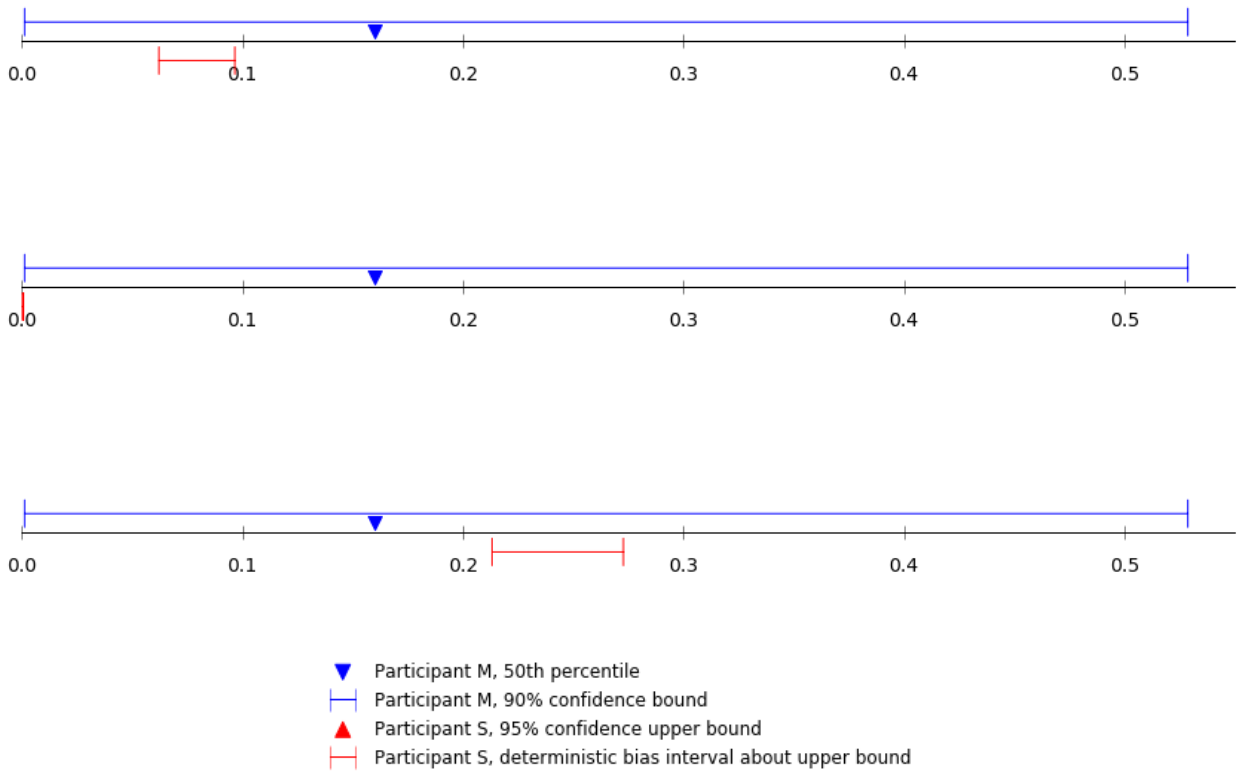

Figure 4: Proposed estimates of the probability of exceedance, part A.2.a for supplied sample sets 1 (top), 2 (middle), and 3 (bottom)

3.1.2 *Part A.2.b – Including Uncertain Systematic Bias and Random Error*

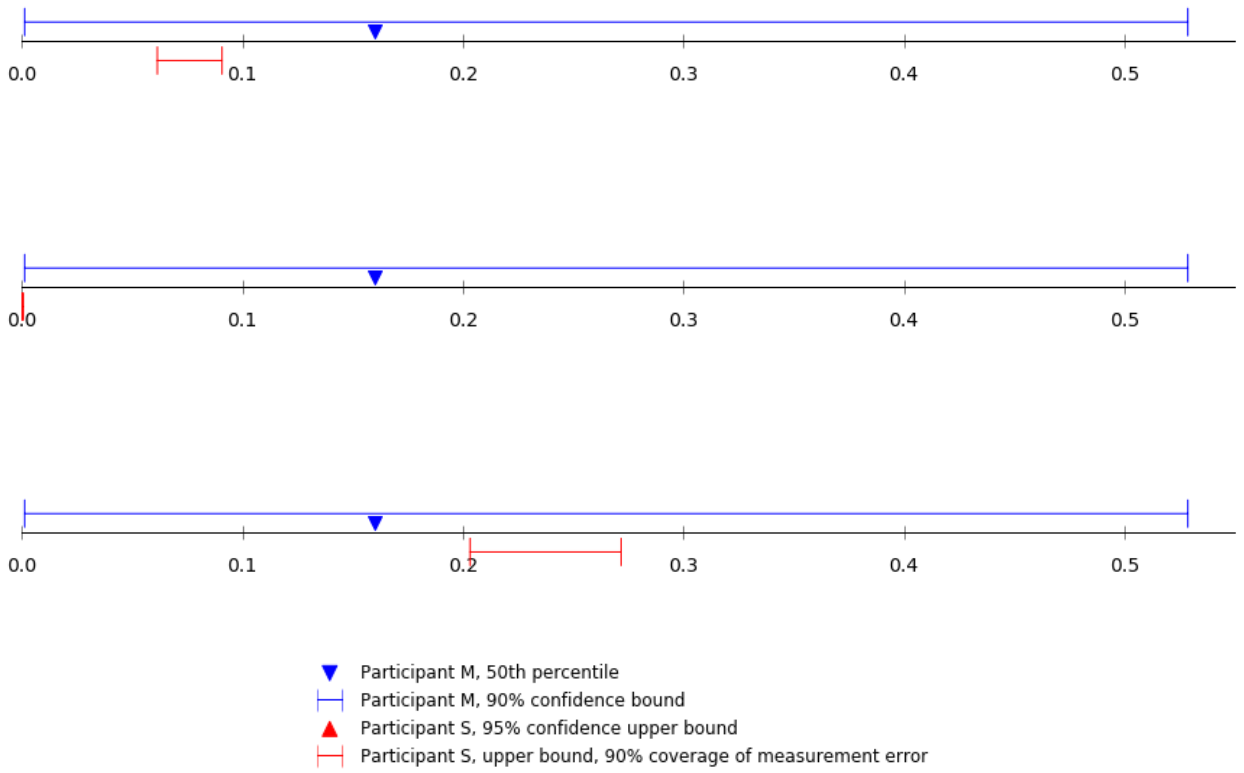

Figure 5: Proposed estimates of the probability of exceedance, part A.2.b for supplied sample sets 1 (top), 2 (middle), and 3 (bottom)

3.2 Case Study: Participant M

3.2.1 *Background*

Participant M has a PhD in a statistics-related field and was in a faculty position at a university for two years before joining the technical staff at SNL. M had been working at SNL less than a year before this study.

M emphasized the cultural emphasis in their statistics community on clearly stating the assumptions associated with any specific analysis when publishing.

3.2.2 *Proposed Solution Methods*

Participant M did not provide an estimate of D with uncertainty, but did provide an estimate of the exceedance probability.

They did this by first mapping all of the sample deflections provided to a discrete domain with two possible values: exceeding the threshold of interest, or not exceeding the threshold.

The resulting samples were then modeled in a non-parametric fashion using a binomial distribution, for which there are many methods for estimating p (the probability of exceeding the threshold, in this case) with confidence bounds.

Because none of the sample sets included a value of D that exceeded the given threshold, estimates were the same for all three sets, including both a point estimate at the 50th percentile and a one-sided 90% exact binomial confidence bound. Moreover, because the threshold was beyond maximum magnitude of measurement error to be considered from all samples, M's estimates remained the same for parts A.2 and A.3.

3.2.3 Reflections

As a statistician, M was very clear from the beginning that they would need either (a) more data, or (b) a source of engineering judgment before being able to make any parametric distributional assumptions concerning the population variability with respect to beam deflection under the load P_0.

M indicated that their choice of method stemmed at least partially from the fact that none of the samples provided exceeded the given threshold. In contrast, they pointed out, if *m* out of *n* samples failed, the "point estimate" would naturally be *m* divided by *n.* With zero failing samples, the question then becomes where to locate the point estimate, which must fall somewhere on an interval bounded by 0 on the low end and *1/n* on the high end. M chose the 50th percentile of the estimated binomial distribution, in keeping with historical precedent at SNL.

While a Bayesian method could potentially be used to estimate the distribution of the exceedance probability, M found it less preferable in this case because it seems difficult to identify a truly non-informative prior when you are faced with this issue of trying to estimate probability of failure without any actual samples of failure.

M didn't encounter any particular surprises while working on part A. However, on part B of the problem, M initially assumed that the uncertainty in the elasticity parameter, which part B asks participants to infer via model calibration, was only epistemic, i.e. that elasticity would not vary from beam to beam, and was thus surprised to find out that this was not the case. M observed that making this initial assumption was reflective of the general difficulty that analysts encounter in remaining cognizant of the points where they are inserting engineering judgment into their analyses, despite how important it is to do so.

Due to the minimal, relatively innocuous assumptions made by the nonparametric method they employed to estimate the probability of exceedance, M expressed high confidence in the resulting estimate. In contrast, M found some of the other answers proposed difficult to evaluate because it was hard to clearly identify their inherent assumptions.

Generally, M shared, the less informative the answer, the higher the confidence one will tend to have in the answer. A related concern that M expressed was that when presenting more informative answers in which one has correspondingly less confidence, there is a risk that the answers are what will be retained while concerns about confidence may not be.

3.3 Case Study: Participant S

3.3.1 Background

Prior to coming to SNL to work as a postdoctoral fellow, participant S obtained a PhD in Chemical Engineering, focusing on validation and uncertainty quantification.

Concerning the relevance of their graduate studies, S observed that while, similar to in the test problem, sparse experimental data was characteristic of their graduate research domain, experimental data was in fact *so* sparse in their graduate research domain that data did not end up being used for inference procedures such as calibration.

S had been at SNL for only a few months before this study, during which they participated in refining the test problem to be used in this study. In particular, S was responsible for generating the sample sets of measurements provided for the problem.

S studied Bayesian methods in graduate school, but ultimately never applied them outside of the pedagogical context. S suggested that having learned Bayesian methods before learning any frequentist methods resulted in a significant bias toward Bayesian methods on their part.

3.3.2 Solutions

S employed Bayesian inference for part A, following a procedure that had been documented in a research paper by a colleague.

A key assumption in their analysis is that the distribution of deflections one would observe across the entire population of beams would be reasonably approximated by some normal distribution. The specific shape of this normal distribution would then correspond to specific values of μ and σ, the mean and standard deviation of the distribution. The aleatory variability in the deflection of any particular beam is then represented by the normal distribution, while the epistemic uncertainty in the specific location and variance of the normal distribution is represented by distributions over mu and sigma.

Based upon the likelihood of the given points being generated a normal distribution with mean, μ, and standard deviation, σ, S used a Markov Chain Monte Carlo (MCMC) algorithm (implemented in Python) to sample the posterior distributions over μ and σ after observing each of the three sets of samples. In addition to performing this inference for the datasets in the problem, S also tested the method by drawing samples from both normal and uniform distributions, and then using the Bayesian method to then infer these known distributions.

Measurement uncertainty was, in both the systematic and aleatory cases, treated by effectively "backing out" the measurement error from the given sample sets before running MCMC. For systematic error, the process is deterministic: you can run MCMC for the two sides of the interval. Accounting for the normally-distributed, aleatory error introduces another level of sampling, where prior to beginning each MCMC chain, you first subtract from each point an error term with magnitude sampled from the given distribution of error.

3.3.3 Reflections

Concerning engineering judgment and assumptions one might make in solving the test problem, S reflected on a tension inherent in the design of the problem. From an engineering standpoint, there are several assumptions one might reasonably make in considering the magnitude of deflection to be expected in particular scenarios across a population of manufactured beams. At the same time, one would not expect sample observations to be so sparse in a real-world problem concerning such beams. Although not directly stated in the text of the problem, it is clear that the reason the data provided are so sparse is that the problem is designed to be representative of scenarios faced by analysts at SNL, where instead of beams, one is asked to make inferences concerning the properties of a complex, engineered component. In contrast to the case with beams, it is much more difficult to apply engineering judgment in scenarios where the properties of components depend on physical principles that span a diverse range of engineering disciplines.

S further stated that while Bayesian methods might be grounded in an elegant philosophy, given the current state-of-the-art they cannot typically be applied in an "off-the-shelf" manner; practitioners are obliged to a variety of assumptions. Such assumptions concern not only prior distributions, but arise also at the algorithmic level, e.g in choosing values for parameters governing the specific behavior of your MCMC algorithm.

S appreciated M's point that estimating deflection with uncertainty from only 4 points required making assumptions for which no explicit basis is provided in the test problem statement. In light of seeing M's proposed nonparametric approach to estimating the exceedance probability, S reasoned that this would be a good approach to apply before any others. It is possible that the bounds on the resulting estimate might provide a basis for a decision concerning the given design requirement without having to inject additional assumptions, in which case significant computational effort would also be saved. Moreover, comparison between the nonparametric estimate and that resulting from another method would provide a sense of the magnitude of the differential in certainty which might rest on the assumptions in inherent in the other method.

Nonetheless, based upon empirical findings from a range of other problems, S estimated that the assumption of normality was fairly safe and conservative, barring specific reasons to exclude it.

While S expected some variability to occur in interpretations of the specific wording of the test problem, S was surprised at the number of questions that arose, particularly concerning the precise effects of measurement uncertainty.

Concerning solution confidence, S drew a distinction between confidence in methods and confidence in specific answers. S expressed a fair amount of confidence in their methods, but somewhat less in their specific answers at the time of the interview. The differential, they explained, was due to the range of variations in procedures and assumptions that they had explored in the course of their analysis, particularly given that the real focus of the End-to-End UQ Frameworks study was ultimately on a comparative analysis of methods rather than on the specific, numerical answers to the test problem. S expects to reach a high level of confidence in their answers as they continue to refine their analysis.

3.4 Case Study: Participant T

3.4.1 Background

Participant T has a PhD in Aerospace Engineering, focusing on computational fluid dynamics during graduate studies. While issues of model validation and uncertainty quantification were certainly an important consideration in T's community of practice in graduate school, there has been a clear progression toward more formal treatment of these issues over the course of T's years as a member of the technical staff at SNL prior to participating in this study.

T's technical expertise is deepest in code and solution verification, and has significant experience in the application of SNL's Dakota, a comprehensive software framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis [7]. While T has experience in validation and, more specifically, uncertainty quantification, they made clear that they did not have formal training in statistics. While expressing significant comfort with the big picture view of the problem, they expected to learn significantly from the particulars of performing the analysis.

3.4.2 Solutions

On part A, T worked closely with a colleague who suggested specific methods of analysis. T ultimately applied three different methods on part A to arrive at a set of (potentially) different estimates. As with the analysis of S, all three of these methods were based on an assumption that the true population of deflections could be reasonably bounded using a family of normal distributions consistent with the epistemic uncertainty caused by sparseness of the data samples.

Two means of estimation that T applied were based upon first generating, for a given set of 4 samples, 5,000 samples of μ and σ representing 5,000 possible normal distributions from which the 4 points may have been sampled. Samples of μ were generated using Student's *t*-distribution and samples of σ were generated using the chi-squared distribution, both parameterized to represent samples of size 4.

Each normal distribution in the resulting ensemble represents a hypothesis concerning the variability one might observe if one were able to inspect the full population of beams, i.e. the irreducible, aleatory uncertainty concerning the deflection of any particular beam under load P_0. The ensemble of all of these hypotheses, then, constitute a sample from the overall distribution

of epistemic uncertainty concerning the precise location and variance of the (presumed) normal distribution of variability of deflections.

This ensemble of normal distributions was then used in two different ways to produce estimates of the exceedance probability. One method was to ascertain, for each distribution in the ensemble, the fraction of the distribution which exceeded the critical threshold, thus yielding a sample of 5,000 exceedance probabilities. These 5,000 samples, then, represent the epistemic uncertainty over the true fraction of the population that would be expected to deflect by more than the threshold value, from which 5^{th} and 95^{th} percentile values were considered bounds on a 90% confidence interval on the probability of exceedance.

The second method was to generate 1,000 samples from each normal distribution in the ensemble, yielding a "super-sample" of 5,000,000 deflection values. The fraction of this super-sample which exceeded the critical threshold was then taken as a highly-likely conservative bound estimate of the exceedance probability that incorporates both the aleatory variability in the population and the epistemic uncertainty concerning that variability.

The third and computationally least expensive method employed by T was based on the calculation of *tolerance intervals*. A tolerance interval is an interval in which a particular range of the population, say the 5% to 95% range, may be expected to be found, with confidence $1-\alpha$. For each set of four samples, T used a statistical table to calculate a tolerance interval to cover 95% over the population with 90% confidence. They then selected a normal distribution with the same coverage, i.e. a normal distribution with the mean at the midpoint of the interval and a standard deviation such that each endpoint of the interval was 2σ from the mean. Then, using this normal distribution as an approximation of the population distribution, they inferred highly-likely conservative bound on the probability of exceedance by determining the fraction of the chosen normal distribution that exceeded the threshold.

T applied all three of these procedures to each of the three sets of 4 samples specified in part A.1. Next, T and their collaborating colleague decided to forego further application of the super-sample method, judging that it provided what was essentially the same estimate as the distribution-of-distributions method, but without any associated bounds on confidence. Application of the remaining two methods to the data sets specified in parts A.2 and A.3 was then similar, but involved additional levels of sampling to yield sample hypotheses that each corresponded to a magnitude of systematic bias sampled from the interval specified for part A.2, and then to additional samples of aleatory error drawn from the normal distribution specified for part A.3.

3.4.3 Reflections

T made clear that if they were working on this problem as a real problem, they would start by consulting with multiple colleagues to solicit not only specific advice with respect to solution methods, but also general thoughts and impressions concerning related problems the colleagues might have encountered. One reason it is valuable to solicit multiple opinions, T reasoned, is that while there are many individual experts, it's rare that any one individual has all the expertise that's relevant when considering a rich, real-world problem.

Another perspective that T emphasized concerned interpreting the results from different methods of estimation, particularly concerning possible discrepancies between them. Discrepancies are unlikely to be truly surprising, as T pointed out, if one consider how differences at the computational level ultimately imply differences in the precise questions to which each result represents an answer.

T was confident in expecting large uncertainty in any solution, due to sparse data combined with a focus on exceedance probabilities, which typically focus on the tails of a metric distribution where epistemic uncertainty can be high even when you have a significant amount of data.

T expressed agreement with their colleague's judgment that the ensemble-of-normal-distributions method was superior to the super-sample method in that the former naturally provided a distribution of uncertainty concerning the quantity of interest, while the latter did not.

Not having significant familiarity with Bayesian inference, T was surprised to see how similar S's method was to their own at the level of the computation that was actually being performed, despite greater differences in the language most naturally used to describe them.

Regarding confidence in their methods and the resulting estimates, on one hand T had concerns that methods such as the Bayesian inference procedure implemented by S, having a richer basis in theory, were likely to be applicable across a broader range of scenario variations. In particular, T was concerned that if an assumption such as normality turned out to be unjustified, it was unclear what means their method might provide for detecting this condition, whereas the Bayesian framework might provide some means of doing so, e.g. by comparing prior and posterior distributions. Nonetheless, given that only 4 samples were available for consideration in this particular scenario, T was reasonably confident that their method(s) would perform as well as anything else.

Further, T shared that although they weren't certain, they had at least some sense that normal distributions were a good match with the underlying truth model used to generate the sample data for the test problem.

At the overall level, T suggested that a central question in this exercise concerned weighing the possible reductions in uncertainty one might seem to gain with more elaborate methods against not only increased complexity and computational effort, but also potential dependence on stronger assumptions. In choosing a method that makes stronger assumptions, it's important to communicate clearly about the extent to which the assumptions appear to be justified, and not imply (by omission, perhaps) that they are more justified than they seem.

Concerning what they would do differently, T reiterated that they would consult with colleagues on a problem such as this one, particularly being conscious of not having a technical background that's directly related to the specifics of part A.

T further emphasized that sparse data concerns are critical and need to be addressed up front. While a clear understanding of what you hope to get out of a given analysis is always important,

it’s of particularly importance to consider in light of sparse data. With few data points available, any information that might be available about the points, or choices you might get to make with respect to which points are to be tested might have a large impact. Similarly, your analysis may be even more sensitive to distortions of any measurement(s), particularly those that you are unaware of.

4. DISCUSSION AND CONCLUSIONS

4.1 Tracing Solution Variability to Differences in Methods

A central goal of this study has been to trace solution variability to differences in computational methods, and then differences in computational methods to difference in individual differences in reasoning, beliefs, and assumptions.

Differences in computational methods here are well documented. With a few more details, such as specific version numbers of particular software packages and perhaps a few low-level numeric parameter values, the information here should be sufficient to replicate the computations performed by the analyst participants. Complete procedural specifications, however, are of limited value toward understanding what the essential differences are between differing results, toward understanding, as participant T put it, the possibly subtle differences in the precise questions to which each result represents an answer.

To the extent to which analysts *generally* agree on the question they seek to answer, and to the extent to which they manage to carry out the computational procedures they intend to perform (i.e. avoid errors in executing their intended computations), remaining discrepancies between results ultimately may come down to differing assumptions made in each analysis.

The central difference in assumptions made by analysts in part A concerned whether or not variability in deflection over the true population of beams might be reasonably approximated by a normal distribution. Making this assumption results in a substantial reduction in uncertainty, and essentially explains the large discrepancy between the upper bound computed by participant M and those of the others.

4.2 Tracing Differences in Methods to Individual, Human Factors

More uniquely, in this study, it is possible to trace choice of methods to participants' reasoning, assumptions, and background:

- With a background in statistics, participant M was highly conscious of both the parametric distributional assumptions necessary to characterize uncertainty in deflection given such sparse data, as well as a need for explicit guidance that would be needed to justify such an assumption.

- Use of Bayesian methods was a natural choice for participant S, given extensive study of such methods in graduate school. Moreover, with an engineering background, S was able to justify the assumption that the magnitude of deflection one would observe across the population of beams could be reasonably approximated by a normal distribution.

- With the least formal training in statistical methods, participant T had little bias toward choosing any one method over any other, and found the non-Bayesian methods described by their colleague accessible and reasonable. Furthermore, more extensive, "big picture"

experience with simulation-based analyses may have supported taking an explicitly empirical approach applying multiple methods.

4.3 Hypotheses Explaining Analyst-to-Analyst Variability

Our observations do seem to demonstrate a clear linkage between analysts' background and experience and their particular choices of methods. While perhaps unsurprising, this dependence is not one that tends to be *explicitly* documented or discussed. We may consider this observation as support for one hypothesis concerning analyst-to-analyst variability, i.e. that variability in choice of methods simply stems from variability in analysts' experience.

Less obviously, there is at least some small amount of evidence here for another causal factor in analyst-to-analyst variability, seen most clearly in (but not limited to) participant S's statement that seeing M's nonparametric analysis led them to believe that it would be best to perform such an analysis before deciding whether or not to go forward with a computationally more expensive analysis that depends on a stronger assumption. While S's statement clearly reflects an open-minded rationality which is essential to progress in quality of analysis, it may also be seen as one clear instance of a natural (and necessary) phenomenon of unknown but potentially large significance: the dependence of our analytic choices on what we see and experience from day to day, or even on what does or does not "occur to us".

4.3 Conclusions

In common with responses to V&V and UQ challenge problems presented elsewhere, we observe here some clear variability across both methods and solutions. In contrast to some of these other contexts, we have not (thus far) observed uncertainty bounds on solutions that fail to overlap.

This is not an unexpected result for a few reasons. One is simply that uncertainty bounds were necessarily large given the sparseness of the data provided. Another has to do with the fact that participants had a long time to work on part A, coupled with many discussions. With ample time to compare and consider methods, participants were less likely to unconsciously embed differing assumptions in their analyses.

This clarity regarding the assumptions made in these analyses provides an explanation for the analyst-to-analyst variability we observe. In the absence of such clarity, our difficulty in explaining analyst-to-analyst variability indicates limitations in our understanding of our own analyses, limitations which raise doubt concerning our ability to evaluate or interpret the results of these same analyses in a valid manner.

The clarity needed to reduce the uncertainty associated with limited insight into our analyses extends beyond clarity in assumptions. It includes factors such as clarity concerning what is truly being asked for or is expected when one begins an analysis, and clarity with respect to the constraints imposed on an analysis by limitations in resources such as quantity or quality of data, time, or expertise. Hopefully, the line of research to which this report belongs can lead us toward a more powerful framework to support such insights.

5. REFERENCES

[1] G. E. Box, N. R. Draper, and others, *Empirical model-building and response surfaces*, vol. 424. .

[2] W. L. Oberkampf, M. Pilch, and T. G. Trucano, *Predictive capability maturity model for computational modeling and simulation*. Sandia National Laboratories, 2007.

[3] K. T. Hu, B. Carnes, and V. Romero, "Introduction: The 2014 Sandia Verification and Validation Challenge Workshop," *J. Verification Valid. Uncertain. Quantif.*, vol. 1, no. 1, p. 010301, 2016.

[4] R. G. Hills *et al.*, "Validation Challenge Workshop," *Comput. Methods Appl. Mech. Eng.*, vol. 197, no. 29–32, pp. 2375–2380, May 2008.

[5] S. N. Aksan, F. D'Auria, and H. Staedtke, "User effects on the thermal-hydraulic transient system code calculations," *Nucl. Eng. Des.*, vol. 145, no. 1, pp. 159–174, 1993.

[6] G. D. Westwater, "Variation Studies on Analyst Contribution to FEA Stress Analysis Uncertainty," presented at the ASME 2014 V&V Symposium, Las Vegas, NV, 2014.

[7] M. S. Eldred *et al.*, "DAKOTA : a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis," Sandia National Laboratories, Technical Report SAND2011-9106, Dec. 2011.

APPENDIX A: TEST PROBLEM[1]

This problem concerns stochastic physical systems as defined below. The analyst is encouraged to work as many of the data UQ characterization, model calibration, validation, and risk assessment (exceedance probability prediction and associated uncertainty) tasks in Sections A-C in the allotted time. Written responses are requested for the questions and tasks in yellow background highlighting. It is recommended that no more than 1/3 of the allotted time be spent on any item until an attempt has been made to complete the other two items.

The context in the following is that an experimental, modeling, and analysis project team needs your analysis help in working the various elements of this problem. The work is being performed for a manufacturer as a customer (or you are all employees of the manufacturer). The manufacturer requires the experiments, modeling, and analysis to assess beam failure risk in various loading conditions to meet safety objectives and define loading conditions under which it will certify its beams, but project resources and constrained and expenses and profits are also exceedingly important to the manufacturer and its survival.

Description of Cantilever Beam Physical System

The case study problem involves a cantilever beam's deflection D at the free end of the beam in Figure 1 and in the direction of a vertically applied downward load P there. Assume zero deflection (D=0) and zero slope (dD/dx = 0) of the beam where it horizontally protrudes from a rigid unyielding vertical wall. Other important parameters of the problem are the beam's geometry as specified by its length L, and height H and width W of the rectangular beam's cross-sectional area normal to its length dimension. Beam height is measured upwards from the bottom of the beam as shown in the figure. Beam width is measured perpendicular to the height and length directions. The beam is made of a homogeneous isotropic material that has strength parameter E(T) which is a function of temperature as discussed in the modeling sections B and C. A model for beam deflection is also presented there for calibration and then validation. The physical and modeled beams are spatially uniform in temperature, but different temperature regimes will be investigated. The model is not needed in section A. The analyst is asked not to use knowledge of the model and its solution when addressing the items in section A, beyond the phenomenological information provided in A.

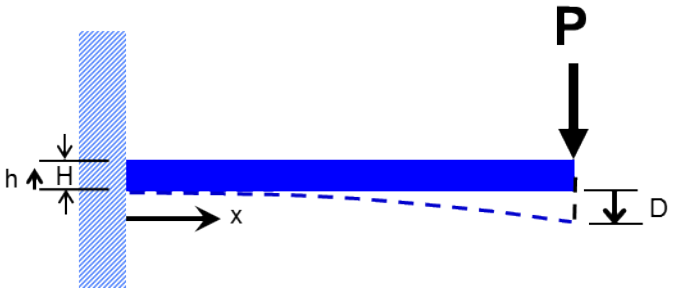

Figure 1 – Cantilever Beam Case Study Problem

[1] This is the version of this test problem used in the experiment described in this report. There will likely be further revisions to this problem before it is considered to be in its final form. Contact: Vicente Romero, 1544.

UQ Problem for Stochastic Physical Systems (i.e., Population Ensembles of Deterministic Physical Systems having Small Variations from Each Other)

We consider a population of cantilever beams for which the relevant parameters L, E(T), H, W vary among the population of beams. Let the variations for each parameter be governed by a probability density function (PDF) of unknown shape. Let the beam height H, width W, and length L be machined from three different types of machines, with control and measurement errors and uncertainties being independent (no correlation) among the three different types of machines. Let the material strength parameter E(T) also vary independently from the beam geometry parameters.

By testing randomly selected beams in nominally identical "replicate" tests in the configuration in Figure 1 and at certain temperatures and other loading conditions, it is desired to use an affordably small number of beams and tests to infer response variability in the large full population of beams (asymptotically ∞). Ultimately this characterization will be used for predicting response variability (and uncertainty thereof) in different loading conditions and at different temperatures than in the characterization tests. The stochastic behavior considered in this document is confined to that where phenomenological mechanisms governing behavior are invariant: behavior is the same for different beams that have the exact same geometric and material attributes, and any behavior variations among beams that have small geometric and material differences is governed by a smooth deterministic function of the geometric and material attributes. That is, loading is kept to regimes where no bifurcations or effectively discontinuous or other anomalous behaviors occur over the variations of beam geometric and material attributes considered in the following.

The loading set point in the replicate tests is a target value P_o, which is representative of the upper range of allowable eventual service loads for the population of beams. But small control variations from this set point exist among the tests, as measured by a load gage. In versions of this scenario, uncertainties exist in the accuracy of the gage. In the rather simple problem posed here, it is unrealistic that boundary condition (BC) control variations and/or measurement uncertainties would be as large as posed here. But in testing of many real systems, BC control variations and measurement uncertainties can be substantial— among the largest and most important uncertainties in propagated effect on system response, if not *the* largest and most important. So what may seem like unrealistically large uncertainties are used here to be more representative of uncertainty magnitudes in testing of more complex systems.

Somewhat larger geometry variations and measurement uncertainties (including deflection measurement uncertainties) exist in the present exercise than might actually exist for the simple geometries involved, but variations and measurements of system attributes and responses in real experiments can involve significant uncertainties, so this is reflected here.

Items A: Experimental Data UQ, with data-based exceedance probability estimation

1. *Aleatory response variability.* There are no input or output measurement errors (or negligible measurement errors) in this item. Input loads and beam geometry and material properties vary randomly and independently in the tests. Table A.1 lists the displacement results of three sets of tests where each set has four replicate tests yielding four random samples of response (beam displacement) per the table's title. All samples come from the same population of response but each set is to be considered separately for analysis purposes. For each set, describe the UQ methodology and results that characterize and express response aleatory variability and/or epistemic uncertainty based on the four samples[2]. If one could perform 12 tests (the total in the table) they would not normally be segregated into three sets of four because combining all 12 gives the best resolving power for characterizing response. We present the test results in segregated form so UQ results from the three sets of samples can be compared. This will provide an idea of the differences in perceived/inferred uncertainty that can arise when only four replicate tests are conducted to characterize a significantly varying random quantity. The analysts are asked not to use realizations across sets in order to have more than 4 samples to refine estimates of response uncertainty.

   The analyst is asked to use their uncertainty characterization to estimate probability of exceeding a critical response level of 0.1813 displacement units. The analyst is asked to report the three exceedance probabilities (one for each column in Table A.1) and any associated uncertainties and explain their interpretation and the process for arriving at them.

Table A.1 –Realizations of beam deflection per Figure 1 where each sample (test) involves a randomly selected beam from the beam population and a random point load that varies about the target load Po.

Realizations of deflection D (when no measurement error exists)	Set 1	Set 2	Set 3
Beam1/Test1	0.1730	0.1636	0.1589
Beam2/Test2	0.1589	0.1577	0.1583
Beam3/Test3	0.1564	0.1506	0.1582
Beam4/Test4	0.1599	0.1535	0.1531

[2] Highlighted text marks where analysts are asked to provide specific answers.

2. Now consider any change to the UQ characterization that might occur when additional information is supplied on response (deflection) measurement error as follows.

   a) *Uncertain systematic error in response measurements.* Consider a "systematic" error in the deflection measurements. This comes from a source of response measurement error that is effectively the same in all the replicate tests. For example, if the same biased sensor is used in all the replicate tests, then all the measurements may have similar bias error. The bias error associated with a given sensor is often unknown but reasonably limited to a given range or distribution of uncertainty. For example, the sensor may be a random pick from a population of sensors characterized by the manufacturer to have accuracies described by a distribution or range of error. Then this information can be used for an uncertainty estimate on the given sensor's measurement error in the tests. Systematic error that biases measurements can also come from other experimental effects and from processing of the measurements. Assume that information exists from the manufacturer and/or calibration lab and/or experimental characterization and/or theoretical analysis that systematic error associated with beam deflection measurement is expected to lie within the following range, with 95% confidence (measurement error = measured value minus true value).

$$U[\text{defl\_err\_sys}] = [-2\%, 0\%] \text{ of measured value} \qquad \text{(Eqn. A.1)}$$

This error ranges from some negative amount to zero, so measured value ≤ true value. Table A.2 entries includes a systematic error of -1.5% relative to the values in Table A.1 (= 98.5% of the values in Table A.1). These errors have magnitudes that are within the range [-2%, 0%] of the Table A.2 measured values, consistent with Eqn. A.1. Although the entries have the same %errors, the error magnitudes are slightly different for each entry in Table A.2. The error magnitudes have small perturbations from each other but are dominated by a systematic error component and can be treated accordingly. The analyst is asked to provide an appropriate UQ characterization for the values in Table A.2 in view of the uncertainty information in Eqn. A.1. The UQ approach should also be summarized. The analyst is asked not to use the information at the top of this paragraph for UQ characterization, as this information would not be available in a real problem.

If the UQ characterization of the data changes, exceedance probability estimates would presumably also change. Hence, three new displacement exceedance probabilities and any uncertainty associated with them should be provided and interpreted, along with any modifications to the exceedance probability estimation and UQ procedure. Given that each data set has essentially the same magnitude of systematic error, are their new

exceedance probability estimates and uncertainties changed by the same amount for each set relative to their results from the data in Table A.1?

Table A.2 – Beam deflection measurement realizations from Table A.1 but with added systematic measurement errors sampled from Eqn. A.1.

Realizations of deflection D (with systematic meas. error in these results)	Set 1	Set 2	Set 3
Beam1/Test1	0.1704	0.1612	0.1565
Beam2/Test2	0.1565	0.1553	0.1560
Beam3/Test3	0.1540	0.1483	0.1558
Beam4/Test4	0.1574	0.1511	0.1509

b) *Aleatory random error added to response measurements.* In this section the measurements are further subject to random measurement errors that are consistent with the following error information supplied by the sensor manufacturer.

U[defl_err_rand] = Normal(mean=0, stdev=0.5% of measured value) (Eqn. A.2)

Random measurement errors consistent with Eqn. A.2 are added to the results in Table A.2 to yield the revised deflection data in Table A.3.

Because the data in Table A.3 is subject to both systematic and random measurement errors, an amended UQ procedure might be applied to the values in the Table. The analyst is asked to provide an appropriate UQ characterization for the values in Table A.3 in view of the supplied uncertainty information in Eqns. A.1 and A.2. The UQ approach should also be explained. Although the systematic and random error components can be determined from comparing results in Tables A.1 - A.3, the analyst is asked not to use this information, as it would not be available in real problems. How close do the UQ characterizations come to the results from item A.2a? That is, can the known source of variability Eqn. A.2 be effectively used to "deconvolve" this known source of variability out of the column data in Table A.3 and thereby reduce their variance and accompanying uncertainty representations toward those for Table A.2?

A UQ option at the other extreme is to make no use of the measurement error variability information in Eqn. A.2. This option might be particularly attractive when an exact uncertainty description like Eqn. A.2 is not available, even if significant measurement error variability over the replicate tests is suspected. If one chooses to "do nothing" in the presence of non-negligible random measurement-error variability, is the uncertainty or risk associated with this choice expressed? If so, how is it expressed?

In view of the new conditions in this subsection, three new displacement exceedance probabilities and associated uncertainties should be provided and interpreted, along with any modifications to the exceedance probability estimation and UQ procedure.

The data in Table A.3 reflects only one realization of random errors over the 4 tests that could occur in accordance with Eqn. A.2. An auxiliary file is provided that has 100 random versions of Table A.3. By applying their UQ methodology to these 100 versions of the problem, the analyst can get a sense of the different results that would be yielded under a broad array of random measurement error realizations. It is anticipated that the spread of analysis results might be surprising large for the relatively small individual errors of Eqn. A.2.

Table A.3 – Beam deflection measurement realizations from Table A.2 but with added random measurement errors sampled from Eqn. A.2.

Realizations of deflection D (w/systematic and random meas. errors in these results)	**Set 1**	**Set 2**	**Set 3**
Beam1/Test1	0.1685	0.1594	0.1548
Beam2/Test2	0.1573	0.1562	0.1568
Beam3/Test3	0.1538	0.1481	0.1556
Beam4/Test4	0.1560	0.1497	0.1494

3. Change of scope of stochastic system being characterized, and associated data normalization for boundary condition variability

Now consider a case where the system to be analyzed is defined to include the population of beams but not the loading apparatus; we want to characterize variability of response due to beam variability without any loading variability. To the previous information the loading variability information figuratively illustrated in Figure A.2 is added. Given the quantitative information below, the analyst is asked (as specified more precisely later) to provide UQ characterizations of deflection for the full population of beams if each beam is subjected to the same fixed load Po=750,000 (the target loading value).

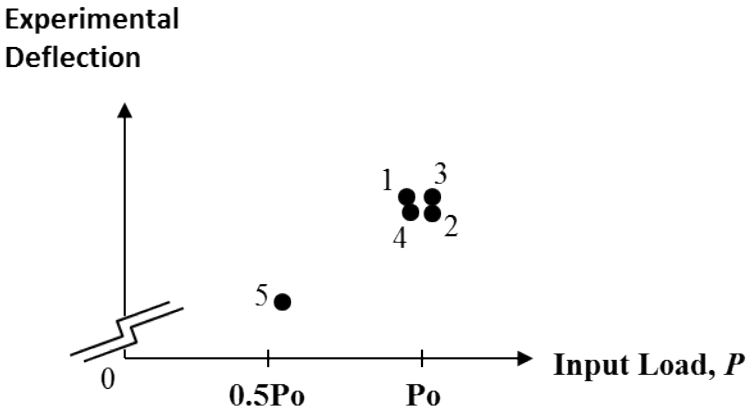

Figure A.2 – Illustration of generic variability of beam loading and associated deflections in tests at target loading values Po and 0.5Po

a) Normalizing response variability for small BC differences in replicate tests—here ignoring any BC measurement errors

Figure A.2 shows a generic illustration of load variations and associated deflections in tests at target loading values of Po and 0.5Po. The four replicates shown about the target load Po are generic versions of the four replicate tests whose measured deflections are listed per set/column in Table A.3. Recall that the deflections in the table vary not only because the input loads vary, but also because the beam geometry and properties vary randomly over the four replicate tests and because deflection measurement error varies from test to test.

Table A.4 lists the load variations involved in the tests/results reported in Tables A.1 – A.3. Table A.4's information, along with information about the local functional

relationship between response and load magnitude (see below) can be used to approximately reverse the effects on response variability contributed by imperfectly controlled loading or boundary conditions in replicate tests. Here this is termed "normalizing out" the effect of known BC variations. The expectation (on average) is more accurate response mean, variance, and exceedance probability results for the specified load Po applied to the beam. In the normalization objective below, Po is the nominal or reference value about which measured loads in the tests are characterized to vary.

In real experiments, boundary condition variations are often spatial and temporal in nature, and parametric descriptions of the variations, e.g. as scalar PDFs or random fields of the variations or estimated variations, are usually not available. So here we do not offer PDF information for variability of the imposed scalar loading in the replicate experiments. Instead we offer what would be available in most real experimental settings: the variations of applied boundary conditions as measured in the replicate tests. The further difficulties of reconstructing field BCs from spatially sparse sensor data are not visited here.

Table A.4 – Cantilever Beam loading variations in replicate tests whose deflections are given in tables A.1 - A.3.

Varying input load P (as measured, subject to meas. errors)	**Set 1**	**Set 2**	**Set 3**
Beam1/Test1	800,000	725,300	762,200
Beam2/Test2	771,300	777,100	765,600
Beam3/Test3	736,400	730,900	759,500
Beam4/Test4	773,600	772,100	769,100

Load-deflection relationship information is needed in order to normalize-out response variations due to the known load variations about Po. Experiments can be used to estimate the relationship's variability and uncertainty (each beam in the population could have a slightly different load-deflection relationship because geometry and material variations affect deflection as a function of load). In view of the nontrivially varying loading conditions in the tests (Table A.4), we next consider how this might be addressed if one auxiliary test can be afforded and negotiated for, and then another test as a hypothetical proposition.

The testing (experimental) design in Figure A.2 features one experiment to characterize response at a substantially different level of loading. The figure represents a situation where no two tested beams are the same beam. This gives information on load-deflection response that is partially confounded by the response differential from using different beams (with somewhat different material properties and dimensions) at Po and 0.5Po. This type of confounding is common in experiments that alter test units or are destructive tests altogether, so that units cannot be tested twice without raising substantial questions about the second result. Furthermore, procedures to pair units and their tests and responses must be instituted at the time of those tests. Unfortunately this is neglected all too often in real projects. Then the unit-result correspondence is lost. This undermines later possibilities if a need is later identified to test some of the same units at different conditions.

If the beams tested at 1.0Po are not damaged in any way, the test at 0.5Po could use one of the beams tested at 1.0Po. Then there would be no confounding from using different beams. (Although in general, some confounding would still exist if non-negligible random measurement errors exist on measured input loads and deflections at target values 0.5Po and 1.0Po.) The reference load value Po is representative of the upper end of design service loads for the population of beams. So "just in case" any damage occurred in the four beams already tested, a new beam is used for the 5^{th} test in Figure A.2.

A rough uncertainty analysis by the project team determines that a 50% change in loading will conservatively assure that any confounding "noise" factors will add relatively small uncertainty to estimation of the load-deflection relationship. So a set point of 0.5Po is chosen for the new test. Resident project knowledge about the beam materials and beam behavior indicates that similar phenomenological mechanisms of behavior exist at this lesser loading; a similar regime of physics applies. But a tradeoff exists. If the load-deflection relationship is significantly nonlinear between 0.5Po and 1.0Po, then such a large difference in loading will undermine accuracy of inferred load-deflection slope at the reference load 1.0Po where normalization is to be performed.

The other factor here is that only one test is carried out at the target load 0.5Po. This does not reflect the various load-deflections relationships of the differing beams in the population. Uncertainty procedures might be employed to approximately address this. Table A.5 lists the measured deflection at target load 0.5Po. Systematic and random measurement errors from Eqns. A.1 and A.2 exist in the measurement. The deflection result is also subject to a small loading deviation about the target 0.5Po, per Table A.6.

Table A.5 – Cantilever Beam deflection response in test at target load 0.5Po.

	measured deflection D (subject to measurement errors)
Beam5/Test5	0.0803

Table A.6 – Cantilever Beam loading deviation from target load 0.5Po and whose measured deflections are given in Table A.5.

	measured input load
Beam5/Test5	380,600

In this subsection assume there are no measurement errors associated with the load data in Tables A.4 and A.6. The data in Tables A.3 and A.5 contain random and systematic measurement errors consistent with the uncertainties stated in Eqns. A.1 and A.2

The analyst is asked to use the data in Tables A.3 - A.6 and the information above to normalize set 1, 2, and 3 data to be more consistent with the specified target load Po for which response uncertainty and exceedance probability are requested. If the analyst chooses not to normalize the data, please state the reasons. In either case, please provide for each data set 1,2,3 a characterization of beam deflection variability and any associated uncertainty inferred for the full population of beams if each beam is subject to the same load Po. Please also provide corresponding exceedance probability estimates and uncertainties, and a summary of any new UQ analysis and procedures.

How do set 1, 2, 3 results (including exceedance probability estimates) change vs. section A.2b results?

Estimating the value of potential new information and where to test

In hindsight, or even prior to the results in Table A.5 and A.6, for the objective of estimating exceedance probability at the specified analysis load Po, can it be reasonably determined whether it would be better to use the test #5 for another replicate at the target load 1.0Po instead of at 0.5Po for normalization of the data (if normalization was done)? Please explain any reasoning and justification underlying the answer.

If yet another test was feasible in the project, at what target loading value would the analyst recommend the sixth test be performed to add to the five in Figure A.2 to best support the objective of estimating exceedance probability at the specified analysis load Po? Please explain any reasoning and justification underlying the answer.

b) *Normalizing response variability for small BC differences in replicate tests, accounting for aleatory and systematic/epistemic errors in the measured BCs*

Here we consider load measurement errors (random and systematic). These errors are resident in the measured BC data in Tables A.4 and A.6. The errors are consistent with the following supplier error information. For random errors:

U[load_err_rand] = Normal(mean = 0, stdev=1% of measured value). (Eqn. A.3)

Systematic measurement errors are expected to lie within the following range, with 95% confidence.

U[load_err_sys] =[-2%, +2%], as a % of measurement (Eqn. A.4)

When accounting for the errors in the UQ procedures, does the uncertainty in the load-deflection relation increase significantly vs. case A.3.a? Does this have a significant effect on response UQ characterizations and exceedance probability results? What are the new UQ characterizations and exceedance probability results? What are the changes to the UQ approach?

If aleatory and/or epistemic measurement uncertainties in input load come from different populations or estimates for 0.5Po and 1.0Po loads because the respective loading apparatus are configurationally different, is the UQ approach readily extensible to this case?

Items B: Model Parameter Calibration for Various Model Forms, Information Sets, and Prediction Scenarios, with model-based exceedance probability estimation

In this section B the analyst is asked to calibrate model parameters for various model forms, prediction purposes, and experimental input and response information. For certain cases the analyst is asked to use the calibrated parameters in prediction models to estimate exceedance probability with uncertainty, and compare to data-based exceedance probability estimates from section A.3b.

Physics Models and Parameters for Calibration

An ordinary differential equation (ODE) for beam deflection derived from a balance of forces and moments in the classical beam problem we are considering is (e.g. [19]):

$$\frac{d^2}{dx^2}\left(\frac{1}{12}WH^3E\frac{d^2}{dx^2}D(x)\right) = q(x) = P\delta(x\text{-}L) \qquad \text{(Eqn. B.1)}$$

Here x is a horizontal coordinate that starts at the wall (x=0) and runs along the length of the beam to its free end at x=L as indicated in Figure 1. Geometry parameters of the beam were described at the beginning of this document. E=E(T) is the beam's modulus of elasticity, a material stiffness/strength property that is a function of temperature in the present problem. The model is written for beams with isotropic and spatially uniform modulus of elasticity, which here requires spatially uniform temperature in the beam so that E is not a function of x. Uniform beam temperatures exist also in the experiments considered in this document. The generalized loading case involves a general distributed load $q(x)$ on the beam. In the present case the point load P in Figure 1 is represented by $q(x)$ being a delta function $\delta(x\text{-}L)\cdot$P that mathematically recovers the point load P at x=L (see [19]).
Equation B.1 together with the relevant geometry and material property values and initial and boundary conditions constitute the model for beam deflection behavior in our problem. An analytic solution to the governing equations and parameter variables of the model is ([19]):

$$D = 4PL^3/(EWH^3)\ . \qquad \text{(Eqn. B.2)}$$

In subsection B.1 we treat the *application* or *instantiation* model of the ODE model Eqn. B.2 as having freely variable values of its geometry and material parameters. This is practical because the instantiation in subsection B.1 has an analytic solution. But in most real problems analytic solutions do not exist, so a numerical solution to a discretized form of the governing ordinary or partial differential equations and specified geometry and boundary and initial conditions must be computed. This often constrains some of the parameter freedoms in the application model. For example, geometries in finite element models are usually fixed instead of parameterized to vary according to actual geometric variations in manufactured devices. We address this very common case of calibration under constrained application model freedom in subsection B.2.

In calibrating the model parameters, the analyst should keep in mind that we will want to use various calibrated parameters in different instantiations of the ODE model Eqn. B.1 to analyze different loading conditions and beam lengths, widths, and heights than in the calibration data

base in section A. For other loading conditions such as distributed loads and one or more point loads not at the end of the beam, the solution Eqn. B.2 will no longer apply, but the model's governing ODE Eqn. B.1 will apply under circumstances discussed later, and can be solved either analytically or numerically for these different loadings.

The existence of an analytic solution Eqn. B.2 allows potential avoidance of some inverse calculations to determine values of calibration parameters and their uncertainty. For example, Eqn. B.2 can be algebraically recast for material property value E on the left hand side as a function of all the other parameters on the right hand side. This could enable direct evaluation for samples of E's uncertainty given the uncertainties of the other variables. Without E separable to the left hand side, iteration would be required to determine samples of E given samples of the other variables' uncertainties. Such separation cannot be done in most real calibration problems, so the analyst is asked to determine samples of the calibration parameter via optimization using the output response variable "forward form" Eqn. B.2 that mimics what would commonly be available from more complex numerical models like finite element models. (Additional information such as direct and adjoint derivatives or sensitivities might also be available from some numerical models, but this is not yet very common in practice so assume this option is not available here.)

The expense of the calibration optimization problems, as measured by the number of model forward runs, can be prohibitive in real cases. So the analyst may wish to demonstrate use of response surface (RS) approximate surrogate models of response as a function of the input variables for quick and inexpensive forward evaluations in the optimization/s. If so, non-negligible effects on the optimized calibration parameter values may be present due to RS approximation. The user should try to assess and effectively eliminate or correct any such RS related effects on the calibrated parameter values, accounting for any remaining uncertainty in the calibrated values. In any case, please keep track of the number of "physics model" forward solutions required (here, number of evaluations of Eqn. B.2).

Two Cases of Supplied Beam Geometry Variability & Uncertainty Information

The analyst is asked to address the following geometry information cases in performing model parameter calibrations as directed in sections B.1 - B.4.

Geometry UQ Case A. Dimensional variations in the large population of beams (hundreds) are controlled and verified in their manufacturing process to meet the following geometric tolerance specifications.

$$\text{Beam Length} = 2. \pm 0.04 \qquad \text{(Eqn. B.3)}$$

$$\text{Beam Height} = 0.2 \pm 0.004 \qquad \text{(Eqn. B.4)}$$

$$\text{Beam Width} = 0.1 \pm 0.002 \qquad \text{(Eqn. B.5)}$$

Geometry UQ Case B. Table B.1 gives additional information on the measured dimensions for the four tested beams in Set 2 in section A, whose experimental conditions and results are to be used in the calibrations prescribed below. Thus, Geometry UQ Case B includes the Case A information in Eqns. B.3-B.5, plus the additional information in tables B.1 and B.2 and Eqns. B.6 - B.11 below.

Table B.1 – Varying cantilever beam dimensions in the four replicate tests associated with Set 2 in Table A.3.

(as measured, subject to meas. errors explained below)	L	H	W
Beam1/Test1	2.026	0.1980	0.1004
Beam2/Test2	2.036	0.2008	0.1015
Beam3/Test3	2.032	0.1978	0.1014
Beam4/Test4	2.028	0.1991	0.1005

Table B.2 – Cantilever beam dimensions for Beam5/Test5 in Figure A.2.

(as measured, subject to meas. errors explained below)	L	H	W
Beam5/Test5	2.026	0. 1987	0.1003

The dimensions in tables B.1 and B.2 are measured values subject to potential measurement errors:

<u>systematic</u> component of measurement error

Beam Length measurement: $-0.01 \leq$ systematic error $\leq +0.01$ (Eqn. B.6)
Beam Height measurement: $-0.001 \leq$ systematic error $\leq +0.001$ (Eqn. B.7)
Beam Width measurement: $-0.001 \leq$ systematic error $\leq +0.001$ (Eqn. B.8)

<u>random</u> component of measurement error

Beam Length measurement: $-0.01 \leq$ random error $\leq +0.01$ (Eqn. B.9)
Beam Height measurement: $-0.001 \leq$ random error $\leq +0.001$ (Eqn. B.10)
Beam Width measurement: $-0.001 \leq$ random error $\leq +0.001$ (Eqn. B.11)

Model Calibration and Prediction Scenarios

For the calibration scenarios in sections B.1 - B.4 below, the experimental conditions and results for Set 2 in section A.3b are to be used, along with any appropriate leveraging of the analyst's UQ processing in that section. For each calibration scenario defined in sections B.1 – B.4, the analyst is asked to perform calibration for geometry specification cases A and B. Does Case B, with its additional more specific (smaller uncertainty) geometry information on the beams tested in the experiments in Set 2 lead to calibrated models that yield smaller uncertainty in exceedance prediction results than the Case A calibrated models?

1. *Model instantiation with relevant parameters as variables.* Here we treat the application model as having freely variable values of its geometry and material parameters. Accordingly, we can take advantage of any auxiliary independent information such as beam dimension uncertainties as defined previously, so that model parameters in the calibration process are ascribed appropriate uncertainties (approximately), according to their correspondence to real/physical uncertainty contributors in the A.3b data being calibrated to.

   The material strength parameter, modulus of elasticity E, is not fundamentally measurable. It is a ***derived*** quantity, determined in tandem with solution of the equation set proposed to represent the beam behavior in specifically designed and controlled experiments for the purposes of such parameter estimation. The experiments in A are used for this purpose here. Assume the beam material responds like a "regular" structural metal to loads in the range applied in section A and is a good candidate for modeling with the force-balance deflection ODE Eqn. B.1.

   The analyst is asked to determine a value or uncertainty description for E along with (possibly changed) final quantifications of the other beam-affiliated parameters/calibration degrees of freedom L, H, W for the following prediction cases a,b,c.

   a) Prediction of end deflection for the whole population of beams. Consider predictions for an applied end load of Po, the nominal loading case investigated experimentally in section A. Quantitative expressions (possibly including uncertainty) for the beam parameters in the model must be arrived at to reflect relevant experimental findings from section A and other information to this point in the document. The analyst can choose any metrics and methods to employ to arrive at suitable parameter expressions. Please sketch the strategy and procedures for arriving at the final values (including any uncertainty descriptions) of the beam-affiliated parameters L, H, W, and E. Provide the parameter values/uncertainties. The analyst is also asked to use these parameter uncertainties in the model to estimate deflection and exceedance probability and associated uncertainties given the target loading value of Po=750,000 discussed in section A. Compare results to the experimental data-based estimates of deflection and exceedance probability from

section A.3b. This provides an indication of the accuracy/conformance of the calibrated model to the experimental results calibrated to. Differences between results inferred directly from the experiments and then from the model calibrated to the experiments should be quantified and commented on, along with any associated implications.

Presumably, the model with appropriate L, H, W, and E characterizations (including uncertainty) from here should be applicable to tip deflection prediction for the population of beams subjected to different loadings, as long as the model ODE Eqn. B.1 can be solved for the loading, and the loading does not change the physics that the ODE and its parameter characterizations adequately capture at the calibration conditions. Unfortunately, the latter is essentially impossible to know without testing calibrated model predictivity at the other conditions. Model validation assessments in section C will give indications for some different application conditions.

b) Prediction of end deflection for one beam to be picked at random from the population. Here a hypothetical beam is picked at random from the population. For predicting uncertainty of its tip deflection to a specified loading, what changes does the analyst foresee in terms of UQ procedures, results, and interpretation vs a) immediately above?

c) Prediction of end deflection for a rectilinear beam of the same material but significantly different dimensions L′, H′, W′ than in the population. The loading may be different from the cases above (not necessarily a point load at the end of the beam) but the loading and geometry of the beam are within a realm where the ODE Eqn. B.1 is reasonable to propose for model predictions. What value or uncertainty description of E would the analyst provide for use with Eqn. B.1 to be applied to the new beam geometry and loading?

2. *Model instantiation with some relevant parameters as fixed values, or absence of independent information on parameter uncertainties.* Here we consider a constraint of set beam dimensions in the calibration model. This is representative of computational models that have fixed geometries instead of parameterized ones. Therefore the beam geometry parameters L, H, W in the model cannot house the variability and/or uncertainty information in Cases A and B. So in this subsection these parameters cannot be assigned uncertainties that approximately correspond to physical uncertainty contributors in the experimental data being calibrated to.

   Even without constraints like the ones here, as a matter of convenience the aleatory and epistemic uncertainties in the calibration data will often be mapped into one or a few selected parameters of the model, regardless of the physical correspondence to the sources of uncertainty in the experiments. This is unavoidable for measurement uncertainties on experimental response because these are not affiliated with any model input parameters.

Furthermore, experimental sources of uncertainty often are not or cannot be characterized to be mapped into corresponding parameters of the model. For example, there could not be an assignment of the geometric uncertainty in Cases A and B to the beam dimensions of the calibration model if this information was not available (i.e., only mean values or nominal estimates were available).

Hence, in this subsection the material strength parameter E is the only calibration degree of freedom to map either or both of epistemic and aleatory uncertainties from experimental sources into. The analyst also has the freedom to dictate the dimensions L, H, W for the beam model to be used in the calibration procedure. The analyst is asked to use these degrees of freedom in whatever way they view will give the best predictions for the same prediction objectives as in B.1 a,b,c. Please provide accompanying reasoning, strategies, procedure summaries, results and interpretation. In the model validation section C an opportunity is given to investigate any degradation of predictivity that might occur from the model calibrated under B.1 freedoms to map experimental uncertainties to related calibration parameters vs. the more constrained B.2 calibration degrees of freedom.

3. *Accounting for effects of model discretization error in the calibration model.* Consider the hypothetical situation where solutions of the model ODE are performed computationally with a discretized finite-element model. Let the discretization effects in the calculations performed to calibrate the model parameters have tip deflection solutions that are biased to a lower computed deflection than a mesh-converged model would predict. Let the bias be 3% such that instead of working with the mesh-converged solution Eqn. B.2 the analyst here obtains solution results from

$$D = 0.97 * 4PL^3/(EWH^3) \; . \qquad \text{(Eqn. B.12)}$$

Note that the magnitude of the discretization related error, which is the deflection result from Eqn. B.12 minus the exact solution Eqn. B.2, varies over the uncertainty ranges (uncertainty space) of the model input variables P, L, E, W, H. Let a solution verification analysis at a selected point in this uncertainty space reveal that discretization-related error in the model biases tip deflections to be less than the converged solution. Let the study provide an estimate that the asymptotic mesh-converged tip deflection is greater than computed deflections from Eqn. B.12 with the following range of uncertainty, to a proclaimed high degree of belief.

mesh-converged deflection = [102%, 105%] of Eqn. B.12 working-mesh defl. (Eqn. B.13)

How does the analyst propose to use the solution verification information captured in Eqn. B.13 at this single evaluation point in the space to account for discretization related error/uncertainty in the prediction model for scenarios B.1a,b,c and B.2a,b,c? Specifically, please demonstrate the approach for B.1c and B.2c for geometry information cases A and B. For Case B, the solution verification evaluation point is partially set by the analyst's choice of the fixed beam dimensions with which the calibration is carried out. Aside from these constraints, at what point or points in the uncertainty space would the analyst select to perform the solution verification analysis?

Keep in mind that the prediction model is chartered to be used to analyze beams of various dimensions and loading conditions, which will involve substantially different discretizations

than used in the calibration calculations. So is not reasonable to assume that the ~same discretizaton error exists between the conditions where the model is calibrated and where it will be used. Therefore the discretizaton error will not systematically cancel out between calibration and usage settings. Also assume the discretization errors and uncertainties are non-negligible as in equations B.12 and B.13. How is this discretization-related uncertainty accounted for in the model calibration procedure and results?

4. Potential Use of Calibration Model to Normalize Experimental Data

A significant complication and source of experimental cost and uncertainty in Section A.3b was the use of auxiliary experimental data at different load 0.5Po to provide load-deflection relationship information for normalizing out the effects of input load variability about the desired service load Po. The theoretically based model might be used instead for the load-deflection relationship to enable data normalization. This might be less complicated than using the auxiliary experimental data at load 0.5Po, and is certainly less experimentally costly. But significant risk might also be involved because the model has not yet been validated, even though it is considered to have a fairly solid theoretical basis.

Consider if the auxiliary experiment at load 0.5Po could not be afforded or conducted in time to move forward in the project. Would the analyst recommend using the model at this early (calibration) stage of its development to normalize the four data points (in Fig. A.1 and tables A.3 and A.4) to an input load of Po, or what alternative approach could or would be used to answer the queries in section A.3b? What uncertainty or caveats would be ascribed to the obtained displacement and exceedance probability results? How do the results compare to those from section A.3b inferred from experimental data alone?

If it was desired by the project to pursue data normalization and the auxiliary experiment at load 0.5Po could be afforded and conducted, can the model be used to reduce the uncertainty in the load-deflection relationship information derived from the experimental data? If so, how would the model be used for this? If the auxiliary experiment would impose a significant cost to the project in time and/or resources, can the analyst make a technical argument that the accuracy traits of the calibrated model reasonably suffice for the specific accuracy needs for data normalization in this situation, such that the experimental cost and time can be avoided with relatively small risk to the project? Explain your reasoning either way.

Items C: Model Validation, Potential associated Adjustment of Prediction Model, and Extrapolative Prediction and Analysis

Background Context

Models are to be used by a beam manufacturer to predict beam tip deflections for a variety of service loadings, temperatures, and rectangular beam dimensions for which experimental characterizations do not exist. Substantial performance and safety consequences could result if

the models used to design beam dimensions and rate/certify them for service temperatures and loading configurations and capacities predict smaller tip deflections than actual physical beams experience. But a competing objective exists. A cost penalty scales with design conservatism. Selling beams of greater strength than truly necessary for given loading and temperature conditions incurs unnecessary expense which reduces market share and thus profits. (Beam strength is controlled by the manufacturer through beam cross-sectional area variables W and H and aspect ratio constraints between these. The customer application dictates the needed beam length L, and the beam material is that in sections A and B.)

Accordingly, strategic model validation assessments are desired to characterize model predictivity and provide usage guidelines to help answer questions like the following. What can the model be used for? Is there a definable parameter space within which the model's predictions are trustworthy? What caveats and/or contextualizations come with the model predictions?

The model parameters have been characterized in a room temperature (20C) calibration setting with beam dimensions and end-loading that are representative or bounding of a large proportion of anticipated service conditions. But some service conditions are expected to extend significantly beyond the conditions where the model parameters were calibrated. Higher service temperatures up to 80C are being contemplated. Restrictions can be placed on this by the manufacturer, but this would reduce the size of the market and hence profits. Past knowledge with generally similar materials and beam/loading configurations indicate that higher temperatures in this regime may significantly lessen beam stiffness, thus increasing deflection to unacceptable levels. Therefore it is considered essential that model predictiveness be tested at a significantly higher temperature than what the beam stiffness material parameter E was calibrated at.

For a total magnitude Q of an integrated load distribution q(x) over the length of the beam, a concentrated point load P=Q at the end of the beam yields greater deflection than any other way the total load Q could be distributed over the beam. So this loading configuration is the most stressing and sensitive one to calibrate models with (as was done in section B), and to validate model predictivity with. A sense of robustness of the model for predicting tip deflection for other loading conditions is also desired; if good predictiveness exists for beams where loading is not concentrated at the end of the beam (a significant portion of the market), then profits can be increased by competitively serving this market with appropriately smaller, more economical beam cross-sections.

Besides these objectives, constraints also exist for the validation assessments. Among the more severe constraints, only three validation tests can be conducted because of resource limitations. The beam dimensions for these validation tests are controlled very tightly to the prescribed values (measurement error/uncertainty is negligible). This makes them relatively costly. Furthermore, tests at elevated temperatures above room temperature (20C) are difficult and costly and can only be conducted up to 60C, short of the anticipated market range to 80C.

Beams up to length L* are available from the manufacturer. These are nominally 10% longer than the popular high-selling beams calibrated to. For a beam length L* and end-load Po, manufactured height and width are set to H* and W* (Table C.1) using industry-standard aspect ratios for maximum recommended deflection-to-length (DTL) allowables at room temperature,

20C. (H* and W* for the given L* are determined from an extensive data base of beam dimension combinations that meet what is considered to be a "safe" maximum DTL ratio of 20% for an end-load Po as established by extensive testing of 20C beams of a very similar material alloy to that of the manufacturer's beams.)

Table C.1 – Dimensions of Beams in Validation Experiments (negligible dimensional variability and measurement errors)

L*	W*	H*
2.20	0.09292	0.18580

Recall that loading of magnitude Po placed at the end of the beams in the calibration activity is a representative upper extreme of allowable loading for beams in this class. The manufacturer also wants to get new experimental evidence and assess model predictivity for tip deflection with the long-beam dimensions (Table C.1) and for the representative upper extreme of deflection given the maximizing case of placing Po at the end of the beam. Ultimately, the manufacturer wants to determine the probability or uncertainty concerning whether the "safe" maximum DTL ratio is met for an end-load Po on the beams. This case effectively bounds the risk of using beams loaded otherwise with integrated load Po. For beam length L* = 2.2 (Table 3), the maximum DTL ratio of 20% corresponds to the following maximum allowable end-deflection.

$$D_{critical} = 0.44 \qquad \text{(Eqn. C.1)}$$

Design of Validation Experiments

The following validation experiments are arrived at in view of the relevant objectives and constraints. A useful model validation hierarchy would attempt to evaluate model predictivity along the following phenomenological aspects that would be useful to resolve independently before considering them jointly for conditions involving mixtures of these aspects.

Phenomenological aspects α are whether the ODE Eqn. B.1 with relevant beam parameter values from the calibrations in section B extend robustly to other beam dimensions and loading configurations at 20C (the calibration temperature).

Phenomenological aspect β is the suspected significant temperature dependence of material stiffness over the contemplated temperature range.

If significant deficiencies in any of these aspects individually are quantified in the validation assessments, then perhaps reasonably robust (extrapolable) model adjustments can be made for each aspect individually, which might approximately scale to conditions where mixtures of these aspects exist. If instead, validations are conducted for experimental mixtures of these aspects, this confounds the effects of any model deficiencies in the individual aspects. Then any

deficiencies associated with each aspect cannot be identified individually to be potentially addressed.

A choice could be made to use beams/dimensions from the calibration population for the validation experiments. Then the validation conditions could entail separate-effect perturbations from the calibration baseline in terms of exploring predictability under very different loading configuration at the same temperature, and in terms of a very different temperature but the same loading configuration. This would be relatively clean for hierarchical validation with respect to these beam loading and temperature factors. But other considerations trump this clean separation of these two physical aspects, as discussed next.

With the three allotted tests, it is decided to use two to assess predictive robustness to the temperature aspect and one to assess the different loading aspect—and to do so with the L*,W*,H* beam, as the longest in the manufacturer's design space and previously untested experimentally. If a large temperature change significantly changes the nominal stiffness of the beam material, it could also significantly change the material's range of stochastic stiffness variation. Two replicate tests with end-loaded beams are dedicated to experimentally explore material variability effects at the extreme testable temperature of 60C and extreme loading case of end-load Po. Given also that the longest beam length L* is involved, any model adjustment with respect to the phenomenological aspects in the categories α and β above will presumably be maximum or nearly so. Then model adjustments could be appropriately down-scaled at less severe factor combinations in the prediction space (or not down-scaled, depending on analyst reasoning).

The remaining experiment is used to assess predictivity at 20C under a significantly different loading configuration and beam geometry than in the calibrations at 20C. This addresses the two items in category α simultaneously, rather than individually, because of the constraint of only one experiment available. But relatively little may be lost, as explained below. For this experiment, a uniform load distribution q(x)=constant is applied downward along the length of the beam, with integrated magnitude Po. An analytic solution to the beam model Eqn. B.1 is supplied by equation for this loading. Stochastic stiffness variation of the beam material will not change with this loading case compared to the calibration loading, but deflection variability could change significantly. Nonetheless, it is reasoned that experimentally sampling the potentially very different material stiffness variability effects at 60C is more important than using multiple tests to sample the deflection variability for a less-extreme uniform-loading case with material stiffness variability that is the same as in the calibrations.

This validation plan does not provide for a clean assessment of model predictivity with respect to all the individual phenomenological aspects listed in categories α and β. In Validation Setting A described below, beam loading and dimensions are different from the calibration baseline. If model performance does not significantly degrade in validation setting A (and high resolution exists in the validation activity to establish this), then confounding is probably not important: it can be reasonably argued that neither loading nor dimension changes degrade model predictiveness significantly because it is highly unlikely that they have individual degradation effects of significant magnitude but counteract each other ~exactly. But if model performance is detected to significantly degrade, it will not be known whether this is from the geometry change from baseline calibration conditions, or is the impact of the beam loading change, or both. In

Validation Setting B, the end-loading configuration is the same as in the calibration baseline, but beam temperature and geometry are both different from the calibration baseline. So any significant degradation in model predictive performance would not be cleanly attributable to geometry change or temperature change factors.

Thus, geometry changes substantially complicate attribution of any model predictive degradation to beam loading and/or beam temperature. But with this tradeoff, greater sensitivity (greater expected deflections than with the beam dimensions of the calibration baseline) is gained in the validation assessments, and a severe edge of the domain of customer interest is experimentally assessed and extrapolative performance of the calibrated model is more severely tested.

If the analyst sees a strong rationale for changing one or more of these three budgeted validation tests, please describe the rationale, advantages and disadvantages, and any recommended changes to the tests.

The analyst is asked to perform validation assessments at the validation conditions A and B outlined below, in whatever order deemed most appropriate—with any reasoning stated for this. This is to be done for the following two tracks of calibration parameters used in the prediction model to be validated:

- Track 1 - use the parameters calibrated in section B.1c for geometry UQ Case B (experimental beams' geometry variations and measurement uncertainties explicitly modelable in the calibration)

- Track 2 - use the parameters calibrated in section B.2c for geometry UQ Case B (experimental beams' geometry variations and measurement uncertainties not explicitly modeled in the calibration--fixed geometry assigned to all experimental beams)

For each track, describe the validation procedure, results, and conclusions at validation settings A and B. Uncertainties regarding the particular realizations of beam material property E, BCs, measurement errors, etc. in the experiments exist and should be accounted for in the validation assessments.

Validation Setting A: One uniformly loaded 20C beam of total load magnitude Po.

This case has a uniform load distribution q(x) = constant = q = Po/L* applied downward along the length L* of the beam with dimensions in Table C.1. An analytic solution to the proposed governing equation Eqn. B.1 for this loading is ([19]):

$$D = 3qL^4/(2EWH^3) \; . \qquad \text{(Eqn. C.2)}$$

To add model discretization error/uncertainty to the validation problem, in analogy with Eqns. B.12 and B.13 the following is given for discretization-related prediction errors for validation setting A. These should be included in the validation analyses.

$$D = 0.98 * 3qL^4/(2EWH^3) \quad \text{calculated with nominal mesh} \quad \text{(Eqn. C.3)}$$

$$\text{mesh-converged deflection} = [101\%, 104\%] \text{ of Eqn. C.3 w/nominal mesh} \quad \text{(Eqn. C.4)}$$

Table C.2 presents the measured deflection in the validation experiment. The deflection measurement is subject to uncertain systematic and random measurement errors per Eqns. A.1 and A.2. The target load of Po stated above Fig. A.2 is exactly met in this experiment, with no measurement error.

Table C.2 – Beam deflection measurement, subject to random and systematic measurement errors per Eqns. A.1 and A.2.

	deflection D (subject to potential systematic and random meas. errors in these results)
ValTest X	0.09768

Validation Setting B: Two Beams at 60C with Target End-Load Po.

This case involves the same target loading Po as in the calibration exercise, but the validation experiments are performed at an elevated temperature of 60C and the beams have dimensions given in Table C.1. The discretization-related prediction error and uncertainty for validation setting B are:

$$D = 0.97 * 4PL^3/(EWH^3) \quad \text{calculated with nominal mesh} \quad \text{(Eqn. C.5)}$$

$$\text{mesh-converged deflection} = [102\%, 105\%] \text{ of Eqn. C.5 w/nominal mesh.} \quad \text{(Eqn. C.6)}$$

It is conceivable that this mesh related uncertainty could be non-negligibly correlated with that in the calibration section if care is taken to mesh the beams "sufficiently" and "similarly". How might these criteria that would possibly strengthen correlation between the discretization related prediction uncertainties in the calibration and validation settings be more concretely defined and implemented? Would the presence of strong correlation affect the analysts calibration and/or validation procedures?

Table C.3 presents the measured deflections and end loads in the validation experiments. The deflection measurement is subject to uncertain systematic and random measurement errors per Eqns. A.1 and A.2. The load measurements are subject to uncertain systematic and random measurement errors per Eqns. A.3 and A.4.

Table C.3 – Beam deflection and End-Load measurements, subject to random and systematic measurement errors as described in the text.

	deflection D (subject to potential systematic and random meas. errors in these results)	**end load P (subject to potential systematic and random meas. errors in these results)**
ValTest Y	0.3880	7.769E5
ValTest Z	0.3840	7.390E5

Prediction Analysis Cases and Potential Adjustment of Prediction Model

The analysis is asked to address the following issues for model Tracks 1 and 2 separately.

From the results of model validation activities A and B, the analyst is asked to decide whether the model should accepted, rejected, or adjusted for the purpose of making tip deflection predictions over the desired 20C – 80C temperature range and for general loadings and beam geometries within the extreme case of concentrated end-load Po and beam geometries like the calibration beams and L*,W*,H* beams whose geometries are figured from this extreme loading case and the maximum DTL guideline of 20%.

Alternatively, the modeler may choose to characterize things in terms of a range of model-use space that the prediction model can be concluded "good enough" or "trustworthy" or "useful" (or however the analysts frames this acceptability or adequacy issue) in a technically defensible and usable way as explained by the analyst. In any case, what adjustments, if any, would the analyst make to the prediction model? How would this change the predictiveness characterization, range/space of acceptable model-use, trustworthiness, etc. versus an un-adjusted prediction model? What "confidence", "credibility", uncertainties, caveats, etc. should be assigned to the

adjusted or un-adjusted prediction model taken forward, and/or to its predictiveness and predictions, as a function of location in the prediction parameter space?

In particular, address these issues for predicting beam end deflection and deflection exceedance probability at the following points in the space:

- at validation settings A and B, before any validation-informed adjustments to the prediction model, and after adjustment if the analyst chooses to adjust. The effect of the latter on the prediction quantities of deflection and exceedance probability should be noted as measures of the predictive change contributed by model adjustment due to any model-form error or potential error quantified by the validation assessment.

- at 80C (extrapolation beyond the validation extreme of 60C) for the beam geometry and loading conditions in validation settings A and B. Extrapolating involves another source of error and uncertainty. The analyst is asked to explain and demonstrate their strategy, if any, for addressing extrapolation error and uncertainty.

Please comment on any degraded model predictivity that occurs from the model parameters calibrated under the Track 1 vs. the Track 2 conditions.

Finally, from a design perspective, what temperature would be considered "safe" for beams in this class that are sized according to the maximum DTL guideline of 20% if "safe" means a population of beams that meet this guideline with 99.9% reliability? What is the uncertainty associated with this "safe" temperature?

APPENDIX B: INTERVIEW QUESTIONS

Note: Due to the technical nature of the work discussed in these interviews, a significant number of follow-up questions were asked verify our understanding of the participants' responses.

Pre-solution interview

- Relevant background and experience
    - *What technical education and experience(s) of yours do you expect to be of particular relevance in solving this problem?*
- Planning
    - *What solution strategies for working the problem occur to you?*
    - *Which strategy do you intend to follow, and why?*
        - *What expectations do you have concerning your overall strategy?*
        - *In which aspects are you more confident, or more uncertain?*
        - *What will your initial step be? What alternatives are there, and what is the reasoning behind your preference?*
            - *Do you have any expectations as to what this initial step will reveal?*
            - *If so, what is your confidence in these expectations?*
- Solution
    - *Can you, at this point, put any kind of bounds on the answer you expect to identify?*
        - If so, what confidence do you have in these bounds?
- Other
    - *Any other impressions to share?*

During-solution interview(s)

- Work to this point
    - *What work have you done on the problem since the last interview? What have been your findings?*
        - To what extent have your findings matched your expectations? Have their been any surprises?
- Planning
    - *What is your next step? What alternatives are there, and what is the reasoning behind your preference?*
        - *Do you have any expectations as to what this initial step will reveal?*
            - *If so, what is your confidence in these expectations?*

    - *What work do you expect is remaining before you identify a solution?*
        - *What confidence do you have in this expectation?*
- Solution
    - *Can you, at this point, put any kind of bounds on the answer you expect to identify?*
        - *If so, what confidence do you have in these bounds?*
        - *Does this projection (if any) differ from any earlier projection, and if so, why?*
- Other
    - *Any other impressions to share?*

Post-solution interview

- Work to this point
    - *What work have you done on the problem since the last interview? What have been your findings?*
        - *To what extent have your findings matched your expectations? Have their been any surprises?*
- Solution
    - *What confidence do you have in the solution you have identified?*
- Reflection
    - *What would you do differently (if anything) if you were to begin work now on a similar problem?*
    - *What would you do (if anything) if you had significantly more time to invest in solving the current problem?*
- Other
    - *Any other impressions to share?*

Post-sharing interview

- Solution
    - *Has your confidence in the solution you have identified changed in any way? If so, how?*
- Reflection
    - *Any impressions to share about the other solutions? Were any aspects of others' proposed solutions surprising to you?*

- *What would you do differently (if anything) if you were to begin work now on a similar problem? Any changes from before, and if so, why?*
- *What would you do (if anything) if you had significantly more time to invest in solving the current problem? Any changes from before, and if so, why?*

DISTRIBUTION

1	MS0899	Technical Library	9536 (electronic copy)
1	MS0359	D. Chavez, LDRD Office	1911


Sandia National Laboratories