DOT/FAA/AM-24/17

Office of Aerospace Medicine

Washington, D.C. 20591

Assessing the Concepts of Best Estimate Plus Uncertainty for FAA Aircraft Seat Certification

Author:

Martin Pilch

Technical Reviewer:

Joseph E. Pellettiere

Civil Aerospace Medical Institute (CAMI)

Federal Aviation Administration

Oklahoma City, OK 73169

NOTICE

This publication and all Office of Aerospace Medicine technical reports are available in full text from the Civil Aerospace Medical Institute's publications website and at the National Transportation Library's Repository & Open Science Access Portal.

Technical Report Documentation

1. Report No.
DOR/FAA/24/17

2. Title & Subtitle
Assessing the Concepts of Best Estimate Plus Uncertainty for FAA Aircraft Seat Certification

3. Report Date
9/1/2024

4. Performing Organization Code
AAM-632

5. Author(s)
Martin Pilch

6. Performing Org Report Number
MPC2024-1

7. Performing Organization Name & Address
FAA Civil Aerospace Medical Institute
P.O. Box 25082
Oklahoma City, OK 73125

8. Contract or Grant Number
6973GH-22-D-00062

9. Sponsoring Agency Name & Address
Aircraft Certification Service (AIR)
Federal Aviation Administration
800 Independence Ave., S.W.
Washington, DC 20591

10. Type of Report & Period Covered
Technical Report

11. Supplementary Notes
This report underwent peer review: William Carroll, David Moorcroft, and Joseph Pellettiere
Technical report DOI: https://doi.org/10.21949/1529682

12. Abstract

This report synthesizes a process for Best Estimate Plus Uncertainty (BEPU) based on lessons learned from the United States (US) Nuclear Regulatory Commission, the US Nuclear Weapons Community, and the emerging Verification, Validation, and Uncertainty Quantification (VVUQ) communities. BEPU supports Certification by Analysis (CbA). BEPU is characterized by an unbiased and objective accounting of all dominant contributors to the uncertainty that informs regulatory decisions. Predictions for untested designs must be bias-corrected for errors in the simulation computational model and model form, and uncertainties in both must be quantified to support regulatory decisions. The process is demonstrated for emergency landing conditions with an application to aircraft seat certification for lumbar injuries. The demonstration includes the steps necessary for certification, identifies what testing is needed, and shows how the model can be accepted. The report discusses four regulatory options for dealing with uncertainties that balance risk tolerance with the maturity of assessment capabilities.

13. Key Word
Best Estimate Plus Uncertainty (BEPU), Quantification of Margins and Uncertainties (QMU), Certification by Analysis (CbA)

14. Distribution Statement
Document is available to the public through the National Transportation Library: https://rosap.ntl.bts.gov

15. Security Classification (of this report)
unclassified

16. Security Classification (of this page)
unclassified

17. No. of Pages
232

18. Price
N/A

Acknowledgments

I want to express my gratitude to everyone who contributed to this report's conceptualization and completion. I thank Joseph Pellettiere, Chief Scientist and Technical Advisor for Crash Dynamic at the Federal Aviation Administration (FAA), for his collaboration and contractual support during this project. I would also like to thank David Moorcroft and William Carroll, FAA Civil Aerospace Medical Institute (CAMI), for their technical support, especially in conducting material characterization tests for this project. I appreciate LaDonna Wilson, Chickasaw Health Consulting, for her contract management, especially for contractual accommodation when I was blindsided by unplanned life events. I am deeply thankful to my colleagues Timothy Trucano, William Oberkampf, Kevin Dowding, William Rider, Thomas Paez, and Benjamin Blackwell for their technical and inspirational collaborations in refining and expanding the principles of Verification, Validation, and Uncertainty Quantification (VVUQ) and their application to computational simulation, which informs high-consequence regulatory decisions. Lastly, I want to acknowledge my wife, Veronica Montaño-Pilch, and our children: Joseph, Martín, Romaldo, Joshua, Angela, and Andrew. You provided understanding, patience, encouragement, and belief in my abilities throughout my professional career. Our love provides meaning to everything we do. I am proud of you!

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A2A	Analyst-to-Analyst
AC	Advisory Circular
AF	Airflex, a type of polyurethane foam
AHP	Analytical Hierarchy Process
AIAA	American Institute of Aeronautics and Astronautics
ASC	Advanced Simulation and Computing
ASME	American Society of Mechanical Engineers
ASTM	American Society for Testing and Materials
ATD	Anthropomorphic Test Dummy
AVS	Office of Aviation Safety
BC	Boundary Condition
BE	Best Estimate
BEPU	Best Estimate Plus Uncertainty
CAMI	Civil Aerospace Medical Institute
CASL	Consortium for the Simulation of Advanced Light Water Reactors
CF	Confor, a type of polyurethane foam
CFD	Computational Fluid Dynamics
CFR	Code for Federal Regulations
COV	Coefficient of Variation
CPU	Central Processing Unit
CSAU	Code Scaling and Uncertainty
CVER	Code Verification
DAX	DAX, a type of polyurethane foam
DCH	Direct Containment Heating
DP	Design Point
DRI	Dynamic Response Index
EASA	European Union Aviation Safety Agency
ePIRT	Extended Phenomena Identification and Ranking Table
FAA	Federal Aviation Agency
FS	Factor of Safety
G2G	Gauge-to-Gauge
GCI	Grid Convergence Index
GPRA	Government Performance and Results Act
HALT	Highly Accelerated Life Testing
HIC	Head Injury Criterion
HOT	Higher Order Term
HT	Heat Transfer
IC	Initial Condition
IEEE	Institute of Electrical and Electronics Engineers
IET	Integral Effects Test
IFD	Indentation Force Deflection
IR	Injury Rate
ISBN	International Standard Book Number
ISO	International Standardization Organization
LHS	Latin Hypercube Sampling

LT	Lower Torso
MC	Monte Carlo
MGW	Maximum Gross Weight
MMAC	Mike Monroney Aeronautical Center
MMS	Method of Manufactured Solutions
MPC	MPilchConsulting
MS	Microsoft
MTOW	Maximum Takeoff Weight
M&S	Modeling and Simulation
NAFEMS	National Agency for Finite Element Methods and Standards
NaRC	National Research Council
NASA	National Aeronautics and Space Administration
NATO	North Atlantic Treaty Organization
NIAR	National Institute for Aviation Research
NNSA	National Nuclear Security Administration
NQA	National Quality Assurance
NRC	Nuclear Regulatory Commission
ODE	Ordinary Differential Equation
PCMM	Predictive Capability Maturity Model
PIRT	Phenomena Identification and Ranking Table
PRA	Probabilistic Risk Assessment
PWR	Pressurized Water Reactor
QMU	Quantification of Margins and Uncertainties
QOI	Quantity of Interest
QRA	Quantitative Risk Assessment
RMS	Root Mean Square
RTS	Regression Test Suite
S2S	Sample-to-Sample
SAE	Society of Automotive Engineers
SET	Separate Effects Test
SME	Subject Matter Expert
SNL	Sandia National Laboratories
SOW	Statement of Work
SQA	Software Quality Assurance
SQE	Software Quality Engineering
SRQ	System Response Quantity
SRS	Shock Response Spectrum
STS	Sustainability Test Suite
SVER	Solution Verification
TMI	Three Mile Island
TRL	Technology Readiness Level
TRLS	Technology Readiness Levels
U	Uncertainty
UB	Upper Body
UQ	Uncertainty Quantification
USNRC	United States Nuclear Regulatory Commission
UT	Upper Torso

VERTS	Verification Test Suite
VV	Verification and Validation
V&V	Verification and Validation
VVUQ	Verification, Validation, and Uncertainty Quantification

Biography of the Author

Dr. Martin Pilch earned a Ph.D. in Nuclear Engineering (1981) from the University of Virginia. Currently, MPilchConsulting specializes in establishing the credibility of computer simulations, verification, validation, uncertainty quantification, and risk assessment for engineering applications spanning the full spectrum of engineering disciplines. (Pilch, 2019) presented an invited ASME keynote, *Cautionary Tale When Using Computational Simulation to Support Regulatory Decisions – Recommendations to Mitigate the Risk*, which was subsequently presented by invitation at Sandia National Laboratories (twice), by invitation once at Johnson and Johnson, by invitation to USNRC senior management, and a second time to staff (over 200 attendees participated in the presentations). The ASME keynote was based on previous experience and material for FAA-invited talks presented at dynamic modeling and simulation workshops in 2017 and 2018. The ASME keynote was the basis for a third FAA-invited talk to EASA in 2019. Pilch gained familiarity with (AC) 20-146 as part of these FAA workshops.

Pilch recently consulted with the CASL Program at Oak Ridge National Laboratory to define and demonstrate a more rigorous code and solution verification methodology for a suite of four codes developed to support performance improvements in existing nuclear power plants. The code capabilities spanned thermal hydraulics, fuel performance, and neutron transport and resulted in three journal publications (Porter et al., 2020a), (Toptan, Porter, Hales, Spencer, et al., 2020), and (Wang et al., 2022). The emphasis was on quantifying numerical errors through systematic convergence studies and developing a matrix that communicated physics-based code verification coverage for the user community and identified gaps to prioritize new verification efforts by the code development team.

Pilch retired in 2016 from Sandia National Laboratories after 35 years of service, having held distinguished staff, management, and program positions. Pilch played a significant role in the first large-scale integration of high-end modeling and simulation with more traditional testing for a nuclear weapon life extension program. Pilch championed the adoption of QMU in the nuclear weapon program and was a charter member of the Laboratories' QMU steering committee. Pilch led SNL/ASC activities supporting an NNSA L1 milestone and related L2 milestones demonstrating VV&UQ principles and processes for high-fidelity weapon safety calculations, including the first demonstration of CompSim- based QMU. Pilch led MTE1 (QMU Methodology) for the Advanced Certification Campaign.

Pilch managed the QMU and Management Support Department and represented the Laboratories position on QMU to the JASONs, SAGSAT, NNSA, National Academy of Sciences (Review of QMU), and the Predictive Engineering Sciences Panel. Pilch mentored multiple classified applications of QMU at Sandia. Some key references highlighting philosophical, programmatic, and methodological leadership, including guest editorship of an issue of Reliability Engineering & System Safety dedicated to

QMU, are (Diegert et al., 2007) (Helton & Pilch, 2011) (Pilch et al., 2011), and (Helton et al., 2004).

For six years, Pilch previously managed the Verification and Validation (V&V) sub-element of the Advanced Simulation and Computing Program at Sandia and was a line manager of the Validation and Uncertainty Quantification Department in the Engineering Sciences Center. Pilch sponsored the Validation Challenge Workshop (Hills et al., 2008), which resulted in a dedicated issue for the journal. One of Pilch's staff members, William L. Oberkampf, authored the definitive book on verification and validation of scientific computing (Oberkampf & Roy, 2010), which is based partly on the work products of the V&V program that Pilch managed.

Pilch spent the first nineteen years of his career developing and validating models for severe accident issues associated with the operation of nuclear power plants. During this time, he participated in and led significant activities using a *risk-informed* approach, which integrated modeling and experiments in a probabilistic framework to address and resolve safety issues arising from the accident at Three Mile Island. These efforts resulted in a dedicated issue of the journal Nuclear Engineering and Design and two journal articles (Pilch et al., 1996) and (Pilch & Allen, 1996). This risk-informed approach was a second (tailored) application of the NRC CSAU methodology, the NRC's first successful application of Best Estimate Plus Uncertainty (BEPU) for regulatory purposes. Through continuous peer review, Pilch was mentored by critical authors of the CSAU methodology.

Abstract/Executive Summary

This report synthesizes a process for utilizing Best Estimate Plus Uncertainty (BEPU) based on lessons learned from the United States (US) Nuclear Regulatory Commission, the US Nuclear Weapons Community, and the emerging Verification, Validation, and Uncertainty Quantification (VVUQ) communities. BEPU supports Certification by Analysis (CbA).

BEPU is characterized by an unbiased and objective accounting of all dominant contributors to the uncertainty that informs regulatory decisions. Decision criteria are the only place where conservatism is encouraged. BEPU supports Certification by Analysis (CbA). Predictions for untested designs must be bias-corrected for errors in the simulation computational model and model form, and uncertainties in both must be quantified to support regulatory decisions.

Eleven potential sources of simulation solution errors and uncertainties are identified, which are applicable to any computational model. Processes and best practices that allow their identification, quantification, and management are discussed. Simulation results must be bias-corrected for known sources of simulation solution errors before being used for model validation or regulatory predictions.

Errors and uncertainties associated with model form are also addressed. Model form errors and uncertainties are assessed by systematically comparing model predictions with a hierarchy of relevant test data. CbA always involves interpolation or extrapolation, and it is critically important that the simulation model correctly predicts trends with variations of design and environment parameters. Testing errors and uncertainties distort and cloud the assessment of model accuracy. Ten potential sources of testing errors and uncertainties are identified. Processes and best practices that allow their identification, quantification, and management are discussed. Test results must be bias-corrected for known sources of testing errors before being used to assess model accuracy.

The process developed here is demonstrated for emergency landing conditions with a specific application to aircraft seat certification for lumbar injuries. The demonstration includes the steps necessary to demonstrate certification, identifies what testing is needed, and shows how the model can be accepted.

The report concludes with comments and recommendations on implementing BEPU. It discusses four regulatory options for dealing with uncertainties that balance risk tolerance with the maturity of assessment capabilities.

1. Introduction

1.1 FAA objectives and project tasking

The Civil Aerospace Medical Institute (CAMI) is the medical certification, education, research, and occupational medicine wing of the Office of Aerospace Medicine (AAM) under the auspices of the Federal Aviation Administration (FAA's) Office of Aviation Safety (AVS). The mission of the Aerospace Medical Research and Safety Assurance Division of CAMI (AAM-600), located at the Mike Monroney Aeronautical Center (MMAC) in Oklahoma City, OK, is "to develop new and innovative ways to support FAA regulatory and advisory missions to improve the safety of humans in civilian aerospace operations." The Aerospace Medical Research and Safety Assurance Division has a laboratory that assesses occupant protection in aviation environments. The Biodynamics Research Team (AAM-632) provides state-of-the-art information, procedures, and equipment evaluations concerning aircraft accident investigation and survivability during normal operations and emergencies.

The use of computational modeling and simulation (M&S) data in the certification of aircraft seats was identified as a key component of streamlining seat certification in 2000 (Wendell H. Ford Aviation Investment and Reform Act for the 21st Century. Public Law 106-181, House Resolution 1000, 2000). This led to the publication of FAA Advisory Circular (AC) 20-146 *Methodology for Dynamic Seat Certification by Analysis for Use in Parts 23, 25, 27, and 29 Airplanes and Rotorcraft* in 2003 and subsequent revision in 2018. Because the field of Certification by Analysis (CbA) for aircraft components was considered underdeveloped, the AC contains safety factors that some consider overly restrictive and prescriptive.

CAMI has initiated a project to expand the use of computational modeling and simulation in cabin safety applications. The project will focus on increasing the use of M&S for certifying aircraft seats and other cabin safety components. The former will be accomplished by evaluating whether more performance-based rules can supplant the current prescriptive guidance. The engineering study documented in this report is the first step in the project. It will provide the information necessary for the FAA to evaluate the use of Best Estimate Plus Uncertainty (BEPU) and Quantified Margins and Uncertainty (QMU) to remove the rigid factors of safety in current FAA guidance.

Key deliverables include:

1. Define a process for using QMU information in the form of BEPU to certify an aircraft seat using computational simulation,
2. Demonstrate the process by a detailed example, which will include the steps necessary to demonstrate certification, identify what testing is needed, and show how the model can be accepted,
3. Develop a training package using the approved process that the FAA will use to train FAA engineers and the industry, and
4. Deliver a final technical report.

Figure1.1 summarizes the FAA's current processes for Certification by Testing (CbT) and Certification by Analysis (CbA) for aircraft seats. CbT is based on the results of a single sled test, even when variability in the form of repeatability and reproducibility is known to exist. Industry bears the risk that a result of a single test would lead to rejection of a design when another nominally identical test might lead to acceptance. The FAA bears the risk that the result of a single test would lead to acceptance of a design when another nominally identical test might lead to rejection of the design. Both industry and the FAA accept these risks in CbT.

The FAA allows CbA as an alternative to CbT for a proposed new design that is near a baseline seat design already certified by testing. However, “safety factors” intentionally stack the deck against CbA. CbA is held accountable for test variability, while CbT is not. This is reflected in reducing the acceptance load of 1500 lb_f for CbT to 1430 lb_f for CbA.

CbA also requires validating the simulation model by comparing predictions to measured lumbar loads for the baseline seat design. If the discrepancy is sufficiently small, $|E_{rel}| < 10\%$, then the model can be used to predict lumbar loads for the nearby design seeking certification. However, predictions for the nearby design must be bias-corrected if the model underpredicts lumbar loads for the baseline design. Otherwise, the applicant must conservatively live with over predictions. However, bias correction of predictions is ambiguous when the discrepancy is based on the difference between a single prediction and a single test result. The discrepancy could reflect either model form error or test variability. Model form error is a candidate for bias correction, while test variability is already accounted for in the reduction of acceptable lumbar loads for 1500 lb_f to 1430 lb_f

The impact of embedded conservatism and the acceptance of known sources of error is hard to quantify and diminishes the value proposition of CbA. The BEPU process developed here will add the necessary formalism when using computer simulation to inform regulatory decisions.

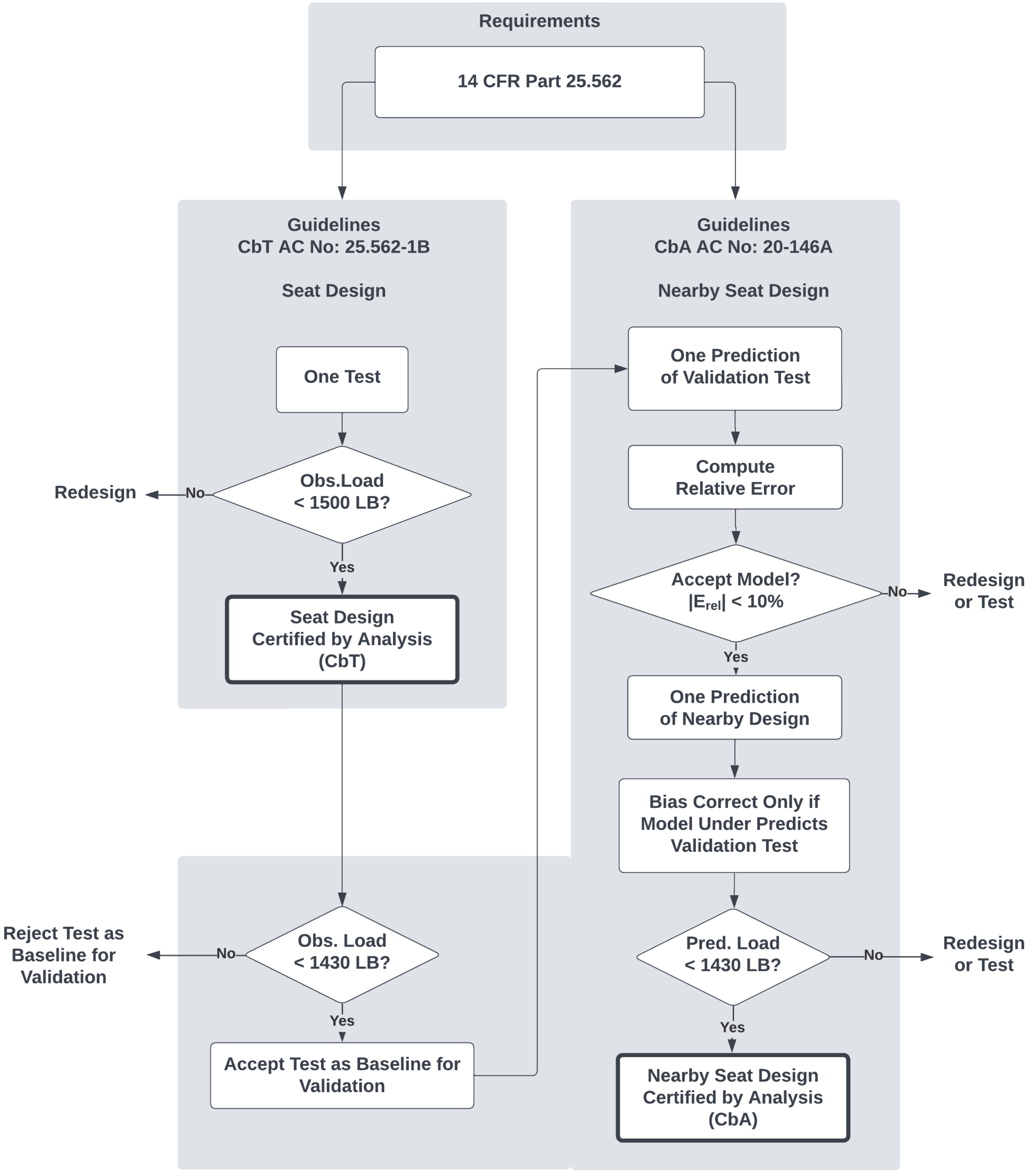

Requirements
14 CFR Part 25.562
Guidelines
CbT AC No: 25.562-1B
Seat Design
One Test
Obs.Load < 1500 LB?
No
Redesign
Yes
Seat Design Certified by Analysis (CbT)
Obs. Load < 1430 LB?
No
Reject Test as Baseline for Validation
Yes
Accept Test as Baseline for Validation
Guidelines
CbA AC No: 20-146A
Nearby Seat Design
One Prediction of Validation Test
Compute Relative Error
Accept Model? |Erel| < 10%
No
Redesign or Test
Yes
One Prediction of Nearby Design
Bias Correct Only if Model Under Predicts Validation Test
Pred. Load < 1430 LB?
No
Redesign or Test
Yes
Nearby Seat Design Certified by Analysis (CbA)

Figure1.1: Current FAA seat certification processes

A simple example will demonstrate the BEPU process and its key concepts. Table 1.1 summarizes the specifications for a hypothetical baseline seat design that was certified by testing. An existing sled test, CAMI test A15008, will be used for the baseline design that was certified by testing. A nearby design seeking CbA has a slightly thicker cushion, all other specifications being the same. Certain features of the demonstration example were chosen to facilitate solutions in MS Excel without the need for finite element modeling. This makes a demonstration of the process practical and transparent.

Table 1.1: Simple example for demonstration of the BEPU process

	Seat Design	**Nearby Design**
Requirements	14CFR25.562	14CFR25.562
Load <	1500 lb_f	1500 lb_f
Environments	14CFR25.562	14CFR25.562
Triangular Pulse Max G	14	14
Rise Time (ms)	80	80
Impact Angle	30^{o}	30^{o}
Passenger	14CFR25.562	14CFR25.562
ATD Weight	170 lb	170 lb
ATD Positioning	FAA-Hybrid III Seated Upright	FAA-Hybrid III Seated Upright
Seat Design		
Frame	Rigid	Rigid
Seating	Single	Single
Monolithic Cushion	CF42 (AC)	CF42 (AC)
Cushion Thickness	2.0”	2.5”
Results	**Test**	**Predicted**
Test	CAMI A15008	
Lumbar Load	1048 lb_f	
Decision Metric		
FoS = L_{req}/L	1.43	

Lessons learned in the application of BEPU will be reviewed in Section 0 to identify essential elements that should be represented in the proposed process. Section 0 will present the proposed process, summarizing the process elements. Section 0 will present key concepts of each process element in detail and demonstrate their application for aircraft seat certification. Section 0 will make recommendations for the future that will align the certification process with the principles of BEPU, strengthening support for performance-based regulation. The FAA will receive an abbreviated training package in PowerPoint slides. This technical report will contain additional background and technical details inappropriate for a training package.

1.2 Value proposition and risks of computational simulation

The value of computational simulation for industry has grown exponentially in recent years, and the trend will not change. This is made possible by the dramatic increase in computing capabilities, as illustrated in Figure 1.2. I received a slide rule in my first-year kit when I entered engineering school in 1970; however, computing capabilities increased by 18 orders of magnitude within one generation. Only a select few have access to the first-of-its-kind capability computers[1]. Still, within about six years, the top 500 organizations adopted that first-of-its-kind capability and became capacity computing[2] a short time later. My laptop, on which I do all my computing for this project, is the equivalent of a supercomputer only 25 years ago. The capabilities of engineering software have also grown exponentially, taking advantage of this computing horsepower.

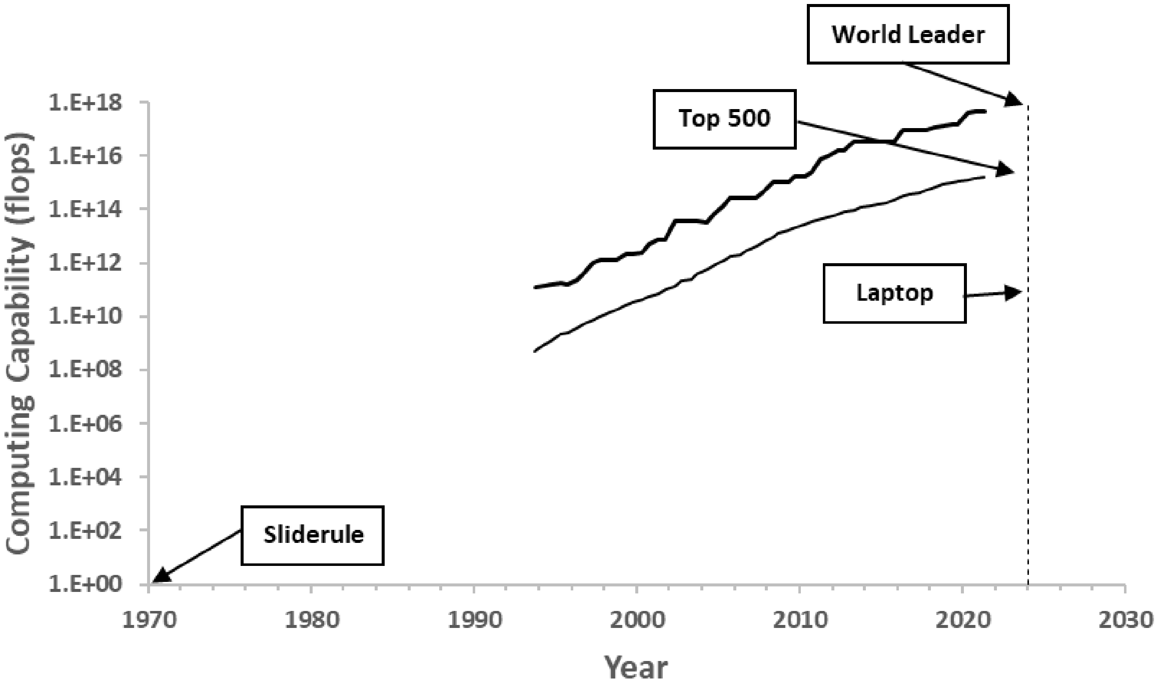


Figure 1.2: Explosion of computing capabilities

Increased reliance on computational simulation challenges regulatory agencies to ensure the credibility of simulation results is appropriate for regulatory decisions. I define credibility in simulation for a specified application as evidence of completeness and correctness, communicated forthright and understandably, and documented for the public record. The process described and demonstrated in this report directly informs this definition of credibility.

(FDA, 2023) offers an alternative definition of credibility: “the trust, established through the collection of evidence, in the predictive capability of a computational model for a context of

[1] Capability computing is the use of the most powerful supercomputers to solve the largest and most demanding problems. The maximum processing power is characteristically applied to a single job.
[2] Capacity computing is the use of computer resources to run many smaller problems at the same time to support design studies or to perform uncertainty quantification studies.

use.” The focus is solely on the collection of evidence. My preferred definition introduces other important attributes, is stated in a goal-oriented manner, and is actionable.

Regulatory agencies have no say in how the industry internally uses testing and computer simulations if regulatory decisions are not involved. Organizations get to set their standards for testing and the use of computer simulation (simulation governance). Organizational standards are based primarily on perceived benefits, experience, and costs.

Even when certification remains test-based, computational simulation has proved its value to industry. Simulation has effectively optimized system and component design for performance, safety, and manufacturability. Optimized designs often deliver greater performance and safety on the first design iteration, foregoing the need for costly multiple design-build-test cycles, which was the historical norm.

Computational simulation supports testing activities. Historically, environments for certification testing were based on the best judgment of subject matter experts and were often associated with the compounding of rare scenarios and severe assumptions. Simulation increasingly plays a role in exploring system responses to many plausible scenarios and their associated environments to identify challenging but physically realizable conditions for certification testing. Unlike testing, simulation can easily explore margins and where potential “cliffs” might exist. A cliff is generally considered a state where performance or safety suddenly degrades; however, I am aware of a case where simulation led to the *discovery* of a cliff where margins were suddenly improved through a better understanding of system response to realistic environments.

Historically considered conservative, simulation has also been used to derive component environments from external system environments. In many cases, environments on a component internal to a system are less severe than external environments, making design and certification easier, but this is not always the case. I know one case where component environments were amplified relative to external environments. Simulation played a critical role in the *discovery* of this counter-intuitive reality. Subsequent design and certification of components to the proper environment avoided the potential for “surprise,” i.e., component (hence system) failure should the scenario of interest ever occur.

Computational simulation has been used to design on-the-ground test facilities replicating severe complex temporal and spatial loading patterns on systems and components. Testing in these new facilities has replaced expensive flight tests or inferior ground tests where the proper environment could not be fully replicated. Simulation also guides the selection, placement, and interpretation of instrumentation, which helps designers better understand why the system performs the way it does when tested.

Certification by Analysis (CbA) is the holy grail for industry. CbA is a new field in which the industry is eager and regulators are cautious. CbA implies that computer simulation can reduce or even eliminate costly certification testing. Based on a long history of operational experience, industry and regulators are comfortable with test standards.

Standards for simulation-informed regulation are new, in a state of flux, and not widely understood. There are no universally accepted standards for simulation, and more importantly, there is little regulatory experience in their application. How much rigor is required, and when does it become too much? Standards for CbA are sometimes more demanding than standards for CbT. For example, CbA may require quantifying uncertainties and their impact on regulatory decision metrics when similar requirements do not exist for CbT for the same regulatory issue. It’s often sufficient in CbT to manage sources of testing error and uncertainty without a formal quantification of their impact on regulatory quantities of interest (QOI).

The regulatory agency bears the risk of CbA, which requires evidence of completeness and correctness, communicated forthright and understandable, and documented for the public record. This often leads to a disparity in the rigor required for simulation compared to historical testing. For example, regulators might look to simulation for a formal Quantification of Margins and Uncertainties (the topic of this report) when neither is required with test-based regulation.

The transition from CbT to CbA is a learning experience for industry and regulatory agencies. Regulatory agencies should seek corroborating evidence through a balance between testing and simulation. Are simulation results consistent with what is already known through testing?

A potential risk to both industry and regulators is that CbA could call into question regulatory decisions that have already been made (Kelly et al., 2011), especially if an objective quantification of uncertainties is expected for simulation and not for testing. The challenge of a new process is that it must complement the FAA‘s current approach to CbT and CbA.

2. Background and Lessons Learned

2.1 Terminology: variability, uncertainty, unknown/unknowns, errors, and intangibles

Terminology must be addressed before discussing formalism in the “uncertainty” (U) in BEPU. Two types of uncertainty need to be addressed: aleatory and epistemic. The terminology is addressed in (Helton & Sallaberry, 2009).

Aleatory uncertainty is the most intuitive. Aleatory uncertainty refers to randomness (perceived) in the occurrence of future events or variability in an attribute of a population of nominally identical units (e.g., size variations of a part associated with manufacturing tolerances). To avoid confusion and to reduce wordiness, it is common to refer to aleatory uncertainty (randomness or variability) as “variability.” Variability is an attribute of a future state or population with a frequency interpretation. The term “frequency” and the symbol “f” are typically reserved for variabilities in the risk assessment community.

Variabilities should be represented and propagated without controversy using the standard rules of probability theory. Conceptually, it should not matter how variabilities are propagated, provided the rules of probability theory are not violated. Common examples of propagation methodologies are the method of moments, Monte Carlo, and polynomial chaos; however, there may be underlying assumptions or practical limitations (computational costs) associated with any method. A fundamental assumption, however, is that the frequency distribution is accurately known, which is rarely, if ever, the case.

Epistemic uncertainties refer to a lack of knowledge with respect to an appropriate value to use for a quantity that has a fixed value in the context of a specific analysis. Epistemic uncertainties have a confidence or belief interpretation. To reduce wordiness, epistemic uncertainties are typically referred to as “uncertainties,” which is why it is essential to minimize confusion with aleatory uncertainties by referring to the former as “variability.” The term “probability” and the symbol “p” are typically reserved for uncertainty in the risk assessment community.

There is no single universally accepted method for representing and propagating epistemic uncertainties. One well-established method (NaRC, 2009) is to represent and propagate epistemic uncertainties using the same rules of probability theory used to represent and propagate frequencies; however, a separation is maintained between uncertainty (p) and frequency (f) in a framework referred to as probability of frequency or second order probability. Bayesian methods have been used to update representations of uncertainty when new information becomes available (Garrick, 2008). The probability of frequency framework has been used in many high-consequence risk studies (Breeding et al., 1992; Helton, 1994; Helton, 2003; Helton et al., 2011).

Uncertainty distributions (measures of belief) are in the eye of the beholder, not an attribute of a future state or population. Different beholders might have different perceptions of what the uncertainty distribution should be. Methods exist to aggregate alternate plausible uncertainty distributions to represent consensus. One method relies on averaging, i.e., weighing the experts. Another relies on weighing the evidence in the hopes that a pool of experts can form a consensus distribution. In either case, there is no expectation that a correct distribution exists, and any result represented as an uncertainty distribution must be viewed as conditional on the judgment and credentials of those who generated the input distributions.

Aleatory uncertainties are conditional on frequency distributions, which are rarely known with great accuracy. Sampling errors are a common source of epistemic uncertainty even when relevant data exist. Epistemic uncertainties are represented by uncertainty distributions characterizing degrees of belief of the beholder. There is no expectation that there is a correct uncertainty distribution. Methods exist that are less restrictive, such as p-boxes (Ferson et al., 2002), Evidence Theory, and Belief and Plausibility. These methods are intended to represent and propagate aleatory and epistemic uncertainties more consistent with the nature of imprecise information. These tend to be research topics, and it is uncommon to see them applied in regulatory applications. In some cases, math can be daunting, and the expertise to use these methods and correctly interpret results typically resides with individuals in the research community, not industry or the regulatory community. This creates confusion and undermines confidence in the results.

Bayesian methods have also been used when both uncertainty and frequency are relevant. The separate concepts of frequency and uncertainty are blended into a single concept termed “likelihood.” Likelihood limits to frequency or uncertainty when one or the other dominates.

Proponents of every method for representing and propagating epistemic uncertainties can be very assertive, and each method has conceptual and practical advantages and disadvantages. Regulatory agencies are challenged because they oversee many regulatory issues, and there are many submittals for the same regulatory issue over time. If different methodologies are used, regulatory agencies will have difficulty discussing and comparing results internally or representing results to their external sponsor (e.g., congress). As a practical matter, regulatory agencies seeking consistency should provide guidance on how aleatory and epistemic uncertainties should be represented within their agency. In the absence of regulatory guidance, best practice is that members of the assessment community explicitly declare their methodology and be consistent when representing, propagating, and interpreting results. Regulatory agencies should be able to approve the proposed method before significant work is performed (see Section 0).

The distinction between variability (frequency, randomness) and uncertainty (probability, belief) matters! Consider an example where national security hangs in the balance when deploying a new weapon system.

On the one hand, the new weapon system might have a tested failure frequency of 1% (99% success rate), and we know this with confidence approaching 100% if the tested population is huge. The correct interpretation is (on average) that 1 in every 100 weapons will fail. This may be acceptable, but for high-value targets, it may be desirable to deploy two weapons to ensure that at least one performs 99.99% of the time.

On the other hand, the new weapon system might have an assessed failure probability of 1% and has never been tested. The correct interpretation is that there is a 1% belief that *all* the weapons will fail. It is plausible that the entire stockpile of new weapons is worthless, and redundant deployment of the same system has no value! The availability of alternate weapon systems based on fundamentally different design concepts is one way to hedge against this risk.

Unknown/unknowns cannot be quantified, and as such, they cannot be explicit contributors to BEPU. (Taleb, 2007) refers to Black Swans, which are unknown/unknowns with the attributes of surprise and high consequence. Although unknown/unknowns are not quantifiable, they can be proactively managed, thus increasing regulatory confidence that the assessment of uncertainties is more complete. Examples of unknown/unknowns and their consequences can be easily found on the Web by searching for engineering disasters.

It is common to assert “an act of God” when a surprise occurs with catastrophic consequences. Who could have known? In reality, some unknown/unknowns are should-have-been-knowns because similar events or similar precipitating events may have happened previously in the same or other industries. The Challenger accident (Vaughan, 1996) is a disastrous example of should-have-been-knowns. Organizational memory (both documented and oral), lessons learned, and formal peer review by subject matter experts with a spectrum of experiences are effective ways of bringing to light should-have-been-knowns.

The risk of unknown/unknowns can also be managed through better design. One way is to reduce the potential for human errors by reducing or eliminating the need for process or procedural safeguards. For example, don’t allow communities to build in the flood zone below a dam. No need for warning systems and evacuation procedures!

Another way is to reduce sensitivity to common-mode failures by adding redundancy (e.g., independent backup systems or diagnostic voting for critical inputs) or adding diversity (e.g., making the system incompatible with certain initiators). Reducing system complexity through simplifying notions, passive designs, fail-safe designs, or firewalls so failures cannot propagate can also reduce the potential for unknown/unknowns.

Lastly, defense-in-depth strategies can mitigate the consequences of unknown/unknowns. For example, the licensing of nuclear power plants requires that the frequency of initiating events leading to core damage be extremely low *and* that a robust reactor containment building be constructed to prevent the escape of radioactive materials to surrounding populations should core damage occur. It could be that overall safety goals could be achieved through small core damage frequencies alone or by having a very robust containment alone. Yet, both are still required to hedge against surprise in assessing either.

Robustness to unknown/unknowns can be created by adopting conservative acceptance and intermediate requirements. Conservative system requirements subjectively demand more margin than objective BEPU would suggest (e.g., FoS > 3 instead of 1). Setting requirements for intermediate steps or barriers (gates) creates balance by avoiding the situation where all your eggs are in one basket (see the defense-in-depth example above). Conservative and intermediate requirements can be accomplished while maintaining a BEPU philosophy for assessment.

Nature will often reveal its unknown/unknowns if given the opportunity. Full system testing for radically innovative designs is usually prudent, allowing Nature to reveal its unknown/unknowns, regardless of the perceived maturity of computer simulation. A test is one full reality partially revealed[3]. A full system test is often a test of functionality with only a limited understanding of why the system behaved the way it did. Hierarchal testing (components, subsystems, and systems) provides insight into why the system behaves the way it does. Highly accelerated lifetime testing (HALT) can identify nearby failure cliffs for components and subsystems. Full system tests are often over-tested beyond required environments to look for nearby failure cliffs. Acceptance and surveillance testing look for unexpected manufacturing defects or aging effects.

Lastly, the risk of unknown/unknowns can be managed through increased technical understanding using computer simulations. The exploration of many scenarios, made possible by computer simulation, are many partial realities fully revealed[3]. *Fully revealed* because computed information can be examined for every computed time and location. However, computer simulations are only as good as the physics and algorithms in the code, so at best, simulation results can only be an approximate reality. BEPU focuses on predictive models (defensible technical basis) and objective quantification of uncertainties (requiring ensemble computing to explore all the scenarios), which is inherently a hedge against unknown/unknowns. Examples can be cited where full system computer simulations have revealed previously unexpected system interactions with environments, leading to a substantial increase in risk in some cases or a substantial decrease in risk in others.

Testing and computer simulation complement each other's potential shortcomings. Consequently, integrating testing and computer simulation is often the best strategy to manage the risk of unknown/unknowns, especially for radically innovative designs.

Errors are deviations from the truth. Errors are not random (i.e., aleatory), but their quantification can be uncertain (i.e., epistemic). Known sources of error should be acknowledged and preferably corrected. Still, there might be pragmatic constraints (e.g., time, money, or computer resources) that limit the ability to correct errors in a timely manner. Known errors can be ignored

[3] I owe these insights to the late Sheldon Tiezen, a colleague at Sandia National Laboratories.

if they have an insignificant impact on decision metrics or if the error is negligible relative to other contributors to variability or uncertainty. Otherwise, assessment results should be bias-corrected for known sources of error; however, it can be difficult to propagate bias-corrected results up an assessment hierarchy.

A code bug is an example of an error that should be corrected based on priority and the availability of resources. Numerical errors resulting from the solution of the physics equations using finite discretization are another example of an error. Discretization errors can be ignored if they are insignificant, minimized with additional refinement, or the results should be bias-corrected.

Sometimes, there is an acknowledged source of error or uncertainty for which there is no objective or subjective evidence to support its quantification or for which there is a conscious choice *not* to quantify for whatever reason. I refer to these as intangibles. Industry should document intangibles as part of their submittal for a regulatory decision. Regulatory agencies can respond by:

1. Requiring more information before making their regulatory decision, or
2. Explicitly accepting the risk without the need for additional information or quantification, or
3. Increasing the acceptance threshold for regulatory decision metrics notionally requiring additional margin on the margin.

2.2 Lessons learned from blind predictions

Blind predictions are the best test of predictive capability. Figure 2.1 summarizes the results of nearly 100 round-robin studies[4] where blind predictions are made a priori and subsequently compared to a referent representing truth. Although results are available for many engineering disciplines, I filtered the studies for solid mechanics, the class of engineering codes used for aircraft seat certification.

Each study (round-robin entry) has vertical dots representing the participants' blind predictions. The narrow horizontal green bands represent the FAA's +/-10% validation acceptance limits. Most participant submittals are outside this band; however, given enough participants, someone will get it about right in most studies. There is no way of knowing if this “wisdom of the masses” is a repeatable skill of a few talented analysts or just luck.

The broader black bands are 95% confidence bounds, i.e., 95% of the dots are contained within these bounds. Predictions can be as much as a factor of three, too high or too low. I refer to this as analyst-to-analyst (A2A) uncertainty. The predictive capability of computational simulation is not mature enough to accept blind predictions for most regulatory applications unless the margin is huge.

[4] Currently unpublished data collected by M. Pilch in collaboration with W. L. Oberkampf (wloconsulting) and J. Wood (jwanalysis).

Data in Figure 2.1 span about three decades (left to right), during which period computing capabilities (Figure 1.2) increased by about eight orders of magnitude. Increasing computing capabilities alone does not ensure better predictions.
A2A uncertainty has the potential to dominate all other sources of uncertainty in any specific application unless it is managed. Three broad lessons learned result from a root cause analysis to understand the source of this A2A uncertainty, which lead to the following recommendations for the management of A2A:

1. Buy down the uncertainty for specific applications by applying a common modeling approach that specifies the whats and not necessarily the hows (no evidence that round-robin participants followed a formal modeling approach beyond analyst judgment) that
2. Reduces and manages numerical errors (numerical errors are rarely quantified in round-robin studies) and that
3. Anchors predictions to a relevant validation hierarchy (alternate plausible models are common in round-robin studies).

The process proposed in Section 0 and the key concepts in Section 0 address these recommendations. The proposed process is focused on the “whats”, not the “hows”, with a goal of getting (about) the right answer for the right reason; consequently, the process is expected to reduce, but not eliminate, A2A uncertainty.

Of course, an organization’s simulation governance can also significantly reduce A2A uncertainty by prescribing *how* a class of analyses will be performed. Simulation governance will have its greatest value when it reflects the key concepts discussed in Section 0; otherwise, the simulation governance could steer *all* analyses to a potentially wrong answer.

I observed a presentation where three new analysts were given the organization’s simulation governance for a class of analyses. The three analysts still produced different results for a representative problem. This is a reminder that an organization’s simulation governance, or the process and key concepts proposed here, can never drive A2A uncertainty to zero. In the regulatory arena, A2A uncertainty will always remain unassessed, and its potential impact on decisions will likely be overlooked. The potential for A2A uncertainty is intangible and requires that regulatory agencies consider it.

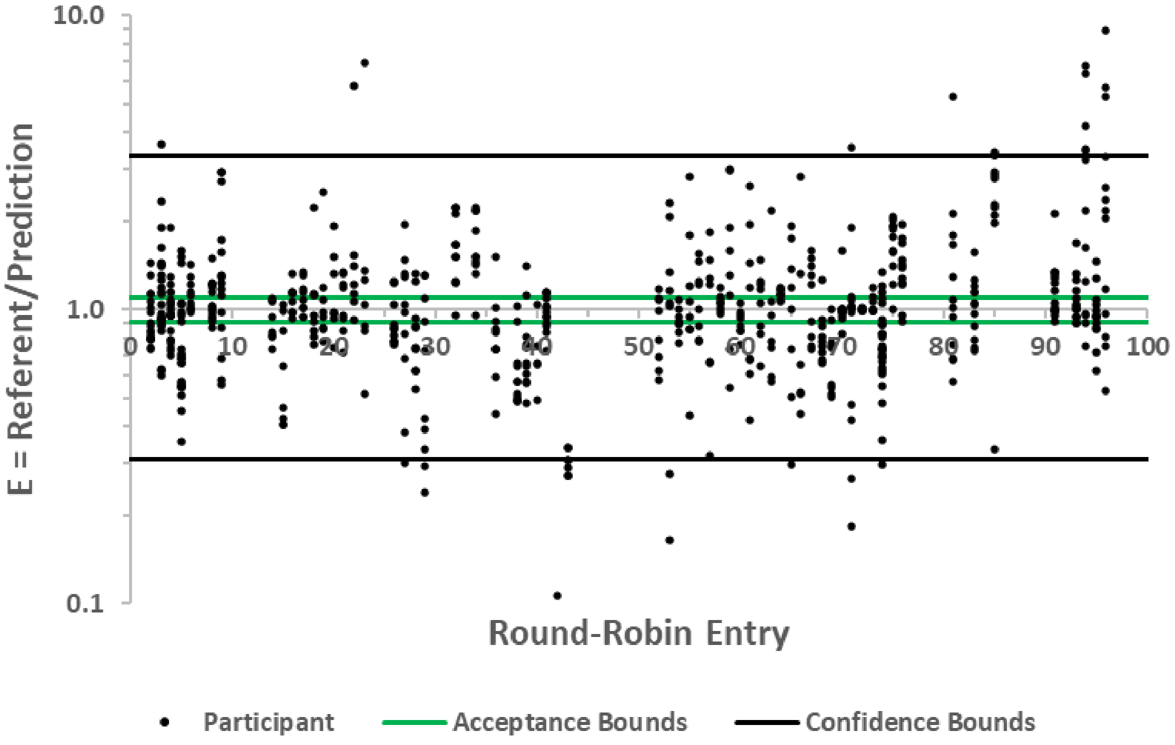


Figure 2.1: Blind predictions are the best test of predictive capability

2.3 Lessons learned from Best Estimate Plus Uncertainty (BEPU)

The United States (US) Nuclear Regulatory Commission (USNRC) championed the concept of Best Estimate Plus Uncertainty (BEPU) as an alternative to conservative regulatory requirements, which created operational and economic penalties for the nuclear power industry. In 1989, the USNRC modified rules and allowed, as an alternative, the best estimate methods if they are coupled with defensible measures of uncertainty in predicting safety parameters.

Assessments must be unbiased, with an objective accounting of all dominant contributors to uncertainty. This allows for discovery because all interactions are represented unbiasedly, and the results of their interactions are not always intuitive[5]. Sensitivity analysis (SA) becomes a powerful tool for efficiently prioritizing future investments that could improve assessment results. These two principles, unbiased assessments with objective quantification of uncertainties, define BEPU.

The benefits of BEPU are undermined when “conservatism” is sprinkled around various elements in an assessment. Introducing these conservatisms is often not transparent to regulators, with the potential for misunderstanding or misrepresenting results. Ad hoc conservatism often leads to the compounding of conservative or unrealistic assumptions, which leads to scenarios that may never occur or hyper-conservative results that negatively

[5] I was exposed to three issues during my career where simulation led to discovery and better understanding of system response. In two cases, risk was significantly greater than previously believed; in the third case, risk was significantly less than previously believed.

impact the economics of the system. The introduction of ad hoc conservativism throughout assessments skews sensitivity analyses. This could lead to ineffective or excessive resource allocation by steering resources to extremely unlikely or physically impossible scenarios and processes.

Ad hoc conservatism undermines the potential for discovery because what leads to conservative results is not always intuitive, especially when physics is competing or nonlinear, when threshold phenomena are present, or when physics exhibits resonance behavior. For example, containment buildings for nuclear power plants were designed to accommodate steam blowdown following a large break loss of coolant accident. This was judged "conservative" for containment pressurization. The accident at Three Mile Island (TMI) and subsequent analyses showed that small break loss of coolant accidents could lead to core melt, failure of the reactor coolant system while still at pressure, and dispersal of core materials into the containment, leading to a phenomenon known as Direct Containment Heating (DCH). DCH could produce containment pressures higher than previously asserted as "conservative[6]."

System requirements are often in conflict. For instance, performance requirements (small and light design) frequently conflict with safety requirements (big and robust design). What is conservative for one requirement is often not conservative for another, and design is challenged with optimizing between the two, which is impossible if ad hoc conservatisms are sprinkled throughout assessments.

Decision criteria are the *only* place where conservatism is encouraged, e.g., requiring additional margin. When confined to decision criteria, conservatism is explicit and transparent, does not distort assessments or sensitivity analyses, and adds robustness to decisions already made should added information become known.

Quantifying Reactor Safety Margins (NRC, 1989) was the USNRC's first demonstration of BEPU. A process called Code Scaling and Uncertainty (CSAU) was explicitly defined for this application and the available codes at the time. Key concepts and many process elements can be extended to other applications. There are three main elements and fourteen process steps associated with CSAU (see Figure 2.2). The three main elements are:

1. Element 1. Requirements and code capabilities: This element addresses the applicability of a computer code for a target application by specifying a scenario, the geometry, and associated initial conditions. Physics requirements are identified and ranked. The product is called a PIRT (Phenomena Identification and Ranking Table). Code readiness is assessed through available documentation and the capability of the code to model the required physics.
2. Element 2. Assessment and ranging of parameters: CSAU aggregates uncertainties from design, operating conditions, and separate effects of physics[7]. This element addresses their quantification with particular emphasis placed on understanding and bias-correcting scale distortions if they exist, i.e., how separate effect models behave at full physical scale when characterized and validated at smaller physical scale. This element also addresses the validation of the model against integral effects data.

[6] M. Pilch played a central role in identifying DCH as a new risk for nuclear power plants. He also played a central role in resolving the issue for the USNRC through an integration of testing and modeling in a probabilistic framework. M. Pilch has also observed other examples of where the principles of BEPU have revealed either greater risk or lower risk in other application areas during his career at Sandia National Laboratories.

[7] These are the three key elements of what we now call the simulation conceptual model (see Section 0).

3. Element 3. Sensitivity and uncertainty analysis: This element computes the total uncertainty for the QOI for the target application by propagating uncertainties from design and operating conditions through the hierarchy of separate effects models with associate uncertainties. Sensitivity analysis determines which input uncertainties are dominant contributors to total uncertainty.

The CSAU process has since been applied many times by the nuclear power industry worldwide with modifications to fit each circumstance. PIRT and emphasis on scaling are unique contributions of CSAU.

The PIRT process ensures sufficiency, efficiency, and transparency to physics requirements and capabilities and is an effective tool to drive research needs and communicate "coverage." PIRT development is now a cornerstone activity in many nuclear power applications. The V&V Program at Sandia National Laboratories has also effectively adopted PIRT for its nuclear weapons applications. The NRC recognized that PIRT is usually a subjective process involving subject matter experts and stakeholders and felt that formal "scaling analysis" could bring more rigor to PIRT development.

Explicit emphasis on scaling is another important contribution of CSAU. In CSAU, scaling does not just refer to physical scale (size) but more broadly to physics scaling, i.e., the combination of physical scale, initial and boundary conditions, system state, and physics into non-dimensional terms representing the relative importance of various processes. For example, Reynolds number is a dimensionless term involving dynamic pressure to shear stress ratio for flow over an aircraft wing. Wing performance at full physical scale can be inferred from observations in wind tunnel tests (small physical scale) if the Reynolds number is matched between the two. The NRC (Zuber et al., 1998) formalized the process for physics scaling for NRC applications by taking a top-down and bottom-up perspective for complex systems. Physics scaling supports the design of new test facilities, identifies potential scale distortions in existing facilities, and provides a quantitative means for assessing phenomena in the PIRT.

Key lessons learned from a review of BEPU:

1. The essential characteristics of BEPU are that assessments are unbiased with an objective accounting of all dominant contributors to uncertainty. Decision criteria are the only place where conservatism is encouraged.
2. Physics scaling is essential in establishing the relevance of tests and models for target applications.

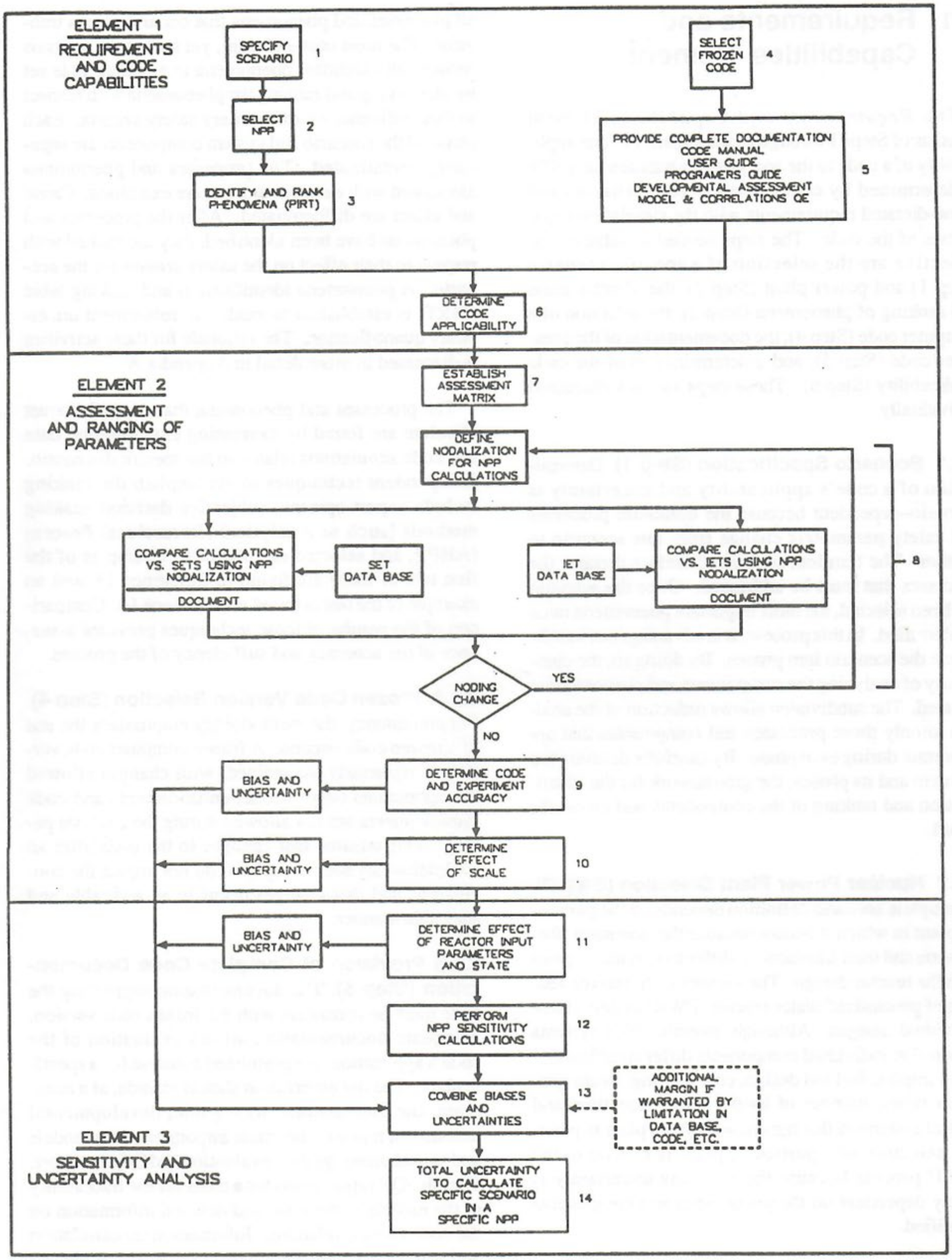

ELEMENT 1
REQUIREMENTS AND CODE CAPABILITIES
SPECIFY SCENARIO 1
SELECT NPP 2
IDENTIFY AND RANK PHENOMENA (PIRT) 3
SELECT FROZEN CODE 4
PROVIDE COMPLETE DOCUMENTATION CODE MANUAL USER GUIDE PROGRAMERS GUIDE DEVELOPMENTAL ASSESSMENT MODEL & CORRELATIONS QE 5
DETERMINE CODE APPLICABILITY 6
ELEMENT 2
ASSESSMENT AND RANGING OF PARAMETERS
ESTABLISH ASSESSMENT MATRIX 7
DEFINE NODALIZATION FOR NPP CALCULATIONS
COMPARE CALCULATIONS VS. SETS USING NPP NODALIZATION
DOCUMENT
SET DATA BASE
IET DATA BASE
COMPARE CALCULATIONS VS. IETS USING NPP NODALIZATION
DOCUMENT
8
NODING CHANGE
YES
NO
BIAS AND UNCERTAINTY
DETERMINE CODE AND EXPERIMENT ACCURACY 9
BIAS AND UNCERTAINTY
DETERMINE EFFECT OF SCALE 10
BIAS AND UNCERTAINTY
DETERMINE EFFECT OF REACTOR INPUT PARAMETERS AND STATE 11
PERFORM NPP SENSITIVITY CALCULATIONS 12
COMBINE BIASES AND UNCERTAINTIES 13
ADDITIONAL MARGIN IF WARRANTED BY LIMITATION IN DATA BASE, CODE, ETC.
ELEMENT 3
SENSITIVITY AND UNCERTAINTY ANALYSIS
TOTAL UNCERTAINTY TO CALCULATE SPECIFIC SCENARIO IN A SPECIFIC NPP 14

Figure 2.2: Code scaling, applicability, and uncertainty (CSAU) evaluation methodology

2.4 Lessons learned from Quantification of Margins and Uncertainties (QMU)

The fundamental concepts of Quantification of Margins and Uncertainties (QMU) were initially developed concurrently at several national laboratories supporting nuclear weapons programs in the late 1990s, including Los Alamos National Laboratory, Lawrence Livermore National Laboratory, and Sandia National Laboratories. QMU focuses on identifying, characterizing, and analyzing performance thresholds and their associated margins for engineering systems that are evaluated under conditions of uncertainty, mainly when portions of those results are generated using computational modeling and simulation (M&S). Examples of QMU outside the nuclear weapons community include NASA interplanetary spacecraft and rover development, missile six-degree-of-freedom simulation results, and characterization of material properties in terminal ballistic encounters.

QMU is typically associated with a decision metric called the confidence factor (CF = Margin/Uncertainty). CF>1.0 implies sufficient margin to overcome an objective assessment of uncertainties. The confidence factor is intuitive and easy for decision-makers to understand. It is also dimensionless, enabling comparing and prioritizing many regulatory decisions using a standard decision metric.

There is no accepted process for QMU, and early reviews noted a lack of formalism in assessing margins and uncertainties. The National Research Council of the National Academy of Sciences (NaRC, 2009) reviewed the implementation of QMU by the three national laboratories for nuclear weapon applications. The NaRC made specific recommendations that add the necessary rigor to QMU.

The NaRC stated:

- QMU provides input for *a risk-informed* decision-making process.
- QMU can benefit from the structured methodology and discipline of quantitative and Probabilistic Risk Assessment.
- In characterizing uncertainties, it is important to pay attention to the distinction between those arising from incomplete knowledge (“epistemic,” or systematic) and those arising from device-to-device variation (“aleatory,” or random).
- The probability-of-frequency approach is the best format for representing aleatory and epistemic uncertainties in Probabilistic Risk Assessments.

The conceptual and computational basis for QMU is described (Helton, 2011) for the general case where both aleatory and epistemic uncertainties are present. The mathematical framework presented by Helton is the probability of frequency approach recommended by (NaRC, 2009). Examples of QMU with analyses from reactor safety and radioactive waste disposal involving the probability of frequency approach are presented in (Helton et al., 2011).

The certification of nuclear weapons has always been risk-informed based on the best available information from testing and computational simulation. Integrating QMU into nuclear weapon assessments provides a quantitative metric, CF=Margin/Uncertainty, to inform decisions. Even with quantitative metrics, (Pilch et al., 2011) reemphasize that nuclear weapon decisions should remain risk-informed and that decision-makers should not abdicate their responsibility to a computed value of the CF. Subjective factors such as the judgment of SMEs, risk tolerance, political environment, the credibility of assessments, and the specter of unknown/unknowns will continue to inform nuclear weapon decisions. Acceptable values of CF range from 2 to 10 for nuclear weapon decisions as a subjective hedge against these factors.

2.5 Lessons learned from Probabilistic Risk Assessments (PRAs)

The USNRC has increased the use of Probabilistic Risk Assessment (PRA) in regulatory matters to the extent supported by state-of-the-art PRA methods and data and in a manner that compliments the USNRC's deterministic approach and supports the USNRC's traditional defense-in-depth philosophy. This has been in response to the Government Performance and Results Act (GPRA) law enacted by Congress in 1993. PRA evaluations in support of regulatory decisions should be as realistic as practical and should be used with appropriate consideration of uncertainties. PRA is yet another implementation of the BEPU philosophy. NUREG-1150 (Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants, 1990) (NRC, 1990) was an early and significant implementation of the PRA methodology with quantitative consideration of uncertainties. Best estimate computer codes, judgment, and uncertainties were used to make key assessments where appropriate.

Congressional motivation and the USNRC's evolving experience with PRA have led to a plan for risk-informed and performance-based regulation (SECY-07-0191). Full-scope PRA is outside the scope of FAA's current regulatory approach and this proposal, but some framework considerations have value in informing the new process.

PRA is a framework that expands the concepts of BEPU and QMU by providing a more holistic view of the system. This leads to a balanced understanding of risk and where regulation needs to be improved or relaxed.

Although the scope, depth, and application of PRAs vary widely, they all follow the same basic steps (NaRC, 2009) and (Garrick, 2008):

1. Define the system being analyzed and what constitutes normal operation to serve as the baseline reference point.
2. Identify and characterize the sources of danger and the hazards (e.g., stored energy, toxic substances, hazardous materials, acts of nature, sabotage, terrorism, equipment failure, and combinations of each, etc.)
3. Develop "what can go wrong" scenarios to establish levels of damage and consequences while identifying points of vulnerability.
4. Based on the relevant evidence, quantify the likelihood of the different scenarios and their attendant levels of damage.
5. Assemble the scenarios according to damage levels and cast the results into the appropriate risk curves and priorities.
6. Interpret the results to guide the risk management process.

PRA is almost always an exercise of predictive uncertainty focused on alternate plausible models, as defined in Appendix A, because there are no system-level data to validate predictions, although data exist to quantify uncertainties for elements of the conceptual model. (Helton et al., 2011) present an example taken from radioactive waste disposal where both aleatory and epistemic uncertainties are present (see Figure 2.3). Each strand represents aleatory uncertainty for one realization of epistemic uncertainty. The strands of the horsetail represent many realizations of epistemic uncertainty. The latter can be interpreted as alternate plausible models.

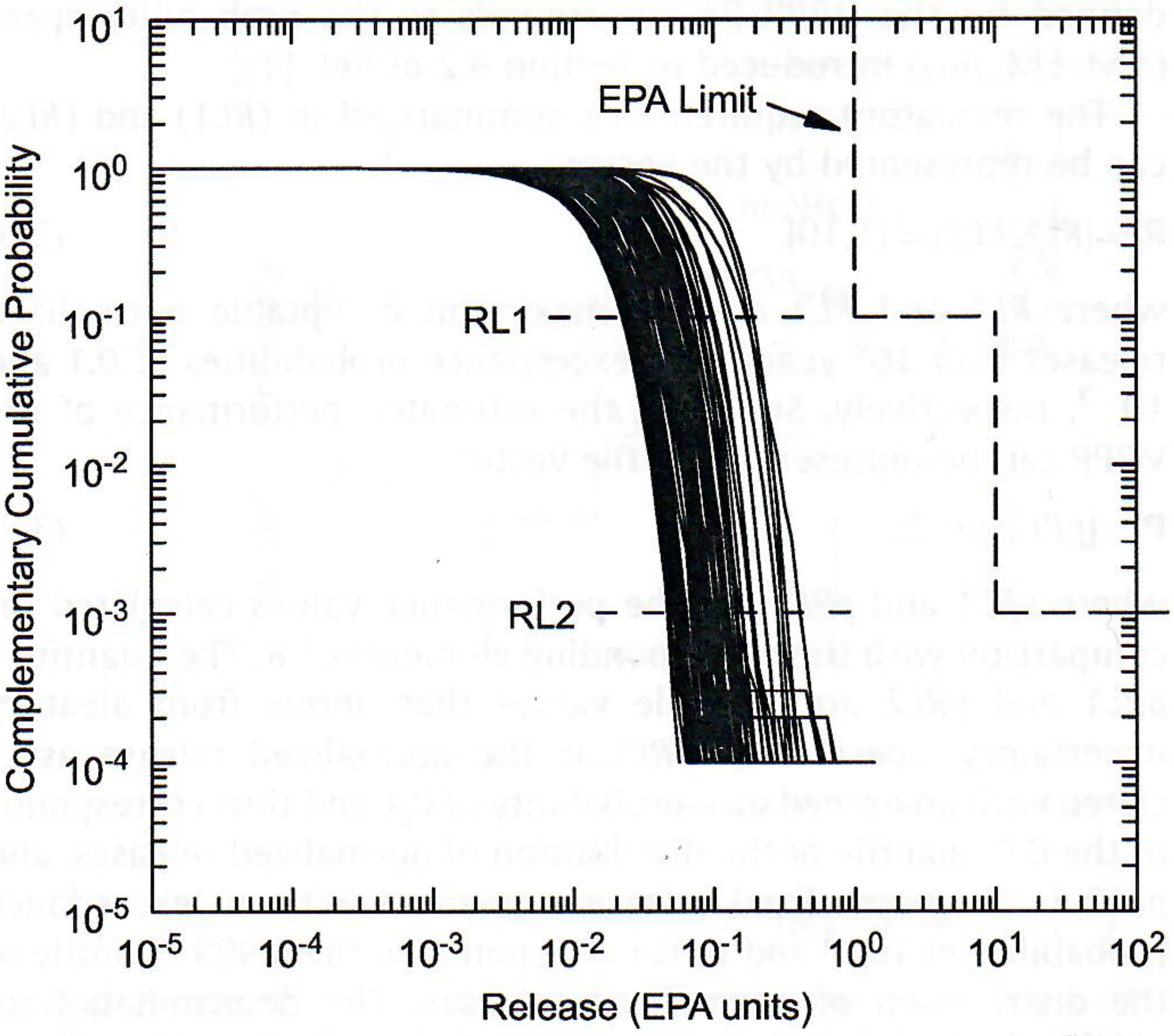


Figure 2.3: Prediction uncertainty for radioactive waste disposal

2.6 Lessons learned from verification and validation

One strength of BEPU, QMU, and PRA is that a common framework can integrate testing and computational simulation to the extent that either is available. The basic concepts apply equally to data-rich issues where computational simulation is not required. However, a common feature is increased reliance on computational modeling and simulation, which introduces new sources of uncertainty not seen in test-based certification.

The formalism of verification and validation (V&V) must complement the formalism of uncertainty quantification (UQ) to identify, quantify, and manage errors and uncertainties in the results of computer simulations. The importance of V&V in the nuclear weapons program is recognized with an explicit program element as part of the Advanced Simulation and Computing (ASC) program at the three national laboratories. The National Nuclear Security

Agency (NNSA, and sponsor of the ASC program) funded a study by the National Research Council of the National Academy of Sciences aimed at assessing the reliability of complex (computer) models (NaRC, 2012), which was a follow-on to the review of QMU (NaRC, 2009).

(NaRC, 2012) stated that the appropriate level of confidence in the results from computer simulations must stem from an understanding of a model's limitations and the uncertainties inherent in its predictions. Ideally, this understanding can be obtained from three interrelated processes that answer key questions:

1. *Verification.* How accurately does the computation solve the underlying equations of the model for the quantities of interest?
2. *Validation.* How accurately does the model represent reality for the quantities of interest?
3. *Uncertainty Quantification* (UQ). How do the various sources of error and uncertainty feed into uncertainty in the model-based prediction of the quantities of interest?

(NaRC, 2012) summarized three fundamental V&V principles as a first step toward identifying best practices.

1. VVUQ tasks are interrelated. A solution verification study may incorrectly characterize the accuracy of a code's solution if code verification is inadequate. A validation assessment depends on assessing numerical error produced by solution verification and on the propagation of model-input uncertainties to computed QOIs.
2. The processes of VVUQ should be applied in the context of an identified set of QOIs. A model may provide an excellent approximation to one QOI in each problem while providing poor approximations to other QOIs. Thus, the questions VVUQ must address are not well posed unless the QOIs have been defined.
3. Verification and validation are not yes/no questions with yes/no answers but rather are quantitative assessments of differences. Solution verification characterizes the difference between a computational model's solution and that of the underlying mathematical model. Validation involves quantitative characterization of the difference between computed QOIs and true physical QOIs.

Specific to verification, the (NaRC, 2012) committee identified several guiding principles and associated best practices. Some of the more important principles and practices are summarized here:

1. Principle: Solution verification is well defined only in terms of specified quantities of interest, which are usually functionals of the total computed solution.
    a. Best practice: Clearly define the QOIs for a given VVUQ analysis, including the solution-verification task. Different QOIs will be affected differently by numerical errors.
    b. Best practice: Ensure solution verification encompasses the full range of inputs employed during UQ assessments.
2. Principle: The efficiency and effectiveness of code and solution verification can often be enhanced by exploiting the hierarchical composition of codes and mathematical models, with verification performed first on the lowest-level building blocks and then on successively more complex levels.
    a. Best practice: Identify hierarchies in computational and mathematical models and exploit them for code and solution verification. It is often worthwhile to design the code with this approach in mind.

    b. Best practice: Include in the test suite problems that test all levels in the hierarchy.
3. Principle: The goal of solution verification is to estimate and, if possible, control the error in each QOI *for the problem at hand*.
    a. Best practice: When possible, in solution verification, use goal-oriented a posteriori error estimates, which give numerical error estimates for specified QOIs. In the ideal case, the fidelity of the simulation is chosen so that the estimated errors are insignificant compared to the uncertainties arising from other sources.
    b. Best practice: If goal-oriented a posteriori error estimates are not available, try to perform self-convergence studies (in which QOIs are computed at different levels of refinement) on the problem at hand. These studies can provide helpful estimates of numerical error.

Many VVUQ tasks introduce questions that can be posed and, in principle, answered within the realm of mathematics. Validation and prediction introduce additional questions whose answers require judgments from the realm of subject-matter expertise. For validation and prediction, the (NaRC, 2012) committee identified several principles and associated best practices. Some of the more important of these are summarized here:

1. Principle: A validation assessment is well defined only in terms of specified quantities of interest (QOIs) and the accuracy needed for the model's intended use.
    a. Best practice: Specify the required accuracy and the QOIs that will be addressed before the validation process.
    b. Best practice: Tailor the effort required to assess and estimate prediction uncertainties to the application's needs.
2. Principle: A validation assessment provides direct information about model accuracy only in the domain of applicability that is “covered” by the physical observations employed in the assessment.
    a. Best practice: When quantifying or bounding model error for a QOI in the problem at hand, systematically assess the relevance of supporting data and validation assessments (which were based on data from different problems, often with different QOIs). Subject-matter expertise should inform this assessment of relevance.
    b. Best practice: If possible, use a broad range of sources of physical observations so that a model's accuracy can be checked under different conditions and at multiple levels of integration.
    c. Best practice: Use “holdout tests” to test validation and prediction methodologies. In such a test, some validation data are withheld from the validation process, the prediction machinery is employed to “predict” the withheld QOIs with quantified uncertainties, and finally, the predictions are compared to the withheld data.
    d. Best practice: If the desired QOI was not observed for the physical systems used in the validation process, compare the sensitivities of the available physical observations with those of the QOI.
    e. Best practice: Consider multiple metrics for comparing model outputs against physical observations.
3. Principle: The efficiency and effectiveness of validation and prediction assessments are often improved by exploiting the hierarchical composition of computational and mathematical models, with assessments beginning on the lowest-level building blocks and proceeding to successively more complex levels.

    a. Best practice: Identify hierarchies in computational and mathematical models, seek measured data that facilitate hierarchical validation assessments, and exploit the hierarchical composition to the extent possible.

(NaRC, 2012) also made some important observations concerning prediction.

1. Both extrapolative and interpolative predictions are risky unless the QOI is a smooth function over the domain. Quantifying uncertainties and assessing their reliability for a prediction require statistical and subject-matter reasoning.
    a. MPilchConsulting (MPC) elaboration: The risk is that the nature of the physics can change in a way not represented in the conceptual model, e.g., flows can transition from laminar to turbulent, threshold phenomena may occur (failure, phase change, runaway chemical reactions, etc.), or resonance phenomena may occur. That latter is a key risk of interpolation. Even when uncertainties are small where data exist, they can leverage into large errors in the application domain if extrapolation is excessive.
2. Prediction uncertainty is a vibrant research topic whose methods vary depending on the features of the problem at hand.
    a. MPC elaboration: The discussion of uncertainties must be framed in the context of a proposed methodology for prediction and prediction uncertainty.
3. Methods for expressing model form error and assessing its impact on prediction uncertainty are in their infancy compared to methods for addressing parametric uncertainty.

Regulatory agencies often look to professional societies for guidance when formulating regulatory standards for CbA. (AIAA, 1998) was an early V&V standard. ASME now has nine separate V&V committees organized around engineering disciplines and product areas. VVUQ10 is the ASME committee that wrote the standard for Verification and Validation in Computational Solid Mechanics (ASME, 2012). (SAE, 2021) tailored the VVUQ10 standard specifically for application to aircraft seat design and evaluation.

The VVUQ10 standard provides some formal definitions:

1. Verification is the process of determining that a computational model accurately represents the underlying mathematical model and its solution.
2. Validation determines the degree to which the model accurately represents the corresponding physical experiment from the perspective of its intended use.
3. Prediction uses a model to calculate a response where the modeler does not know the experimental outputs.
4. Uncertainty quantification characterizes all uncertainties in the model or experiment and quantifies their effect on the simulation or experiment outputs.
5. Model form uncertainty is associated with modeling assumptions and approximations.
    a. MPC elaboration: The VVUQ10 standard is referring to all elements of the simulation conceptual model as developed in Section 0. e.g., environments and initial conditions, system state, and physics.
    b. VVUQ10: Model form uncertainty is extremely difficult to quantify.

It is worth noting that other definitions appear in the V&V community, even within the family of ASME V&V standards. Applying V&V principles in applications cannot wait for community consensus. These definitions are sufficient for this report.

As depicted in Figure 2.4, validation is the assessment of model accuracy by comparing model predictions with experiment results. The process emphasizes the need for strong integration between simulation and testing, as well as the need for simulation quality (code and calculation verification) and testing quality (instrument and data assurance).

Accuracy requirements need to be established for each element in the validation hierarchy, and the decision point “Requirements Satisfied” provides an objective decision point for initiating improvement in the conceptual, mathematical, and computational models and in the experimental designs. The VVUQ10 standard states that system-level accuracy requirements should flow down to establish accuracy requirements for each element in the validation hierarchy. My experience is that this is difficult in most applications. More commonly, elements of the validation hierarchy are treated as pass/fail gates or as an opportunity to calibrate the model (see Section 4.4.4 Calibrate the model0 for cautionary notes on calibration). In the absence of system-level data, the validation hierarchy is the only evidence that system-level predictions are not seriously biased, but it falls short of quantifying model form error (bias), which (NaRC, 2012) and (ASME, 2006) agree is very difficult.

Appendix A better frames a discussion of prediction uncertainty. Two approaches to prediction uncertainty are discussed. The first explores alternate plausible models when no system data exist. This involves quantifying uncertainties in all elements of the simulation conceptual model and their computationally expensive propagation through the computational model.
In the second approach, the simulation conceptual model is frozen, and model form errors and uncertainties are quantified by comparing it with relevant system-level data. This approach is computationally practical.

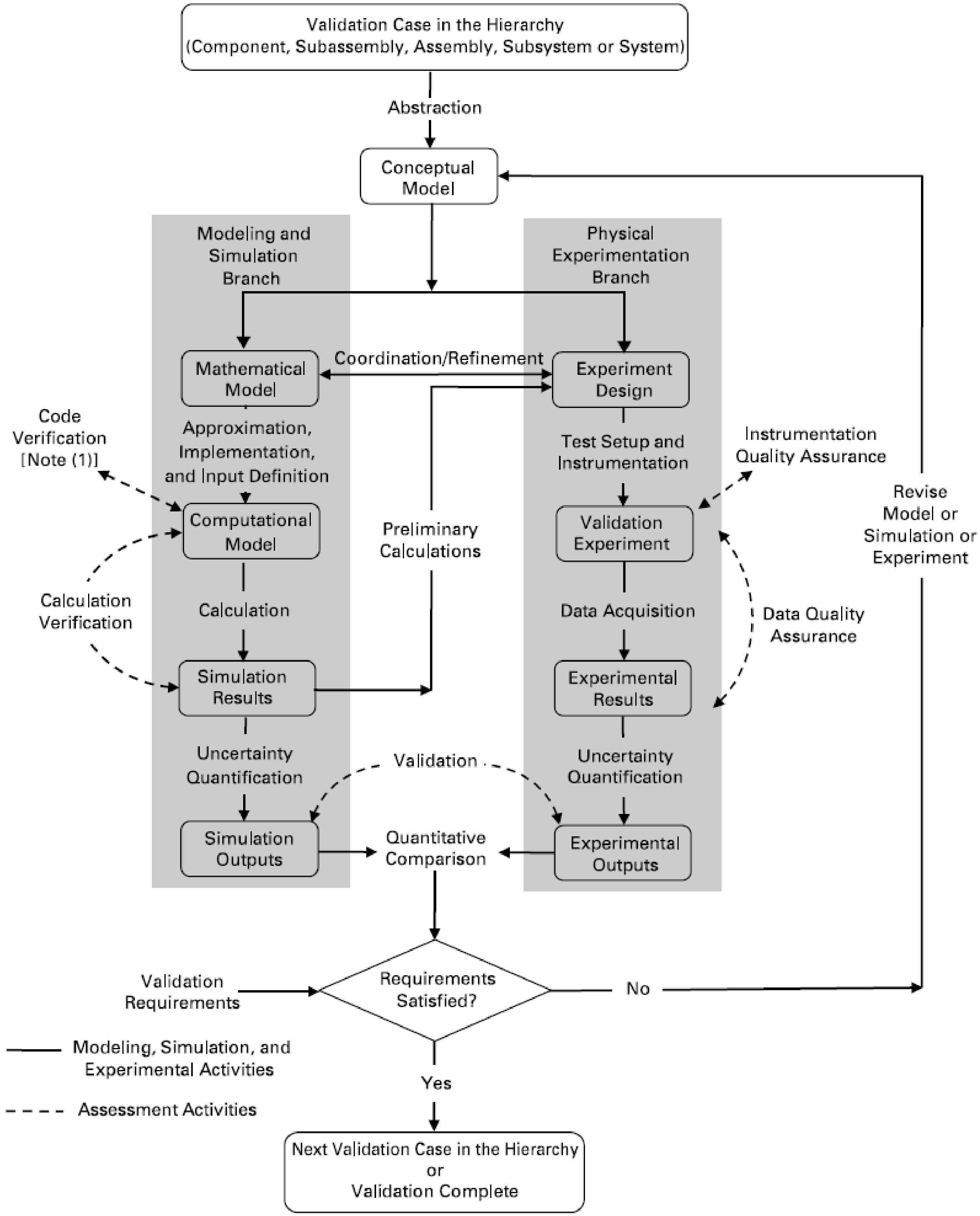

Validation Case in the Hierarchy
(Component, Subassembly, Assembly, Subsystem or System)
Abstraction
Conceptual Model
Modeling and Simulation Branch
Physical Experimentation Branch
Mathematical Model
Coordination/Refinement
Experiment Design
Code Verification [Note (1)]
Approximation, Implementation, and Input Definition
Test Setup and Instrumentation
Instrumentation Quality Assurance
Revise Model or Simulation or Experiment
Computational Model
Preliminary Calculations
Validation Experiment
Calculation Verification
Calculation
Data Acquisition
Data Quality Assurance
Simulation Results
Experimental Results
Uncertainty Quantification
Validation
Uncertainty Quantification
Simulation Outputs
Quantitative Comparison
Experimental Outputs
Validation Requirements
Requirements Satisfied?
No
Modeling, Simulation, and Experimental Activities
Yes
Assessment Activities
Next Validation Case in the Hierarchy
or
Validation Complete

Figure 2.4: Validation process from (ASME, 2019)

2.7 Lessons learned from maturity assessments of computational simulation

The Predictive Capability Maturity Model (PCMM) (Oberkampf et al., 2007) is often cited as providing quality attributes for computational simulation. PCMM was developed to address a challenge by the National Nuclear Security Agency[8] (NNSA), which funded the Advanced Simulation and Computing Program at the three national laboratories. The challenge was *how to measure and communicate progress in predictive capability,* which subsequently was the subject of JASON[9] review. PCMM has repeatedly proven to be a useful framework for discussing simulation quality. The nuclear weapons community at Sandia National Laboratories commonly uses it as a communication framework, but it also has significant limitations.
PCMM has six elements with increasing levels of maturity (see Table 2.1) for each. Each of the six elements of PCMM was intended to answer an essential question about the simulation model. The maturity levels were bounded on the low end by analyst judgment and on the high end with the highest level of rigor imaginable and intermediate grades. The maturity levels were also aligned with the consequence and intended use of simulation results. It was hoped that PCMM would drive analysts to "balance" their efforts across PCMM elements. For instance, geometric fidelity included in the model should not be so detailed that model size renders numerical estimation or uncertainty quantification impossible on the available platforms.

PCMM was later refined by adding much more detail (Hills et al., 2013) in an attempt to remove ambiguity in its application. Elements were added and decomposed into sub-elements, allowing more clarifying language to be developed for each maturity level. The PCMM has spawned other similar efforts by NASA (Blattnig et al., 2013), the DoD (Harmon & Youngblood, 2005), and medical devices (ASME, 2018). There have also been attempts to adopt Technology Readiness Levels (TRLs) to simulation maturity, but (Clay et al., 2007) concluded that TRLs were not well suited for simulation-based activities. Consistent with the original NNSA challenge to "measure," PCMM and most maturity assessment schemes that followed include a scoring scale and an algorithm to aggregate the score across elements.

PCMM was always intended to be evidence-based. Scoring has always proved problematic because it is subjective and often displaced the intended focus on evidence. Creating unique, informative, orthogonal, and, most importantly, unambiguous descriptors was impossible. If such descriptors could be crafted, scoring is just an index scheme for functional descriptive language. Why not just report the functional descriptive language? There is no way to ensure consistent scoring within or across different projects, and "grade inflation" is common. More importantly, there is no way to quantify risk reduction going from one score to the next higher, and the risk reduction, if it could be quantified, is not the same for any two PCMM elements.

[8] The National Nuclear Security Administration (NNSA) is a United States federal agency responsible for safeguarding national security through the military application of nuclear science. NNSA maintains and enhances the safety, security, and effectiveness of the US nuclear weapons stockpile, and has funded the Advanced Simulation and Computing (ASC) Program at the three national laboratories. The ASC Program leverages massively parallel computing to offset the cessation of underground nuclear testing.

[9] JASON is an independent group of elite scientists that advises the United States government on matters of science and technology, mostly of a sensitive nature.

If scoring is problematic, then aggregation of scores within sub-elements and across main elements is even more problematic. (Oberkampf et al., 2007) cautioned against aggregation but recognized that practitioners would not resist the urge. Oberkampf recommended that any aggregation be presented as minimum, average, and maximum scores, but this recommendation was never implemented in applications. It was common that a single summary number (e.g., PCMM = 1.75) would be presented to customers of simulation. Summary scores completely shield decision-makers from the evidence base, contrary to the intent of PCMM.

The community recognized that a graded approach to simulation was required based on the consequence and intended use of simulation results. PCMM tried to reflect this through the maturity levels. However, scoring became counterproductive when scoring targets were associated with usage, e.g., we might have stated that simulation in support of qualification should be performed at Level 2 for each PCMM attribute. The target score, in conjunction with the descriptive language of what was intended by Level 2, effectively told analysts how to do their job; consequently, there was pushback by the analyst community. In addition, aspiring to higher scores also came with cost and schedule implications. In some PCMM elements, state-of-the-art capabilities might not exist to execute at some higher scoring levels. Regardless, customers of simulation often expected Level 3 in simulation results. Are we not the best and brightest? This raises an important question: can simulation results be used even if target scores cannot be achieved for any reason?

Despite these shortcomings, Oberkampf has successfully used Table 2.1 in guided discussion with students taking his V&V training class. The discussion helped students realize how much analyst judgment and how little V&V rigor is present in their analysis activities.
Scoring and targets are distractions that undermine the value of PCMM as a framework for simulation quality and communication. A better approach is to seek descriptive answers to seven open-ended questions.

1. What are the initial and boundary conditions, and why do you believe their implementation is adequate?
2. What is the model's geometric (or representational) fidelity, and why do you believe it is adequate?
3. What are the physics and material models that are needed, what are your capabilities, and what are the research needs?
4. Why do you believe that code bugs and algorithm deficiencies are not corrupting simulation results?
5. Why do you believe numerical errors are not biasing simulation results?
6. How accurate and application-relevant are the models?
7. Do uncertainties in simulation results impact decisions, and what are the dominant contributors to uncertainty?

PCMM and the process proposed here serve different purposes, but the seven PCMM questions are well represented in the regulatory process proposed here (Section 0). The first three questions address the simulation conceptual model (Section 0). Questions 4 and 5 deal with simulation solution errors (Section 0). Question 6 addresses the accuracy of the conceptual model (Section 0). Question 7 addresses the impact of uncertainties on decisions (Section 0).

Table 2.1: Predictive Capability Maturity Model (PCMM)

MATURITY ATTRIBUTE	**Maturity Level 0** **Low Consequence,** **Minimal M&S Impact,** **e.g., Scoping Studies**	**Maturity Level 1** **Moderate Consequence,** **Some M&S Impact,** **e.g., Design Support**	**Maturity Level 2** **High-Consequence,** **High M&S Impact,** **e.g., Qualification Support**	**Maturity Level 3** **High-Consequence,** **Decision-Making Based on M&S,** **e.g., Qualification or Certification**
Representation and Geometric Fidelity **What features are neglected because of simplifications or stylizations?**	• Judgment only • Little or no representational or geometric fidelity for the system and BCs	• Significant simplification or stylization of the system and BCs • Geometry or representation of major components is defined	• Limited simplification or stylization of major components and BCs • Geometry or representation is well defined for major components and some minor components • Some peer review conducted	• Essentially no simplification or stylization of components in the system and BCs • Geometry or representation of all components is at the detail of "as built," e.g., gaps, material interfaces, fasteners • Independent peer review conducted
Physics and Material Model Fidelity **How fundamental are the physics and material models and what is the level of model calibration?**	• Judgment only • Model forms are either unknown or fully empirical • Few, if any, physics-informed models • No coupling of models	• Some models are physics-based and are calibrated using data from related systems • Minimal or ad hoc coupling of models	• Physics-based models for all important processes • Significant calibration needed using Separate Effects Tests (SETs) and Integral Effects Tests (IETs) • One-way coupling of models • Some peer review conducted	• All models are physics-based • Minimal need for calibration using SETs and IETs • Sound physical basis for extrapolation and coupling of models • Full, two-way, coupling of models • Independent peer review conducted
Code Verification **Are algorithm deficiencies, software errors, and poor SQE practices corrupting the simulation results?**	• Judgment only • Minimal testing of any software elements • Little or no SQE procedures specified or followed	• Code is managed by SQE procedures • Unit and regression testing conducted • Some comparisons made with benchmarks	• Some algorithms are tested to determine the observed order of numerical convergence • Some Features & Capabilities (F&C) are tested with benchmark solutions • Some peer review conducted	• All important algorithms are tested to determine the observed order of numerical convergence • All important F&Cs are tested with rigorous benchmark solutions • Independent peer review conducted
Solution Verification **Are numerical solution errors and human procedural errors corrupting the simulation results?**	• Judgment only • Numerical errors have an unknown or large effect on simulation results	• Numerical effects on relevant SRQs are qualitatively estimated • Input/output (I/O) verified only by the analysts	• Numerical effects are quantitatively estimated to be small on some SRQs • I/O independently verified • Some peer review conducted	• Numerical effects are determined to be small on all important SRQs • Important simulations are independently reproduced • Independent peer review conducted
Model Validation **How carefully is the accuracy of the simulation and experimental results assessed at various tiers in a validation hierarchy?**	• Judgment only • Few, if any, comparisons with measurements from similar systems or applications	• Quantitative assessment of accuracy of SRQs not directly relevant to the application of interest • Large or unknown experimental uncertainties	• Quantitative assessment of predictive accuracy for some key SRQs from IETs and SETs • Experimental uncertainties are well characterized for most SETs but poorly known for IETs • Some peer review conducted	• Quantitative assessment of predictive accuracy for all important SRQs from IETs and SETs at conditions/geometries directly relevant to the application • Experimental uncertainties are well characterized for all IETs and SETs • Independent peer review conducted
Uncertainty Quantification and Sensitivity Analysis **How thoroughly are uncertainties and sensitivities characterized and propagated?**	• Judgment only • Only deterministic analyses are conducted • Uncertainties and sensitivities are not addressed	• Aleatory and epistemic (A&E) uncertainties propagated, but without distinction • Informal sensitivity studies conducted • Many strong UQ/SA assumptions made	• A&E uncertainties segregated, propagated, and identified in SRQs • Quantitative sensitivity analyses conducted for most parameters • Numerical propagation errors are estimated and their effect known • Some strong assumptions made • Some peer review conducted	• A&E uncertainties comprehensively treated and properly interpreted • Comprehensive sensitivity analyses conducted for parameters and models • Numerical propagation errors are demonstrated to be small • No significant UQ/SA assumptions made • Independent peer review conducted

2.8 Summary of lessons learned

Key lessons learned are summarized here:

1. A2A uncertainty can be reduced with a common modeling approach but can never be eliminated.
2. The essential characteristics of BEPU, QMU, and PRA are that assessments are unbiased with an objective accounting of all dominant contributors to uncertainty. Decision criteria are the only place where conservatism is encouraged.
3. QMU (and BEPU) can benefit from the structured methodology and discipline of Probabilistic Risk Assessment (PRA). QMU, BEPU, and PRA are frameworks for integrating what is known and what is not and for integrating physical and computational simulation to the extent that either is available. PRA provides the necessary formalism for characterizing, propagating, and interpreting uncertainties.
4. The formalisms of verification and validation must complement the formulism of uncertainty quantification to identify, quantify, and manage errors and uncertainties in computer simulation results. Verification (code and solution) provides evidence that you are solving the equations correctly and that solution errors are understood and quantified for the target application. Validation provides evidence that you are solving the right equations, and that model form errors and uncertainties are understood and quantified for a hierarchy of validation tests of increasing complexity and relevance to the target application.
5. Prediction uncertainty is a research topic whose methods vary depending on the features of the problem at hand. Two approaches to prediction uncertainty exist. The first explores alternate plausible models when no system data exist. The second uses system-level data to quantify model form error and uncertainty.
6. Methods for expressing model form error (model bias) and assessing its impact on prediction uncertainty are in their infancy. Model form errors are extremely difficult to quantify.
7. BEPU, QMU, and PRA provide input to risk-informed decisions. Do not abdicate decisions to a computed number; other subjective considerations must be weighed based on risk tolerance and the maturity of simulation capabilities.

3. Proposed Process

The proposed BEPU process for FAA seat certification applications is illustrated in Figure 3.1. The proposed process addresses the lessons learned summarized in Section 0. The process is comprised of six main elements, which are briefly summarized here. Section 0 discusses key concepts in greater detail and demonstrates their application for aircraft seat certification.

3.1 Define reality of interest and regulatory decision

This element aims to frame the regulatory decision that needs to be made and establish requirements for regulatory acceptance. The reality of interest broadly describes the system and events with corresponding consequences that require a regulatory decision. Regulatory requirements and expectations must be clear and well documented:

1. Identify what QOIs are relevant to the decisions.
2. What requirements constrain the application, the solution approach, rigor, and how the results will be presented and interpreted?
3. What are the qualitative and quantitative regulatory requirements and the supporting decision metrics?
4. What are the expectations for assessment quality, organizational or independent peer review, and documentation?

The product of the element should be a plan submitted for regulatory approval before significant work is performed.

3.2 Develop conceptual models

The product of this element is a fully specified system capable of being solved. This element should identify what is in, what is out, and why. The conceptual model is an abstraction of the application-specific reality of interest. There can be multiple conceptual models: physical simulations (tests), simulation of tests, and simulation of the target application. Conceptual models must be specified sufficiently to allow an unambiguous solution by physical or computational simulation. A conceptual model comprises three elements:

1. The environments associated with initiating events and scenarios.
2. The system state characterizing the design, geometry, material, demographics, etc., which define the system before the initiating event.
3. The governing physics, including constitutive and material models.

A sharp distinction is made between the specification of the simulation conceptual model and the solution of the simulation conceptual model. Grid, solver parameters, algorithm knobs, etc., are all about the *solution* of the simulation conceptual model and should not be confused with the simulation conceptual model. Evidence is required that the specification of the conceptual model is complete, that errors and sources of uncertainties are understood, and that intangibles and key assumptions are documented.

3.3 Assess simulation solution errors

This element's product is a computational model that identifies, quantifies, and manages computational solution errors. The expectation is that solutions will converge to the correct answer for the intended application. This can only be *inferred* from evidence of code verification and solution verification. Simulation results should be bias-corrected for known sources of solution errors unless negligible. Simulation solution uncertainties will cloud validation assessments and application predictions.

3.4 Assess accuracy of simulation conceptual model

The product of this element is a risk-informed decision to accept or reject the model for regulatory predictions and quantification of the model from error and uncertainty. The latter defines the approach to prediction uncertainty adopted in this report (see Appendix A.2). Validation is the model assessment process, from the perspective of intended use, by comparing simulation results with relevant experiment results when both simulation and test results are clouded by uncertainty. Model acceptance is distinguished from model validation and is judged based on the magnitude of the discrepancy between predictions and relevant test data. Model calibration is an empirical adjustment of the model to improve agreement with data with the intent of reducing prediction errors and uncertainties. Confidence in the model is derived by applying the process to a hierarchy of validation tests of increasing complexity and application relevance.

3.5 Integrate risk

This step's product quantifies the required regulatory metrics and identifies dominant sources of uncertainty. Predicted results for the certification design are bias-corrected (positive or negative) for known sources of errors and their uncertainties. Model form errors and uncertainties are quantified in the previous element. Acceptable lumbar loads are prescribed without uncertainty in 14 CFR Part 25.562.

3.6 Make regulatory decision

The product of this element is a risk-informed decision to accept or reject a proposed design based on CbA. Quantitative input to risk-informed decisions is the decision metric (e.g., a factor of safety, FoS) and sensitivity of results to uncertainties. Other subjective factors that could inform the regulatory decision are:

1. Environmental factors of the regulatory agency: Is there congressional oversight and mandates? What is the regulatory tolerance to risk, the complexity of the decision, and experience with this type of decision?
2. Corroborating evidence: Is there a balance between testing and computational simulation? Does the process complement the FAA’s current approach to CbT and CbA?
3. Credibility of the assessments: Is there evidence of completeness and correctness (VVUQ), communicated in a forthright and understandable manner, and documented for the record with sufficient detail that test results and simulation results can be recreated?
4. Findings of regulatory review and independent per review, as appropriate.

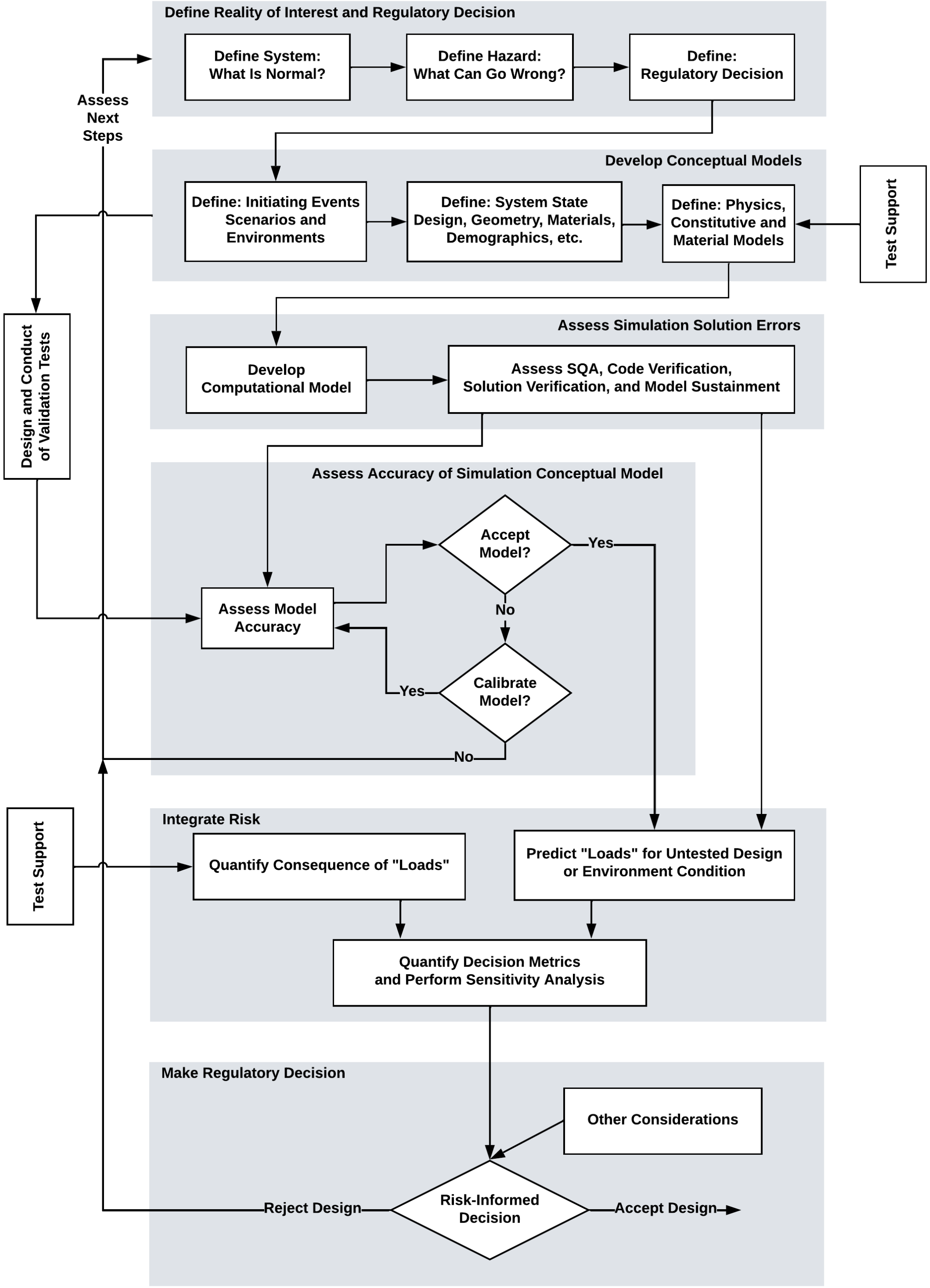

Define Reality of Interest and Regulatory Decision
Define System: What Is Normal?
Define Hazard: What Can Go Wrong?
Define: Regulatory Decision
Assess Next Steps
Develop Conceptual Models
Define: Initiating Events Scenarios and Environments
Define: System State Design, Geometry, Materials, Demographics, etc.
Define: Physics, Constitutive and Material Models
Test Support
Design and Conduct of Validation Tests
Assess Simulation Solution Errors
Develop Computational Model
Assess SQA, Code Verification, Solution Verification, and Model Sustainment
Assess Accuracy of Simulation Conceptual Model
Accept Model?
Yes
No
Assess Model Accuracy
Calibrate Model?
Yes
No
Test Support
Integrate Risk
Quantify Consequence of "Loads"
Predict "Loads" for Untested Design or Environment Condition
Quantify Decision Metrics and Perform Sensitivity Analysis
Make Regulatory Decision
Other Considerations
Risk-Informed Decision
Reject Design
Accept Design

Figure 3.1: Proposed BEPU process for FAA applications

4. Key Concepts and Demonstration of the Proposed Process

The fundamental concepts of each process element and sub-element will be motivated and discussed in detail. Personal experience and best practices will be highlighted where appropriate. The goal is to develop a broad understanding of expectations (the “whats”) and risks when using computational simulation in regulatory decisions.

A detailed example will also demonstrate the process, including the steps necessary to demonstrate certification, identifying what testing is needed, and showing how the model can be accepted. The methods employed in the report are not the only ones that meet the expectations of the process elements. The “hows” are not unique and may be application-specific. There may be better methods than those demonstrated here.

The demonstration offers the opportunity to illustrate the pros and cons of different methods and approaches. This opportunity does not exist in a regulatory submittal. For example, the regulatory decision metric is computed in two ways to illustrate separate ways of considering uncertainties in decisions.

The demonstration will be tailored so that all computer simulations can be performed with MS Excel. Monte Carlo simulations will be performed with @Risk[10], a commercial risk and decision analysis platform that integrates with MS Excel. Using Excel and @Risk adds a desirable element of practicality and transparency to the demonstration. However, real-world seat designs and certification simulations are not expected to be performed with MS Excel.

Finite element codes can be computationally expensive, eight hours for a complete simulation of an aircraft seat design; consequently, the number of computationally expensive simulations to achieve CbA is a critical consideration. Industry believes CbA can be achieved with two simulations: one simulation of a baseline seat design to validate the model and a second simulation of a nearby seat design seeking CbA. This may be optimistic or overly simplistic, but the author is sensitive to the fact that many more simulations can be performed with a simple spreadsheet model than with high-fidelity finite element solutions. The methods demonstrated here are motivated by industries' need to minimize the number of computationally expensive simulations. The demonstration will discuss computational burden as if a spreadsheet solution were a computationally expensive finite element solution.

[10] Palisade.Lumivero.com

4.1 Define reality of interest and regulatory decision

This element aims to frame the regulatory decision that needs to be made and establish requirements for regulatory acceptance. The reality of interest broadly describes the system and events with corresponding consequences that require a regulatory decision. Regulatory requirements and expectations must be clear and well documented:

1. Identify what QOIs are relevant to the decisions.
2. What requirements constrain the application, the solution approach, rigor, and how the results are to be presented and interpreted?
3. What are the qualitative and quantitative regulatory requirements and the supporting decision metrics?
4. Expectations for assessment quality, organizational or independent peer review, and documentation.

The product of the element should be a technical plan submitted for regulatory approval before significant work is performed.

4.1.1 What is normal and what can go wrong?

The scenario of interest is defined in 14 CFR Part 25.562 as an emergency landing creating the potential for lumbar injuries; see Figure 4.1. The focus will be on transport category aircraft with a maximum takeoff weight (MTOW) > 12500 lbs.

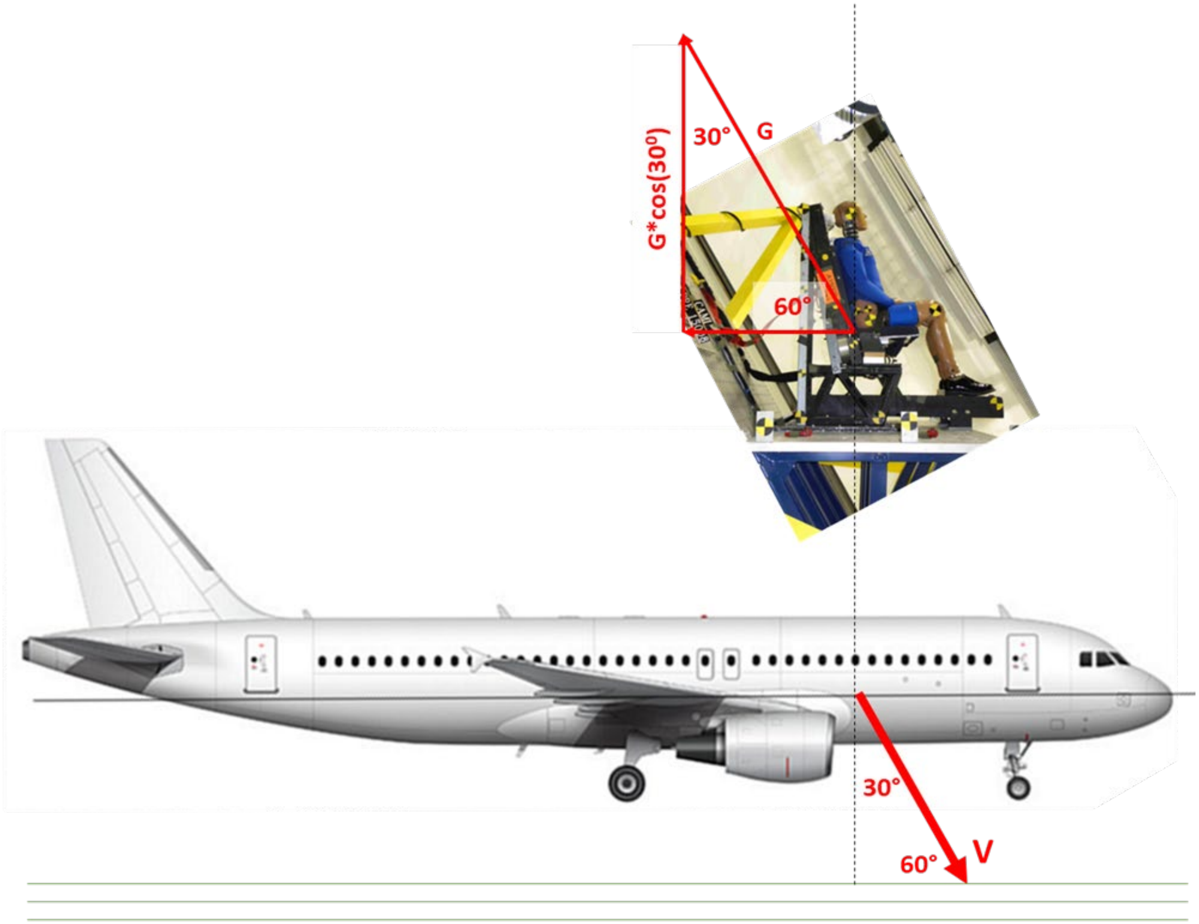


Figure 4.1: Potential for lumbar injuries during an emergency landing

4.1.2 Regulatory decision

The applicant requests certification of a proposed “nearby” seat design using computational simulation (CbA) without testing it. It is assumed that the baseline seat design has already been CbT (CAMI A15008). The nearby design differs from the baseline design in that the cushion thickness was increased from 2.0” to 2.5”.

Qualitative requirements are given in 14CFR Part 25, “Protect each occupant during an emergency landing condition when proper use is made of seats, safety belts, and shoulder harnesses provided for in the design, and the occupant is exposed to the loads resulting from the conditions prescribed.” 14CFR Part 25 prescribes a quantitative requirement related to lumbar injuries that is consistent with the above goal, i.e., the maximum lumbar load, L, as measured in an approved ATD must be less than L_{req} = 1500 lb_f, which can be expressed in terms of a factor of safety, FoS = $L_{req}/L > 1.0$. The origins and interpretation of this requirement are discussed in Section 0.

Table 1.1 summarizes the regulatory requirements and the requested regulatory decision. Note that the demonstration is for a hypothetical seat design. These are consistent with the CAMI

A15008 test, where the maximum lumbar load was measured to be 1048 lbf, and the FoS is 1.43.

Table 4.1: Summary of the regulatory requirements and the requested regulatory decision

	Baseline Design	**Nearby Design**
Requirements	14CFR25.562	14CFR25.562
Load <	1500 lb$_f$	1500 lb$_f$
Environments	14CFR25.562	14CFR25.562
Triangular Pulse G$_{max}$	14	14
Rise Time (ms)	80	80
Impact Angle	30^o	30^o
Passenger	14CFR25.562	14CFR25.562
ATD Weight	170 lb	170 lb
ATD	FAA-Hybrid III	FAA-Hybrid III
Positioning	Seated Upright	Seated Upright
Seat Design		
Frame	Rigid	Rigid
Seating	Single	Single
Monolithic Cushion	CF42 (AC)	CF42 (AC)
Cushion Thickness	2.0”	2.5”
Results	**Test**	**Predicted**
Test	CAMI A15008	
Lumbar Load	1048 lb$_f$	
Decision Metric		
FoS = L$_{req}$/L	1.43	

4.1.3 Technical plan to demonstrate CbA

This section describes planned activities that will generate the evidence necessary to support CbA of the proposed new design. The plan is organized around the process depicted in Figure 3.1. Ideally, a formalized plan should be submitted to the regulatory agency or their independent peer review panel for prior comment and approval before the start of work; however, that was not an option for this demonstration.

Plan to frame the regulatory decision

Figure 4.2 storyboards what the simulation results will look like for a design seeking CbA. Only epistemic uncertainties are represented, so “probability’ is interpreted as the degree of belief. The black line is the computed distribution of uncertain lumbar loads. The green line is FAA’s acceptance threshold.

Once certified, the population of fielded seats will vary. This variability results from variability in materials and manufacturing and is impossible to predict with simulation because the dominant contributors and their quantification do not exist. Aleatory uncertainties are impossible to predict with computational simulation and are important intangibles at the time of a regulatory decision. Seat-to-seat variability can only be estimated by sampling and testing a posteriori.

There are two approaches for generating the distribution shown in Figure 4.2: assessment of alternate plausible models and model form error and uncertainty. The distinction is addressed in greater detail in Appendix A. The first is not computationally practical; consequently, prediction uncertainty based on quantifying model form error and uncertainty is adopted here. The literature has abundant research data for stylized seats, and industry likely has an extensive database for actual seats. These data are necessary to support the adopted approach to prediction uncertainty.

The decision metric selected for this demonstration is the factor of safety, $FoS = L_{req}/L$, where L_{req} is the regulatory threshold and L is the assessed load. The FoS is preferable over the margin of safety, $MoS = L_{req} – L$, because FoS is dimensionless, allowing comparison and prioritization with other regulatory decisions, e.g., head injuries based on assessment of HIC. The FoS must be modified to reflect how uncertainties will impact decisions explicitly. One choice is to regulate on the median, $FoS = L_{req}/L_{50}$, and learn from the uncertainties, where L_{50} is the 50th percentile of the distribution. This is most like the current regulatory approach. The second choice is to regulate with the intent of high confidence, $FoS = L_{req}/L_{95}$, where L_{95} is the 95th percentile of the distribution. The choice of FoS would be prescribed upfront in the typical regulatory environment, but in this demonstration, both will be computed to explore the pros and cons of each. Conceptually, $FoS > 1$ will be used for regulatory acceptance in this demonstration. However, motivation for larger values will be discussed.

The confidence factor, $CF = M/U = (L_{req}\text{-}L_{50})/(L_{95}\text{-}L_{50})$, is another candidate for the decision metric. The confidence factor is commonly associated with the QMU community, especially at the three nuclear weapons laboratories. The FoS is selected over CF because the latter involves two percentiles of the distribution, making it hard to converge for small margins. More importantly, the CF evaluation is incompatible with using Wilks' formula (Appendix F.4), an important method for computing distribution percentiles with confidence when computer simulations are computationally expensive.

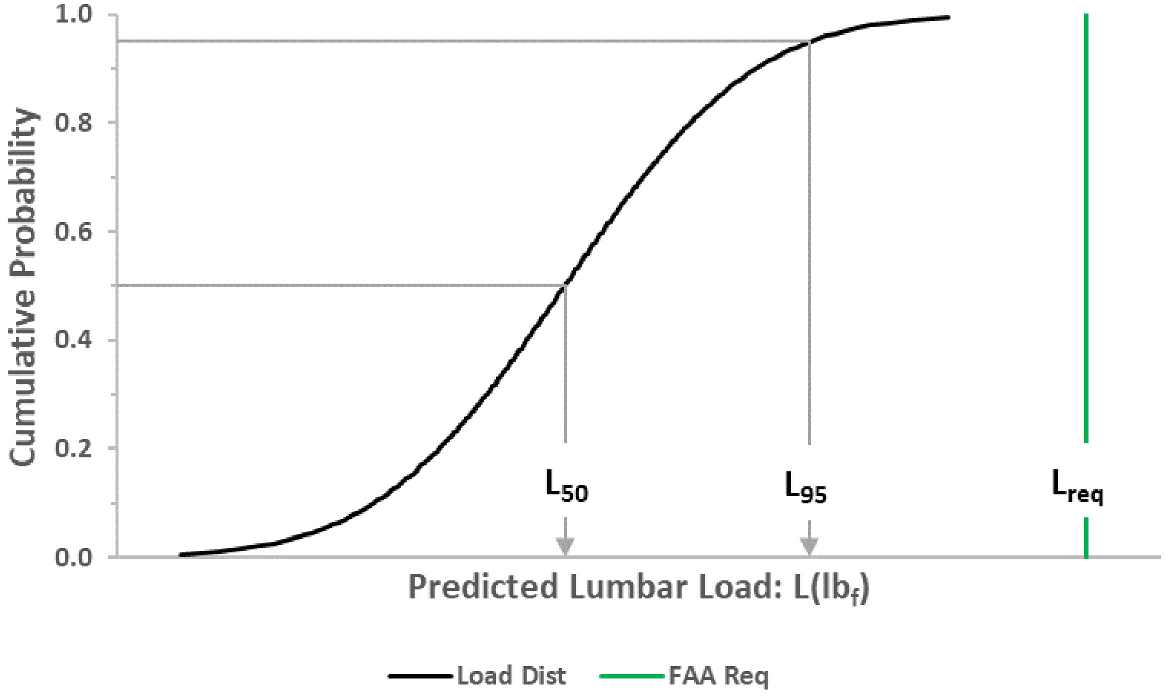


Figure 4.2: Storyboard *of* what the results will look like

Plan to develop conceptual models

The conceptual model is an abstraction of reality. There are three conceptual models to consider: the conceptual model for physical testing, the conceptual model for simulation of system-level validation tests, and the conceptual model for simulation of the certification design. The first two are identical in this demonstration, and the third differs only because the cushion thickness is slightly larger.

The conceptual model is defined in terms of environments, system state, and physics. The first two are fully prescribed by regulations or in the design submittal. Physics models will also be fully specified to quantify model form error and uncertainty. An extended phenomena identification and ranking table (ePIRT) will be developed to guide the level of fidelity needed in the conceptual model.

The FAA has agreed to perform material testing for CF42 (AC) foam to support this project. The test matrix will provide data for the full range of compressions and compression rates relevant to scenarios and environments prescribed by the FAA. The data will be used to fit parameters in a proposed constitutive model. An analytic solution exists for the initial static compression that results when the ATD is first positioned on the seat before a dynamic sled test.

The system will include the ATD and the seat design. The system will be modeled as a one-dimensional spring-damper and mass system. The upper and lower torsos of the ATD will be modeled as rigid masses. The cushion will be modeled as a massless spring and damper. The frame will be treated as rigid, so the boundary conditions prescribed by the FAA will be applied directly to the seat pan, i.e., the bottom of the cushion. The FAA prescribes the scenario and associated boundary conditions. A numerical solution of the system of equations is required; however, an analytic solution exists for the case where the seat has no cushion.

Plan to assess simulation solution errors

A computational model will be implemented into MS Excel to provide numerical solutions to the simulation conceptual model. The solution algorithm will employ a first-order explicit time-integration scheme. Monte Carlo simulations will be performed with @Risk.

A suite of tests will be developed to ensure that coding bugs, algorithm deficiencies, or platform inconsistencies do not introduce errors into simulation results. The tests will include regression tests with known referents, a verification test with an analytic solution, and an application-relevant acceptance test. The verification test is a nearby, but linear, spring-mass system for which an analytic solution will be derived. Even slight deviations between the observed convergence rate and the formal order of convergence are a sensitive indicator of code bugs or algorithm deficiencies. The numerical solution for the baseline seat design will be used as an application-relevant acceptance test to be evaluated over time to ensure that simulation results are stable to code and platform updates.

Discretization errors will be quantified and bounded using Richardson extrapolation and Roache's grid convergence index (GCI) for validation simulations of the baseline design and simulations for the nearby certification design. The goal will be to render discretization errors negligible if possible. Observed and formal order of convergence rates will be compared as a further test of code bugs or algorithm deficiencies.

Plan to assess accuracy of simulation conceptual model

The accuracy of the simulation conceptual model will be assessed for a hierarchy of validation tests of increasing relevance to the application. Model form error will be assessed for each element in the validation hierarchy using the discrepancy measure,

$$E = \ln \frac{M}{P} \,, \qquad \textbf{0-1}$$

which limits to the relative error when predictions (P) are close to measurements (M). This metric honors the physical constraint that neither M nor P should be negative. All validation tests will represent pass/fail gates based on an assessment of model bias, which should be small, $|E_{50}| < 0.10$, at the application design point (DP). In general, E will be uncertain because of uncertainties in either P of M or both, but validation acceptance will be based on the median. The calibration of the CF42 (AC) material constitutive model is at the physics, or foundational, level of the validation hierarchy. The middle, or transition, tier has two validation activities. First, the analytic model for lumbar loads for seats without cushions will be validated against a database of three relevant tests conducted with 14G environments. Second, the analytic model for initial static compression will be validated against a database of four relevant tests with 2.0" or 4.0" CF42 (AC) foams.

Validation with system-level tests (highest tier in the validation hierarchy) will be used to quantify model form error and uncertainty, E_{mf}, at the design point (DP),

$$\mathrm{E_{mf}(DP) = E_{mf}(trend\ bias, scatter)}\ , \qquad \textbf{0-2}$$

which is assumed to be equal to the model form error assessed for a relevant database. Trend bias refers to the ability of the model to predict trends with respect to environment and design parameters. Scatter reflects uncertainty.

Consistent with current FAA guidance for CbA, validation of the model against a single certification test for the baseline seat design provides an estimate of the model's bias without assessing the model's ability to predict sensitivity to design or environmental parameters. A direct assessment of model form uncertainty (scatter) is not possible with a single test; however, scatter (uncertainty) will be inferred indirectly from an assessment of measured lumbar loads in multiple series of replicate tests conducted for assorted designs (cushion material, thickness) and environments.

A second approach to quantifying model form error and uncertainty involves validating the model against a database of nine relevant system-level tests for 0.0”, 2.0”, and 4.0” CF42 (AC) foam conducted with 14G and 19G environments. Trend bias (with design and environment parameters) and uncertainty (scatter) contributions to model form error and uncertainty can be evaluated in a self-consistent manner.

The pros and cons of each approach for quantifying the model form error will be discussed.

Plan to integrate risk

The quantification of model form error and uncertainty will be applied as a correction to the simulation result at the certification design point, S(DP),

$$\mathrm{L(DP) = S(DP)e^{E_{mf}(DP)}}\ . \qquad \textbf{0-3}$$

This is how the uncertainty distribution in Figure 4.2 will be computed, from which the decision metrics can be calculated as described earlier. In general, uncertainties from the simulation solution and model form must be aggregated to calculate uncertainties in the QOI. Monte Carlo simulations will be used in this demonstration. A sensitivity analysis will be conducted to identify the dominant contributors to uncertainty depicted in Figure 4.2.

Plan for making a risk-informed decision

Demonstration of a regulatory decision will be risk-informed based on quantitative inputs and other more subjective considerations. Quantitative inputs include computed decision metrics and the results of a sensitivity analysis. Two subjective considerations will be addressed in the context of this demonstration.

1. Corroborating evidence: Is there an appropriate balance between testing and simulation, and how does the new process relate to current FAA approaches?

2. Credibility of the assessments: Is there evidence of completeness and correctness communicated in a forthright and understandable manner and documented for the record?

4.2 Develop conceptual models

The product of this element is a fully specified system capable of being solved by either physical simulation or computational simulation — what is in, what is out, and why. The conceptual model is an abstraction of the application-specific reality of interest. Conceptual models must be specified sufficiently to allow an unambiguous solution by physical or computational simulation. A conceptual model comprises three elements:

1. Environments associated with initiating events and scenarios.
2. The system state is defined by design, geometry, material, demographics, etc., which define the system before the initiating event.
3. The governing physics, including constitutive and material models.

A sharp distinction is made between the specification of the simulation conceptual model and the solution of the simulation conceptual model. Grid, solver parameters, algorithm knobs, etc., are all about the *solution* of the simulation conceptual model and should not be confused with the simulation conceptual model. Evidence is required that the specification of the conceptual model is complete, that errors and sources of uncertainties are understood, and that intangibles and key assumptions are documented.

Section 0 declared that model form error and uncertainty is the methodology approach adopted for this study. This means that all conceptual model elements are frozen, and that model form errors and uncertainties are estimated by comparing the fixed model with appropriate system-level data.

The environments associated with initiating events and scenarios are already frozen because they are prescribed in regulations without uncertainty. Likewise, the applicant specifies the seat design (geometry and materials) without uncertainty, and other aspects of the system state (ATD and seating position) are prescribed in regulations without uncertainty. Here, we empathize that physics (as described in Section 0) is also frozen without uncertainty. This includes the constitutive model for CF42 (AC), which will be evaluated using best estimate values of the fitting parameters.

There can be multiple conceptual models: physical simulations (tests), simulation of tests, and simulation of the certification design. All three conceptual models are identical for this demonstration, except the certification design has a slightly thicker cushion.

The US Nuclear Regulatory Commission (NRC) championed the use of phenomenon identification and ranking tables, or PIRTs (Boyack et al., 2001; NRC, 1989). Sandia National Laboratories has also had success using PIRTs within its nuclear weapons program and when communicating with external peer review panels. PIRT is an application-specific tool for organizing and communicating:

1. What physics capabilities are needed, and what capabilities are not needed. It is important to identify both and provide evidence or documented rationale justifying the choice.
2. Sufficiency of existing physics capabilities within analysis tools (codes) to meet application needs. PIRT answers the question, do you have the needed capabilities for assessment?
3. What gaps need to be addressed? This drives capability development and research activities.
4. Efficiency of planned activities needed to address gaps in capabilities. Do only what is necessary.

PIRT is commonly a subjective process involving key stakeholders as appropriate for the issue's importance. Key stakeholders might include subject matter experts from the analysis community, academia, code developers, industry, and regulatory agencies. The NRC also championed the use of formal scaling as a means of ranking phenomena (Zuber et al., 1998). Implementing complex models involving multi-physics and disparate time scales can be challenging. PIRTs are living documents that can change over time as understanding changes or new capabilities are developed and implemented in codes.

The original focus of PIRT was phenomena, i.e., physics and material models. The scope of PIRT has been expanded here to formally include all three elements of the simulation conceptual model: environments, system state, and physics. My experience is that elements of environments and system state would creep into PIRTs in an ad hoc manner anyway. The new tool is called ePIRT for extended PIRT. Appendix B describes an ePIRT developed for this project, which guides modeling approaches described in Sections 0, 0, and 0.
Two capability gaps are identified in the ePIRT:

1. Compliance of ATD lower and upper torsos.
2. One-dimensional approach to modeling.

It’s expected that these gaps will not introduce dominant first order effects in the demonstration and would not exist in high-fidelity finite element modeling. The upper torso of the ATD will be treated as rigid in this report.

4.2.1 Define initiating events, scenarios, and environments

Emergency landing conditions is the scenario of interest defined by the FAA. Emergency landing conditions are conditional on the scenario and the class of aircraft. Environments associated with an emergency landing are summarized in Table 4.2. Figure 4.3 compares the acceleration environments, represented as triangular pulses, experienced by the seat for each class of aircraft.

This demonstration assumes a landing accident for a transport category airplane (MTOW>12500 lbs) as described in 14 CFR Part 25.562. These environments apply to all three conceptual models: physical simulations (tests), simulation of tests, and simulation of the certification design.

Table 4.2: Environments by aircraft type associated with emergency landing conditions

14CFR	Max G	Rise Time (ms)	Impact Angle (deg)	Comments
Part 23.562a	19	50	30	Normal, utility, acrobatic, commuter category airplanes, MTOW<12500 lbs, first row seats
Part 23.562b	15	60	30	Normal, utility, acrobatic, commuter category airplanes, MTOW<12500 lbs
Part 25.562	14	80	30	Transport category airplanes, MTOW>12500 lbs
Part 27.562	30	31	60	Part 27: Rotocraft, MGW<7000 lbs, passenger capacity<=9
Part 29.562	30	31	60	Part 29: Rotocraft, above Part 27 limits

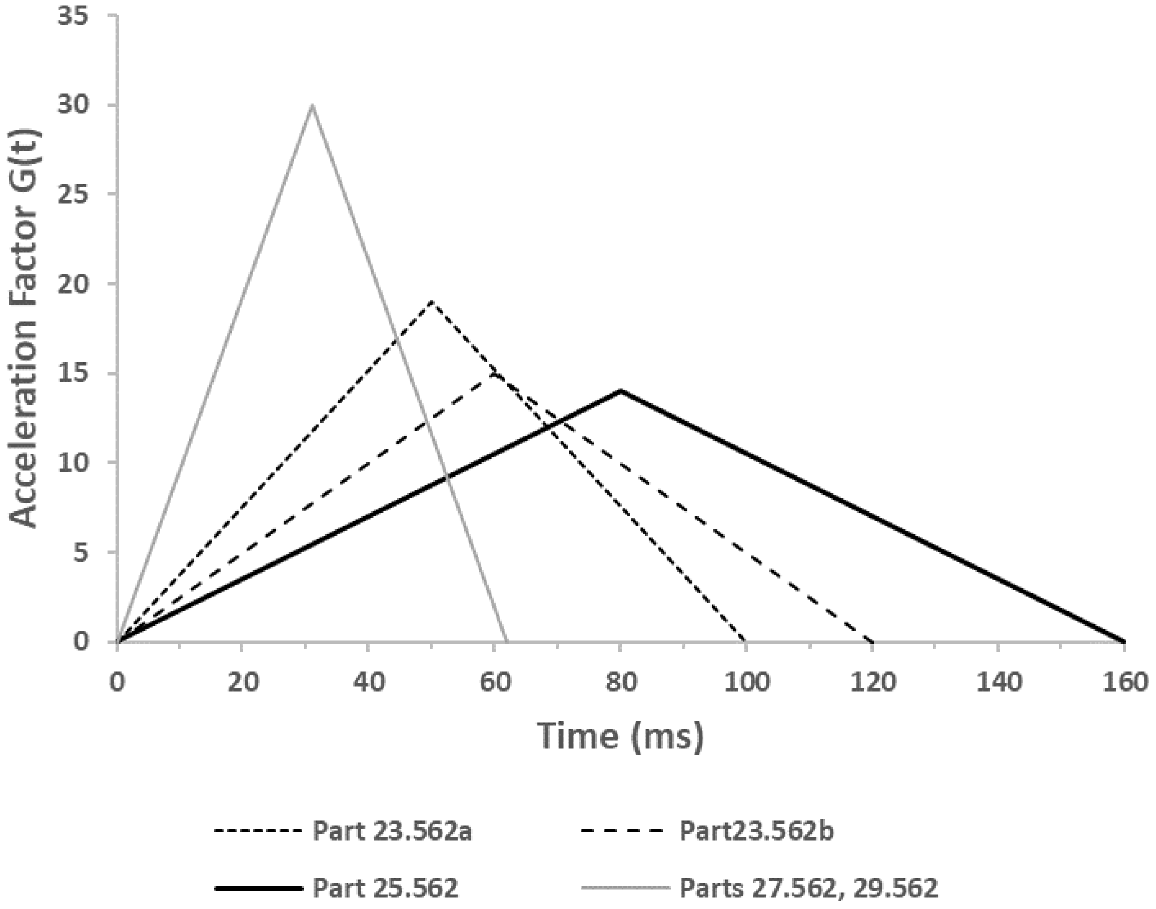


Figure 4.3: Acceleration environments by aircraft type for emergency landings

4.2.2 Define system state

Table 4.3shows that all three conceptual models map directly to the reality of interest for the system state. As prescribed in regulations, an FAA-approved ATD is specified as the reality of interest. The FAA's approach to requirements specification does not assume that an ATD is fully representative of the population of airline passengers. This will be discussed further in Section 0.

Table 4.4 shows the estimated weight distribution for an FAA-Hybrid III ATD. Two weights are of primary interest: the upper torso (UT) weight, which defines loads at the point of the lumbar load cell, and the upper body (UB) weight, which loads the cushion. The bare weights are taken from (Olivares, 2013), where additional judgment is required to parse reported pelvis/abdomen weights into UT and LT weights. (SAE, 2021) provides estimates for clothing and clavicle weights. These weights are slightly less than the 170 lb prescribed by regulations. (SAE, 2021) discusses a process for adding distributed weights to bring the total ATD weight up to the 170 lb prescribed in regulations.

Flight attendants will instruct passengers to assume the brace position if there is sufficient warning; however, the FAA does not consider the brace position to be the best estimate. The FAA states that most crashes will not have sufficient time for passengers to assume the brace position; consequently, the FAA prescribes that the ATD be seated upright when assessing lumbar loads through testing or simulation.

The applicant prescribes the seat design. Here, a hypothetical seat design is specified, making the demonstration practical and allowing connection to a research database of publicly available sled tests. A single seat is assumed to be attached to a single rigid frame without armrests. The cushion is a 2.0" monolithic Confor-42 (AC)[11] foam without a cover. The nearby seat design differs from the baseline only in that the cushion is slightly thicker.

*Table **4**.3: Conceptual models for system state*

		Conceptual Models	
	Reality of Interest	**Baseline Design: Physical and Computational Simulation**	**Nearby Design: Computational Simulation**
Passenger Demographics	**14 CFR Part 25.562**		
ATD	FAA-Hybrid III	FAA-Hybrid III	FAA-Hybrid III
Weight (lb)	170	170	170
Positioning	Seated upright	Seated upright	Seated upright
Seat Design	**Defined by Applicant**		
Frame	Rigid frame no armrests	Rigid frame no Armrests	Rigid frame no armrests

[11] Confor^{TM} is refers to a family of rate dependent foams commonly used in aircraft seats. Confor-42 is one in the class of Confor foams. Confor-42 foam underwent a reformulation to meet new fire safety standards. The reformulated foam is referred to as Confor-42 (AC), which will be shortened to CF42 (AC) in the remainder of this report. Although the properties of CF42 (AC) are like the original formulation, they are not identical, so it is important to maintain the distinction.

Number of Seats	1	1	1
Cushion	CF42 (AC) monolithic foam w/o cover	CF42 (AC) monolithic foam w/o cover	CF42 (AC) monolithic foam w/o cover
Thickness	2.0” baseline design 2.5” nearby design	2.0”	2.5”

Table 4.4: Estimated weights (lb) for FAA-Hybrid III ATD

Contributors to Weight	Bare (Olivares, 2013)	+ Clothing & Clavicle (SAE, 2021)	+ Distributed Weight (SAE, 2021)	Comments
Total	**166.0**	**168.5**	**170.0**	1.5 lbf distributed wgt < +/- 3 lbf spec
Upper Torso (UT)	81.3	81.3	82.1	Above load cell, loads lumbar
Lower Torso (LT)	20.4	20.4	20.6	Below load cell
Upper Body (UB=UT+LT)	101.8	101.8	102.7	UB=UT+LT, loads cushion
Upper Torso Wgt (lbf)	**81.3**	**81.3**	**82.1**	Above load cell, loads lumbar
Head	10.0	10.0	10.1	
Neck	3.4	3.4	3.4	
Upper chest	37.9	37.9	38.2	
Upper arms	8.8	8.8	8.9	
Lower arms (half)	3.8	3.8	3.8	Wgt carried by elbows transmitted to UT
Pelvis/abdomen above load cell	17.5	17.5	17.7	Linear estimation from LT cutaway
Clothing	0.0	0.0	0.0	SAE ARP5765 Rev. B, negligible
Clavicle	0.0	0.0	0.0	SAE ARP5765 Rev. B
Lower Torso Wgt (lbf)	**20.4**	**20.4**	**20.6**	Below load cell
Pelvis/abdomen	37.9	37.9	38.2	
Pelvis/abdomen above load cell	-17.5	-17.5	-17.7	
Clothing	0.0	0.0	0.0	SAE ARP5765 Rev. B, negligible
Lower Body Wgt (lbf)	**64.3**	**66.8**	**67.3**	Does not load cushion under pelvis
Lower arms (half)	3.8	3.8	3.8	Wgt carried by wrists transmitted to LT
Hands	2.5	2.5	2.5	Rests on thigh
Upper legs	34.0	34.0	34.3	
Lower legs	24.0	24.0	24.2	
Feet	0.0	0.0	0.0	
Clothing	0.0	2.5	2.5	SAE ARP5765 Rev. B, Shoes

4.2.3 Define physics and constitutive models

Initial static compression

The cushion is compressed during the seating of the ATD. This initial static compression, f_0, serves as an initial condition for the dynamic event. The procedure is to position the ATD with the seat in the upright position in the lab frame of reference and tension the lap belt (
Figure 4.4). According to (DeWeese & Gowdy, 2002) "Standard practice for setting the belt tension before a sled test is to tighten the belt until two fingers can be comfortably placed between the belt and the ATD's abdomen." In human trials, participants were instructed to tension the lap belt in preparation for an emergency landing. The average tension was 5.55 lb_f, but the range of results was substantial (
Figure 4.4). This is equivalent to a 4.81 lb_f compressive load on the cushion (Table 4.5).

Table 4.5: Lap belt tensioning for 1182 human trials (DeWeese & Gowdy, 2002)

Belt angle (referenced to horizontal) q	60^0
Avg emergency tension (lb_f)	5.55
Avg lap belt compressive load (lb_f)	4.81

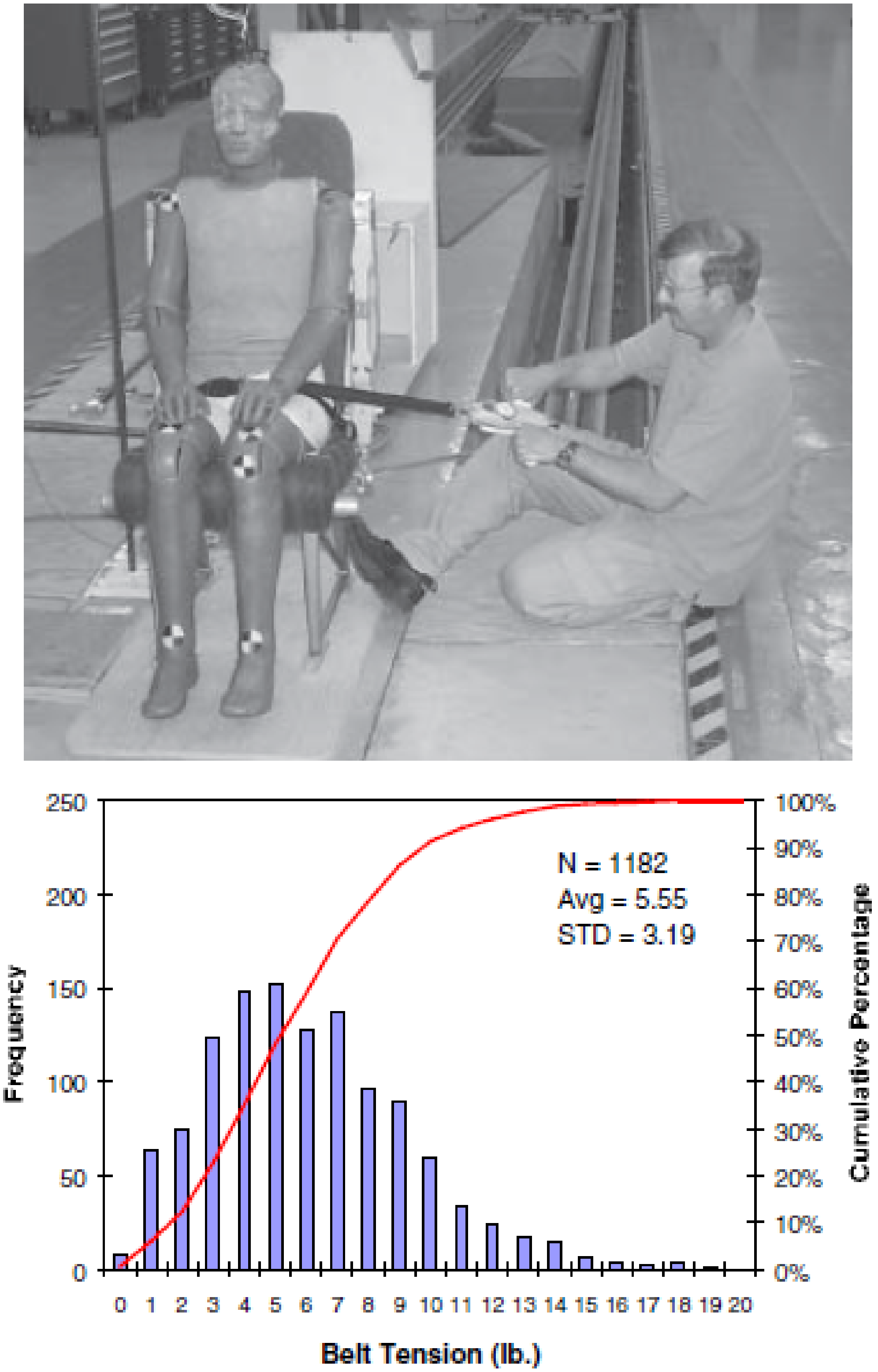


Figure 4.4: Seating of the ATD and tensioning of the lap belt (DeWeese & Gowdy, 2002)

The conceptual model is depicted in Figure 4.5. The initial compressive load (W_0) on the cushion balances the quasi-static resistance of the cushion, $F_{qs}(f_0)$,

$$\mathrm{W_0 = F_{qs}(\varphi_0) = W_{UB} + W_{LapBelt}} \;, \qquad \textbf{0-4}$$

where W_{UB} is the weight of the upper body of the ATD, and $W_{LapBelt}$ is the tension in the lap belt that contributes to cushion compression. Quasi static resistance of the cushion to compression is given by (see Appendix C.3),

$$\mathrm{F_{qs}(\varphi) = \frac{F_0}{\left(1-\frac{\varphi}{\varphi_c}\right)^{a}}}\,, \qquad \textbf{0-5}$$

for f > 10%. Combining Equations 0-4 and 0-5 results in a simple analytic expression for the initial static compression,

$$\varphi_0^* = \frac{\varphi_0}{\varphi_c} = 1 - \left(\frac{F_0}{W_0}\right)^{1/a} = 1 - \pi_1^{1/a}\,. \qquad \textbf{0-6}$$

Normalization of the initial static compression, φ_0^*, results in one physics scaling group,

$$\pi_1 = \frac{F_0}{W_0}\,, \qquad \textbf{0-7}$$

which necessarily must exceed unity because the form of F_{qs} is limited to the plastic regime.

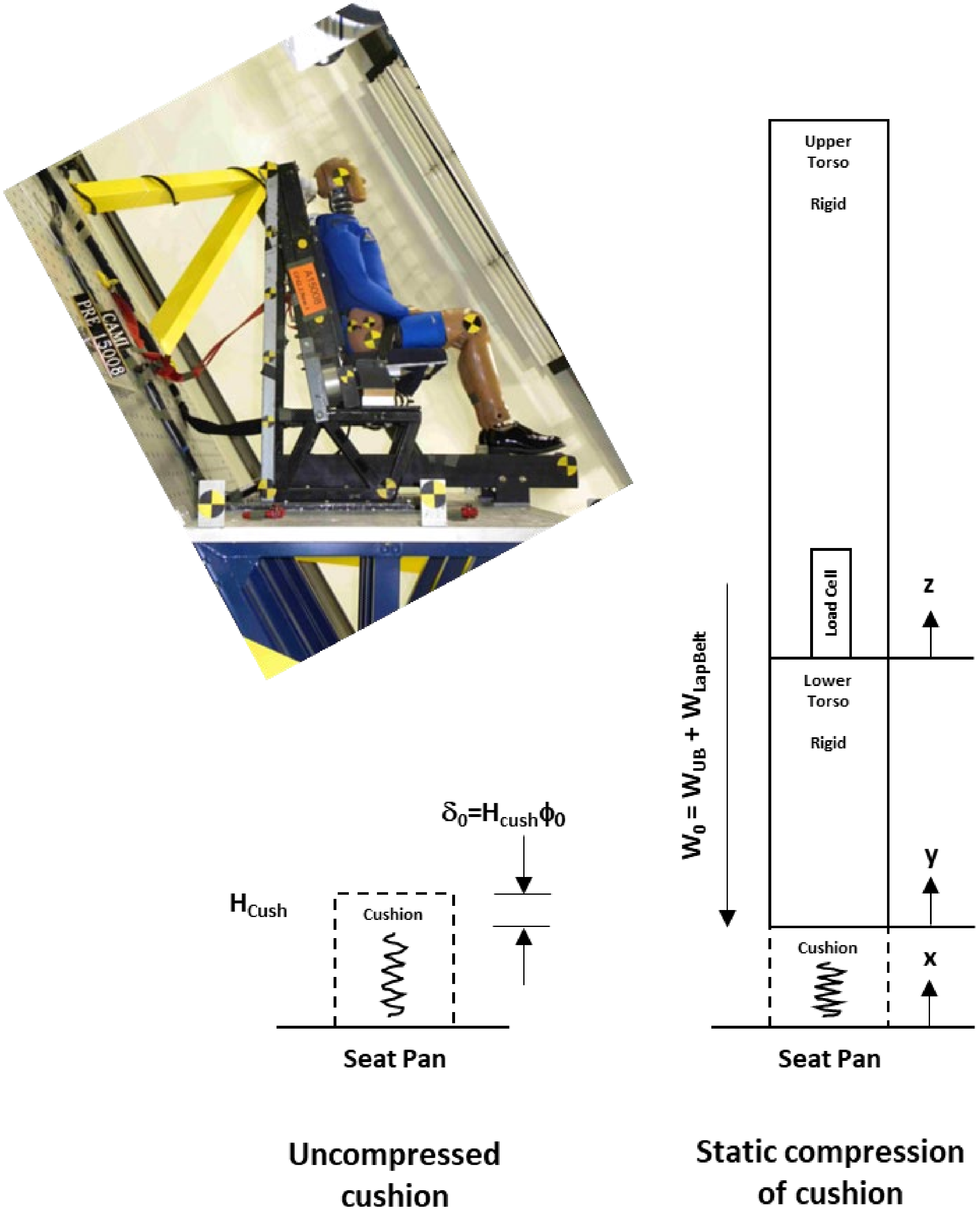


Figure 4.5: Initial static compression of cushion during seating of ATD

Seat pan boundary condition

The seat experiences a strong upward acceleration pulse (in the seat frame of reference) when the aircraft strikes the ground during an emergency landing. The environment experienced by the seat frame is prescribed in 14 CFR Part 25.562 for transport category aircraft and summarized in Table 4.2. Assuming a rigid seat frame, the seat pan under the cushion experiences the same environment. The triangular acceleration pulse (Figure 4.3) is characterized in terms of constant jerk (J), where J is the rate change of acceleration. The environment experienced by the ATD in the direction of the lumbar column (Figure 4.6) is given by,

$$\begin{aligned} \mathrm{J}\cos\theta = \ddot{\mathrm{x}} &= \frac{\mathrm{G_{max}g^0}\cos\theta}{\mathrm{t_{rise}}} \qquad \text{for } \mathrm{t} \le \mathrm{t_{rise}} \\ \ddot{\mathrm{x}} &= -\frac{\mathrm{G_{max}g^0}\cos\theta}{\mathrm{t_{rise}}} \qquad \text{for } \mathrm{t} > \mathrm{t_{rise}}\ . \end{aligned} \tag{0-8}$$

Note that “x” is measured in the direction of the lumbar column.
Equation 0-8 can be integrated directly subject to the initial conditions,

$$\ddot{x}(0) = \dot{x}(0) = x(0) = 0\ , \tag{0-9}$$

with the resulting solutions given by:

$$\text{for } \mathrm{t}^* \le 1 \qquad\qquad\qquad \text{for } \mathrm{t}^* > 1$$

$$\mathrm{a}^* = \frac{\ddot{\mathrm{x}}}{\mathrm{a_{ref}}} = \mathrm{t}^* \tag{0-10}$$

$$\mathrm{a}^* = \frac{\ddot{\mathrm{x}}}{\mathrm{a_{ref}}} = 2 - \mathrm{t}^* \tag{0-11}$$

$$\mathrm{v}^* = \frac{\dot{\mathrm{x}}}{\mathrm{v_{ref}}} = \mathrm{t}^{*2} \tag{0-12}$$

$$\mathrm{v}^* = \frac{\dot{\mathrm{x}}}{\mathrm{v_{ref}}} = -2 + 4\mathrm{t}^* - \mathrm{t}^{*2} \tag{0-13}$$

$$\mathrm{x}^* = \frac{\mathrm{x}}{\mathrm{x_{ref}}} = \mathrm{t}^{*3} \tag{0-14}$$

$$\mathrm{x}^* = \frac{\mathrm{x}}{\mathrm{x_{ref}}} = 2 - 6\mathrm{t}^* + 6\mathrm{t}^{*2} - \mathrm{t}^{*3}\ . \tag{0-15}$$

The reference quantities,

$$a_{ref} = G_{max}g^0\cos\theta \tag{0-16}$$

$$v_{ref} = \frac{1}{2}a_{ref} = \frac{1}{2}G_{max}g^0\cos\theta \tag{0-17}$$

$$x_{ref} = \frac{1}{3}v_{ref} = \frac{1}{6}G_{max}g^0\cos\theta\ , \tag{0-18}$$

normalize the responses to unity when

$$t^* = \frac{t}{t_{rise}} \tag{0-19}$$

is unity. The normalized responses are only a function of t*; there are no physics scaling groups controlling the responses. The normalized response of the seat pan in the direction of the lumbar column is shown in Figure 4.7.

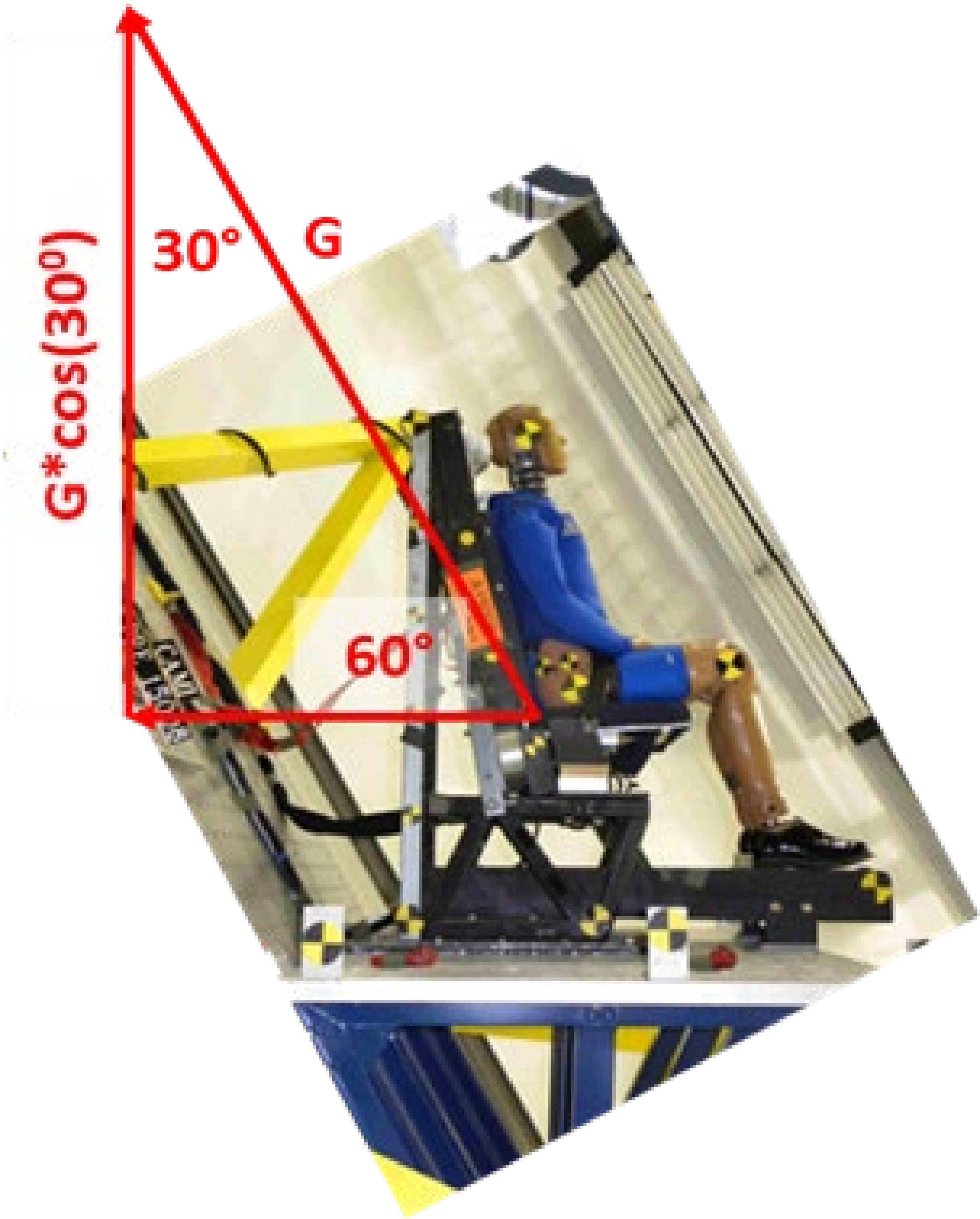


Figure 4.6: Environment experienced by the seat pan and ATD

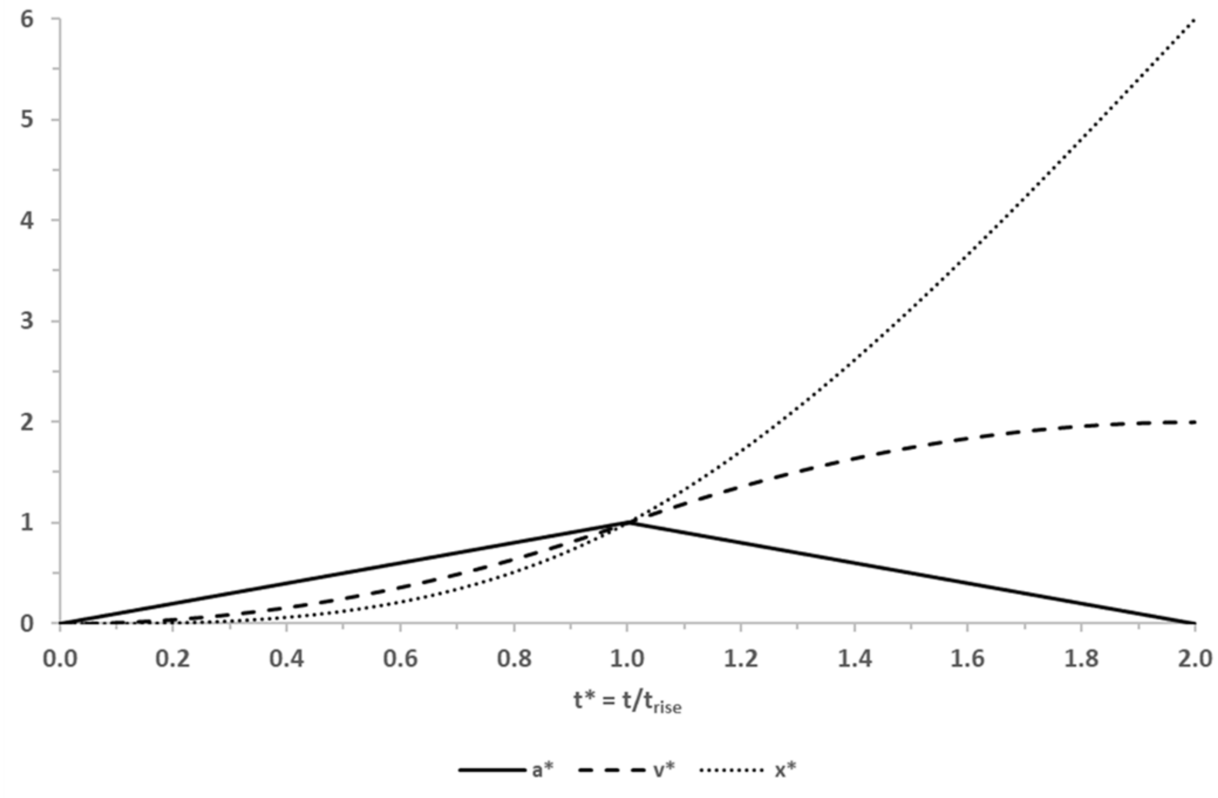


Figure 4.7: Normalized response of the seat pan in the direction of the lumbar column

Lumbar loads for seat without cushions (or rigid cushions)

Consider the case where the aircraft seat does not have a cushion. This case is useful because it serves as a reference for the more realistic case of seats with cushions. This is also a useful case in the validation hierarchy Figure 4.21.

The conceptual model is depicted in Figure 4.8. The upper torso (UT) is treated as rigid. With this assumption, the force balance on the UT is given by,

$$\mathrm{L}' = \mathrm{W}_{\mathrm{UT}} \cos\theta + \mathrm{M}_{\mathrm{UT}} \ddot{\mathrm{Z}}\,. \qquad \textbf{0-20}$$

The load cell in a sled test is tare corrected so that the reported lumbar loads are associated solely with the dynamic event. In addition, the base of the lumbar experiences the accident environment directly, $\ddot{x}$ at the seat pan, when the frame and LT are considered rigid; consequently,

$$\mathrm{L} = \mathrm{L}' - \mathrm{W}_{\mathrm{UT}} \cos\theta = \mathrm{M}_{\mathrm{UT}} \ddot{\mathrm{Z}} = \mathrm{M}_{\mathrm{UT}} \ddot{\mathrm{x}}. \qquad \textbf{0-21}$$

This equation has an analytic solution when $\ddot{x}$ is given by Equations 0-10 and 0-11.

$$L^* = \frac{L}{L_{nc}} = t^* \quad \text{for } t^* \leq 1\text{ , and} \tag{0-22}$$

$$L^* = \frac{L}{L_{nc}} = 2 - t^* \quad \text{for } t^* > 1\text{ ,} \tag{0-23}$$

where

$$L_{nc} = W_{UT}G_{max}\cos\theta\text{ ,} \tag{0-24}$$

is the maximum lumbar load in the no cushion case. Note that the normalized lumbar load (L*) is only a function of the normalized time (t*). There are no physics scaling groups that control the response.

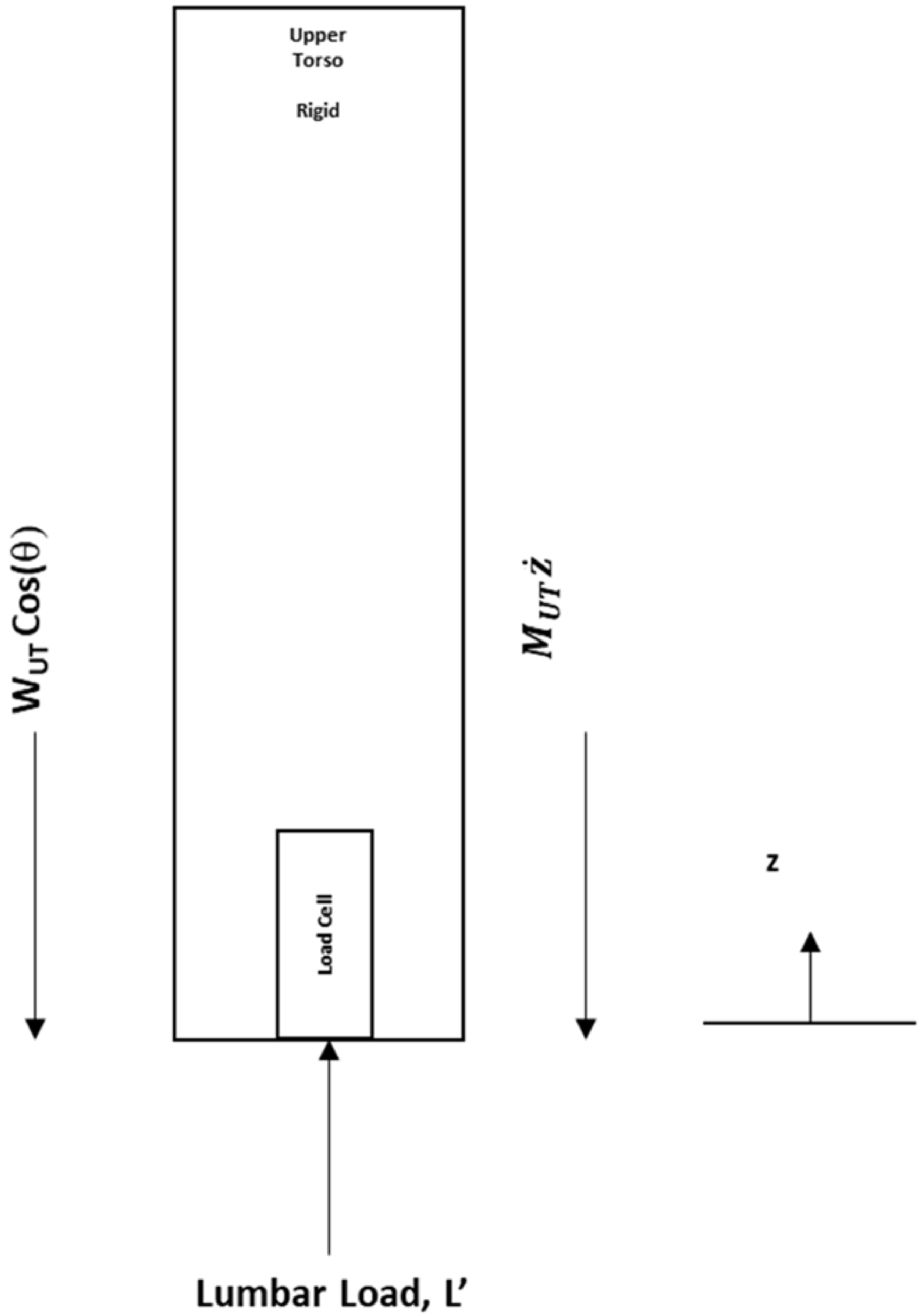


Figure 4.8: Conceptual model for the ATD upper torso (UT)

Foam constitutive model

The foam constitutive model used in this study is a modification of a constitutive model proposed by (Johnson & Cook, 1985) for metals. Details and assessments are presented in Appendix C.3. The foam constitutive model has the form,

$$\frac{F(\varphi,\dot{\varphi})}{F_0}=\frac{1}{\left(1-\frac{\varphi}{\varphi_c}\right)^{\mathrm{a}}}\left\{1+\left(\mathrm{b}+\mathrm{c}\frac{\varphi}{\varphi_c}\right)\max\left[0,\ln\frac{\dot{\varphi}}{\dot{\varphi}_c}\right]\right\}, \qquad \textbf{0-25}$$

There are six fitting parameters in this constitutive model: F_0, f_c, a, b, c, and $\dot{\varphi}_c$. Their interpretation is discussed in Appendix C.3 and values specific to CF42 (AC) are given in Table C.4.

Lumbar loads for seats with cushions

The controlling equations for lumbar loads in the case of seats with cushions are derived from the physics of the ATD upper torso (UT), the ATD lower torso (LT), the ATD upper body (UB), and the cushion, respectively.

The conceptual model for the UT is shown in Figure 4.8. The tare corrected force balance is repeated here,

$$\mathrm{L}=\mathrm{L}'-\mathrm{W}_{\mathrm{UT}}\cos\theta=\mathrm{M}_{\mathrm{UT}}\ddot{z}\,. \qquad \textbf{0-26}$$

Figure 4.9 shows the conceptual model for the ATD LT. Because the LT is treated as rigid,

$$\ddot{\mathrm{z}}=\ddot{\mathrm{y}}\,. \qquad \textbf{0-27}$$

Figure 4.10 shows the conceptual model for the ATD UB. The UB comprises the UT and the LT, and together they load the cushion. A force balance on the UB results in,

$$\mathrm{M}_{\mathrm{UB}}\ddot{\mathrm{y}}+\mathrm{W}_0=\mathrm{F}(\varphi,\dot{\varphi})\,, \qquad \textbf{0-28}$$

where $\mathrm{F}(\varphi,\dot{\varphi})$ is rate-dependent resistance to compression for the cushion (see Equation 0-25).

Figure 4.11 shows the conceptual model of the cushion, which is treated as a massless spring-damper,

$$=\mathrm{x}+\mathrm{H}_{\mathrm{cush}}(1-\varphi)\,, \qquad \textbf{0-29}$$

consequently,

$$\ddot{\mathrm{y}}=\ddot{\mathrm{x}}-\mathrm{H}_{\mathrm{cush}}\ddot{\varphi}\,, \qquad \textbf{0-30}$$

where $\ddot{x}$ is the boundary condition (Equations 0-10 and 0-11) on the seat pan in line with the lumbar.

Combining Equations 0-26, 0-27, and 0-28 yields an equation for the lumbar load,

$$\mathrm{L} = \frac{\mathrm{W_{UT}}}{\mathrm{W_{UB}}}[\mathrm{F}(\varphi, \dot{\varphi}) - \mathrm{W_0}]\,, \qquad \textbf{0-31}$$

and combining Equations 0-28 and 0-30 yields an ODE for the compression dynamics,

$$\mathrm{H_{cush}}\ddot{\varphi} = \ddot{\mathrm{x}} - \mathrm{g}^0\frac{[\mathrm{F}(\varphi, \dot{\varphi}) - \mathrm{W_0}]}{\mathrm{W_{UB}}}\,, \qquad \textbf{0-32}$$

which is driven by the boundary condition on the set pan, $\ddot{x}$, given by Equations 0-10 and 0-11, and subject to the initial conditions,

$$\mathrm{L}(0) = 0\,, \qquad \textbf{0-33}$$

$$\ddot{\varphi}(0) = 0 \quad \dot{\varphi}(0) = 0 \quad \varphi(0) = \varphi_0\,, \qquad \textbf{0-34}$$

$$\ddot{\mathrm{x}}(0) = 0 \quad \dot{\mathrm{x}}(0) = 0 \quad \mathrm{x}(0) = 0\,. \qquad \textbf{0-35}$$

Scaling of the physics is accomplished by defining nondimensional variables,

$$\mathrm{L}^* = \frac{\mathrm{L}}{\mathrm{L_{nc}}}\,, \quad \mathrm{t}^* = \frac{\mathrm{t}}{\mathrm{t_{rise}}}\,, \quad \varphi^* = \frac{\varphi}{\varphi_\mathrm{c}}\,, \quad \ddot{\mathrm{x}}^* = \frac{\ddot{\mathrm{x}}}{\mathrm{a_{ref}}}\,. \qquad \textbf{0-36}$$

With these definitions the controlling equations become,

Constitutive model

$$\frac{\mathrm{F}(\varphi, \dot{\varphi})}{\mathrm{F_0}} = \mathrm{F}^*(\varphi^*, \dot{\varphi}^*) = \frac{1}{(1 - \varphi^*)^\mathrm{a}}\left\{1 + (\mathrm{b} + \mathrm{c}\varphi^*)\max[0, \ln(\pi_2\dot{\varphi}^*)]\right\}, \qquad \textbf{0-37}$$

Lumbar load

$$\mathrm{L}^* = \frac{\mathrm{L}}{\mathrm{L_{nc}}} = \pi_3[\pi_1\mathrm{F}^*(\varphi^*, \dot{\varphi}^*) - 1]\,, \qquad \textbf{0-38}$$

Compression dynamics

$$\pi_4\ddot{\varphi}^* = \pi_5\ddot{\mathrm{x}}^* - [\pi_1\mathrm{F}^*(\varphi^*, \dot{\varphi}^*) - 1]\,, \qquad \textbf{0-39}$$

Initial condition

$$\mathrm{L}^*(0) = 0\,, \qquad \textbf{0-40}$$

Initial conditions

$$\ddot{\varphi}^*(0) = 0\,, \quad \dot{\varphi}^*(0) = 0\,, \quad \varphi^*(0) = 1 - \pi_1^{1/\mathrm{a}}\,, \qquad \textbf{0-41}$$

Seat pan initial conditions

$$\ddot{x}^*(0) = 0\,, \quad \dot{x}^*(0) = 0\,, \quad \ddot{\mathrm{x}}^*(0) = 0\,. \qquad \textbf{0-42}$$

where the scaling groups are given by,

Initial static compression

$$\pi_1 = \frac{\mathrm{F_0}}{\mathrm{W_0}}\,, \qquad \textbf{0-43}$$

Constitutive model

$$\pi_2 = \frac{\mathrm{\varphi_c}}{\mathrm{t_{rise}}}\frac{1}{\dot{\mathrm{\varphi}}_\mathrm{c}}\,, \quad a, \quad b, \quad c, \tag*{\textbf{0-44}}$$

Lumbar load

$$\pi_3 = \frac{\mathrm{W_{UT}}}{\mathrm{W_{UB}}}\frac{\mathrm{W_0}}{\mathrm{L_{nc}}}\,, \tag*{\textbf{0-45}}$$

Compression dynamics

$$\pi_4 = \frac{\mathrm{W_{UB}}}{\mathrm{W_0}}\frac{\mathrm{H_{cush}\varphi_c}}{\mathrm{t_{rise}^2 g^0}}\,, \quad \pi_5 = \frac{\mathrm{W_{UB}}}{\mathrm{W_0}}\mathrm{G_{max}}\cos\theta\,. \tag*{\textbf{0-46}}$$

Formally, the response observed in a test is representative of the conceptual model if all the scaling groups are equivalent. Test results cannot be used directly to inform regulatory decisions if distortions exist; however, if distortions are not excessive, test results can be used as validation benchmarks for computational simulations that can subsequently be used to inform regulatory decisions. Consequently, the scaling groups can play a significant role in test design, and they can be used to assess the adequacy of existing tests as validation benchmarks. Lastly, the scaling groups can be used to assess the degree of interpolation or extrapolation of a new design relative to a baseline design and the validation database.

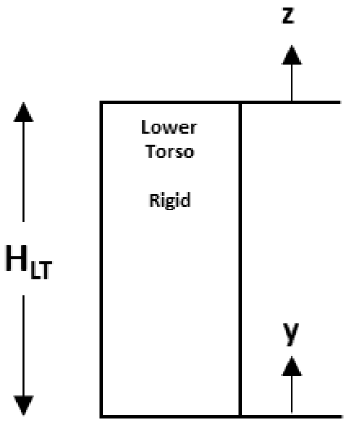


Figure 4.9: Conceptual model of the ATD lower torso (LT)

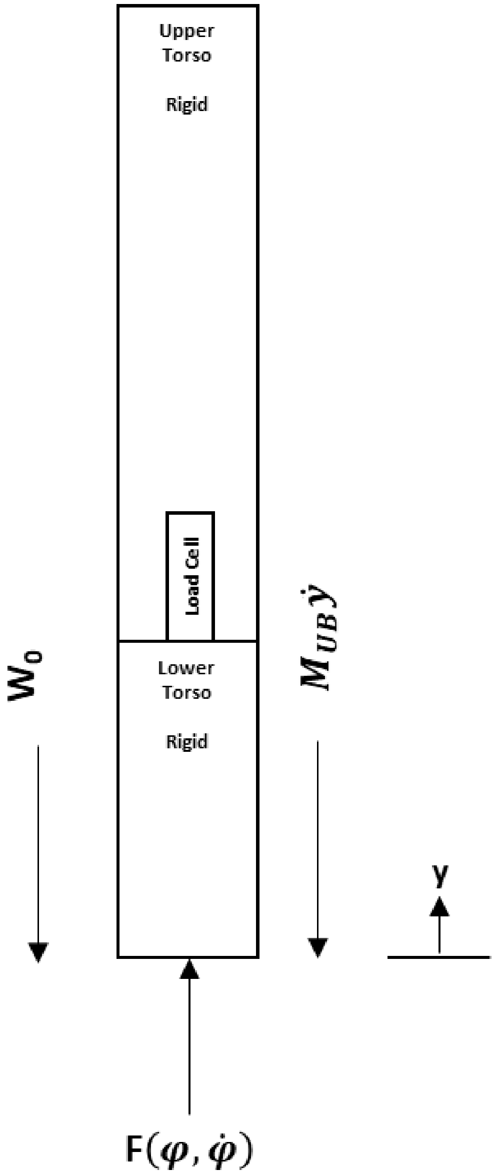


Figure 4.10: Conceptual model of the ATD upper body (UB)

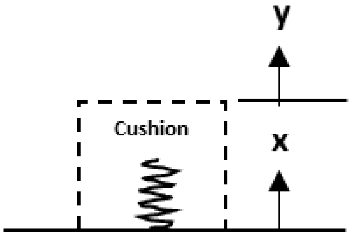


Figure 4.11: Conceptual model of the cushion

4.2.4 Assessment of conceptual models

There are three conceptual models. The first conceptual model is the sled test of the baseline design. In practice, the test environment (i.e., the acceleration pulse characterized by $\mathrm{G}^*_{\mathrm{max}}$ and $\mathrm{t}^*_{\mathrm{rise}}$) rarely reproduce precisely the target environment; consequently, a test is almost always a nearby representation of the target baseline design. It is standard practice to normalize the observed lumbar load to the target $\mathrm{G}_{\mathrm{max|target}}$,

$$\mathrm{L} = \mathrm{L}_{\mathrm{obs}} \frac{\mathrm{G}_{\mathrm{target}}}{\mathrm{G}_{\mathrm{obs}}}, \qquad \textbf{0-47}$$

when there is a discrepancy in G between the test and the target. This may be the primary effect, but a discrepancy in $\mathrm{t}_{\mathrm{rise}}$ may also impact the lumbar load. The practice is to report the normalized lumbar load and the target environment, $\mathrm{G}_{\mathrm{max}}$ and $\mathrm{t}_{\mathrm{rise}}$. Any residual discrepancy is absorbed into the characterization of test precision errors (see Section 0). Excluding physics, Table 4.6 summarizes the conceptual model for the sled test.

The measured lumbar loads in the sled test are used as a benchmark to validate model predictions. The second conceptual model is the simulation model for the sled test. The conceptual model for validation predictions is the same as the physical test. This is not always the case. Sometimes, the test conceptual model is distorted from the target for practical reasons, in which case, the validation conceptual model must reflect the validation test and not the target application. For example, tests can only be conducted at shoe box scale instead of full scale for economic reasons. Formal physical scaling (Table 4.7) becomes important when the test conceptual model is distorted for any reason.

The conceptual model for the nearby design is the third conceptual model. It differs from the conceptual models (target, test, and validation) for the baseline design only by the environment or design perturbations defining the new untested design. As indicated in Table 4.6, the nearby design differs from the baseline design in that the cushion thickness has been increased from 2.0” to 2.5”.

Table 4.6: Empirical comparison of conceptual models

	Baseline Design		Nearby Design
	Reality of Interest	Test: A15008 and Val Sim	Simulation
Environment	14CFR Part 25.562		
G	14	14	14
t_{rise} (ms)	80	80	80
Impact Angle, q (deg)	30	30	30
ATD	FAA-Hybrid III		
Total weight (lb)	170	170	170
UT weight (lb)	82.1	82.1	82.1
UB weight (lb)	102.7	102.7	102.7
$W_{LapBelt}$ (lbf)	4.81	4.8	4.8
Seating position	Upright	Upright	Upright
Seat Design			
Foam	CF42 (AC)	CF42 (AC)	CF42 (AC)
F_0 (lbf)	15.241	15.241	15.241
Comp at lockup, f_c	0.892	0.892	0.892
Shape param, a	1.321	1.321	1.321
Rate param-1, b	0.872	0.872	0.872
Rate param-2, c	1.642	1.642	1.642
Critical rate, $\dot{\varphi}_c$ (1/s)	6.968E-03	6.968E-03	6.968E-03
Thickness, H_{cush} (in)	2.0	2.0	2.5

Table 4.7: Formal comparison of conceptual models based on physics scaling

		Baseline Design		Nearby Design
	Scaling Group	Reality of Interest	Test: A15008 And Val Sim	Simulation
Initial static compression	p_1	0.142	0.142	0.142
Constitutive model	p_2	1.601E+03	1.601E+03	1.601E+03
	A	1.321	1.321	1.321
	B	0.872	0.872	0.872
	C	1.642	1.642	1.642
Lumbar load	p_3	0.086	0.086	0.086
Compression dynamics	p_4	0.690	0.690	0.863
	p_5	11.582	11.582	11.582

4.3 Assess simulation solution errors

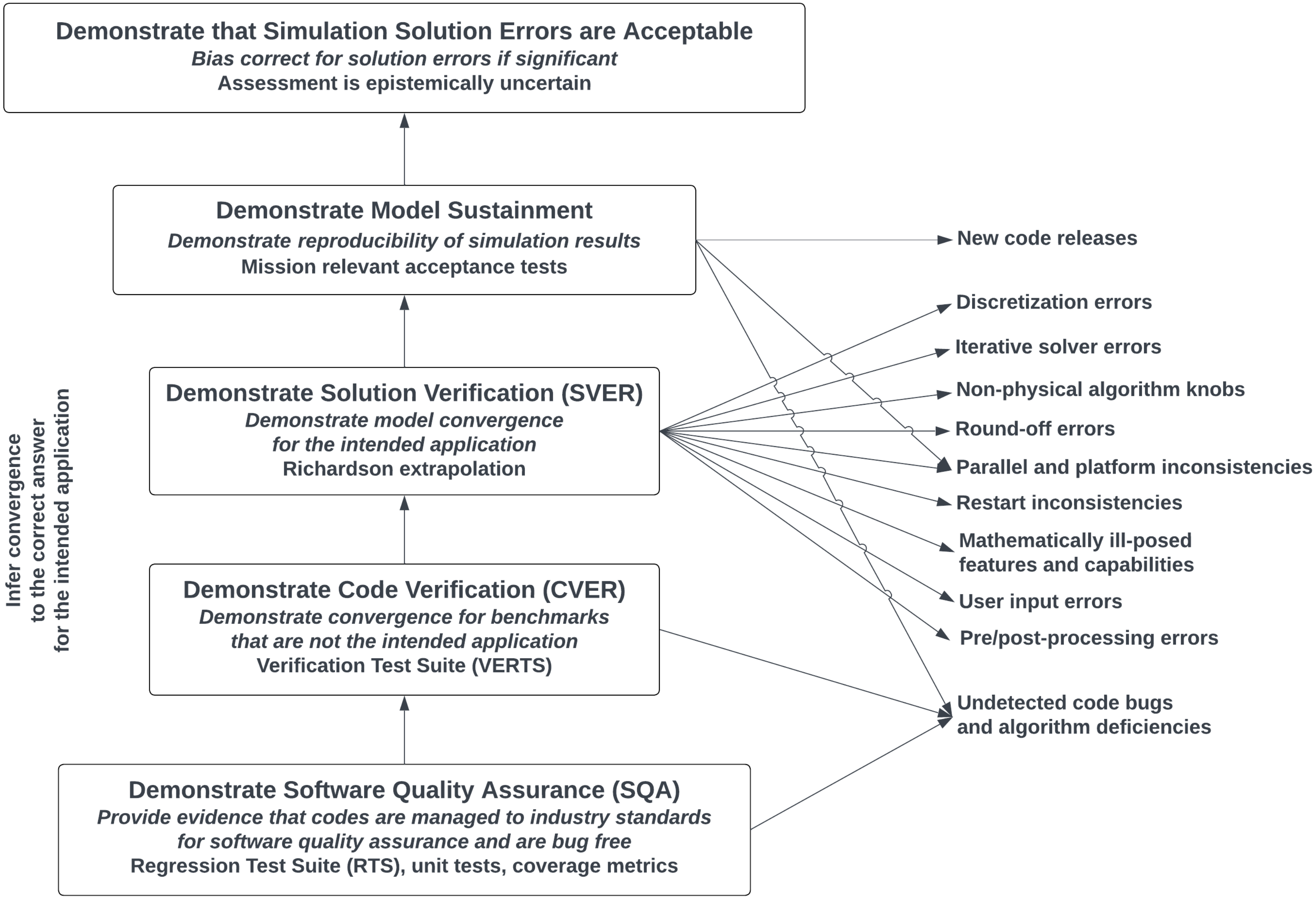
This product of this element is a computational model and the identification, quantification, and management of simulation solution errors. The goal is to demonstrate that solution errors are understood and acceptable in application simulations. Simulation results should be bias-corrected for known sources of solution errors unless negligible. The assessment of simulation solution errors is epistemically uncertain.

One expectation is that codes are managed to accepted software quality assurance standards and are bug-free. Software quality assurance (SQA) practices are the foundation for code development and managing simulation errors. Users can derive confidence that code capabilities are correctly implemented and robust when code development follows accepted software quality assurance (SQA) standards and when code testing is adequate.

A second expectation is that simulation results will converge to the correct answer for the intended application. There is no need for computer simulation if we know the correct answer for the intended application; consequently, satisfaction of this goal can only be *inferred* from evidence of code verification (CVER) and solution verification (SVER). Code verification addresses convergence to the correct answer for benchmarks that are not the intended application. Solution verification addresses convergence for the intended application, but we do not know if the answer is correct.

A third expectation is that current simulation results will be reproducible in the future. If not, which solution is correct? This is referred to as model sustainment.

Eleven sources of simulation solution errors are shown on the right side of Figure 4.12 and are discussed in greater detail in Section 0. The process for managing simulation solution errors is shown on the left side of Figure 4.12 and is discussed in greater detail in Section 0. Section 0 summarizes the assessment of simulation solution errors for the demonstration problem.



Demonstrate that Simulation Solution Errors are Acceptable
Bias correct for solution errors if significant
Assessment is epistemically uncertain
Demonstrate Model Sustainment
Demonstrate reproducibility of simulation results
Mission relevant acceptance tests
Demonstrate Solution Verification (SVER)
Demonstrate model convergence for the intended application
Richardson extrapolation
Demonstrate Code Verification (CVER)
Demonstrate convergence for benchmarks that are not the intended application
Verification Test Suite (VERTS)
Demonstrate Software Quality Assurance (SQA)
Provide evidence that codes are managed to industry standards for software quality assurance and are bug free
Regression Test Suite (RTS), unit tests, coverage metrics
Infer convergence to the correct answer for the intended application
New code releases
Discretization errors
Iterative solver errors
Non-physical algorithm knobs
Round-off errors
Parallel and platform inconsistencies
Restart inconsistencies
Mathematically ill-posed features and capabilities
User input errors
Pre/post-processing errors
Undetected code bugs and algorithm deficiencies

Figure 4.12: Process for the management of simulation solution errors

4.3.1 Develop computational model

In Section 0, the simulation conceptual model was defined as an application-specific abstraction of the reality of interest, specified with sufficient detail to allow an unambiguous solution. The computational model is the framework of hardware and software for solving the conceptual model. The computational model comprises:

1. The computers and operating systems that the codes will run on,
2. The codes used to perform pre-processing, solve the equations, and post-process the results,
3. Algorithms and data structures that form the basis of numerical solutions, and
4. Selection of discretization (e.g., grid), solver parameters, algorithm knobs, and other parameters necessary to solve the equations,

A commercial structural dynamics code, e.g., LSDYNA, will, in general, be used to solve the conceptual model; however, MS Excel will be used in this demonstration as a practical matter.

For the purposes of demonstrating the proposed BEPU process, the conceptual model comprises ordinary differential equations that require numerical solutions. A seven-step explicit time marching algorithm is employed with an expected formal order of convergence, p=1.0.

Step 1: Compute initial conditions

$$t = 0, \;\; L(0) = 0, \quad \ddot{x}(0) = 0, \quad \ddot{\varphi}(0) = \dot{\varphi}(0) = 0, \quad \varphi(0) = \varphi_c \left[1 - \left(\frac{F_0}{W_0} \right)^{1/a} \right] \tag{0-48}$$

Step 2: Take a time step

$$t^{N+1} = t^N + \Delta t = t^N + \frac{t_{rise}}{N_{step}}, \tag{0-49}$$

where the time step size is given by

$$\Delta t = \frac{t_{rise}}{N_{step}}, \tag{0-50}$$

and where N_{step} is a user supplied number of time steps to reach t_{rise}.

Step 3: Update the seat pan boundary condition, which is analytic

$$\begin{array}{cc} \text{for } t^{N+1} \le t_{rise} & \text{for } t^{N+1} > t_{rise} \\ \ddot{x}^{N+1} = G_{max} g^0 \cos\theta \dfrac{t^{N+1}}{t_{rise}} & \ddot{x}^{N+1} = G_{max} g^0 \cos\theta \left(2 - \dfrac{t^{N+1}}{t_{rise}} \right) \end{array} \tag{0-51}$$

Step 4: Update the foam compressive state

$$\dot{\varphi}^{N+1} = \dot{\varphi}^N + \ddot{\varphi}^N \Delta t\,, \tag{0-52}$$

$$\varphi^{N+1} = \varphi^N + \dot{\varphi}^N \Delta t\,. \tag{0-53}$$

Step 5: Update the lumbar load

$$L^{N+1} = \frac{W_{UT}}{W_{UB}} [F(\varphi^{N+1}, \dot{\varphi}^{N+1}) - W_0] \, . \qquad \textbf{0-54}$$

Step 6: Update the compressive acceleration

$$\ddot{\varphi}^{N+1} = \frac{\ddot{x}^{N+1} - g^0 \frac{[F(\varphi^{N+1}, \dot{\varphi}^{N+1}) - W_0]}{W_{UB}}}{H_{cush}} \, . \qquad \textbf{0-55}$$

Step 7: Step through time repeating Steps 2 through 6 until a user-specified time after t_{rise}

$$\frac{t^{N+1}}{t_{rise}} > \text{user specified value} \, . \qquad \textbf{0-56}$$

The computational model for the demonstration is summarized here:

1. The computers and operating systems that the codes will run on (3/19/24):

Platform	Dell XPS 15 9500 Intel® Core™ i9-10885H CPU @ 2.4GHz, 8 Core(s), 16 Logical Processors, 64GB RAM
Operating System	Windows 11 Pro Version 23H2 (OS Build 22631.3296)

2. The codes used to perform pre-processing, solve the equations, and post-process the results (3/19/24):
   a. There are no pre-processing codes required;
   b. Deterministic simulations and post-processing (including charting) are performed with MS Excel; and
   c. Monte Carlo simulations will be performed with @Risk, which is a commercial risk and decision analysis platform that integrates with MS Excel:

MS Excel	Microsoft® Excel® for Microsoft 365 MSO (Version 2402 Build 16.0.17328.20124) 64-bit
@Risk	@RISK 8.5.2. (Build 16)

3. Algorithms and data structures that form the basis of numerical solution:
   a. The solution algorithm is defined by the seven-step process above and Equations 0-48 to 0-56; and
   b. Environment and design parameters are implemented in tables suitable for lookup.

4. Parameters necessary to solve the equations, e.g., of discretization (e.g., grid), solver parameters, algorithm knobs, and other parameters necessary to solve the equations:

c. The user inputs the number of time steps to t_{rise} and the termination criteria, which are the only parameters necessary to solve the equations.

4.3.2 Sources of simulation solution error

The right side of Figure 4.12 lists eleven common sources of simulation solution errors, which are discussed in greater detail below.

Undetected code bugs and algorithm deficiencies

To rephrase George Box's famous quote, "All codes have bugs, some are useful." A broad user community routinely using a code for an intended class of applications is evidence that a code is useful. As evidence that all codes have bugs, commercial codes have an issue tracking system where users can submit problems, and the code development team can resolve them. Problems can be in the form of code bugs, i.e., an error in coding; but they can also be in the form of algorithm deficiencies. An algorithm deficiency is an algorithm that is coded correctly but gives an incorrect answer under some or all conditions; for example, the linearization of highly non-linear boundary conditions (e.g., radiation heat transfer). Consequently, simulation results are in a potential state of flux as bugs and algorithm deficiencies are detected and resolved, and new bugs and algorithm deficiencies are introduced with the addition of new capabilities to the code.

Code bugs and algorithm deficiencies will not impact every application, but my empirical experience[12] with a wide range of codes and applications is they are more common than you think. Often, they are masked by the complexity of the model and their solutions, by calibration activities to empirically minimize discrepancies between predictions and data, a general perception that discrepancies can be attributed to other recognized sources of error or uncertainty, and a widespread belief that codes are adequately tested. The codes were mature, had an extensive regression test suite, and were accepted by an experienced user base in *all* cases where issues were discovered. In *every* case, convergence studies made evident the code bug or algorithm deficiency, but push-back to perform order of accuracy studies is common. I was once told by a young analyst, "I'm doing engineering, not a science project," implying that their judgment reigned supreme. I often heard the phrase, "good enough for government work," after a bug or algorithm deficiency was discovered.

Pre/post-processing errors

Pre/post-processing errors can occur when information needs to be processed before it is useful for code input or to compute relevant QOI from code outputs. As an example of pre-processing, interpolation might be required to map measured boundary conditions in a test onto the computational model. As examples of post-processing, the head injury criterion (HIC, used in crash) and shock response spectrum (SRS, used in structural dynamics) are QOIs that are

[12] I speak generally and without attribution because the applications are either sensitive or the discovery is considered proprietary or embarrassing to the code development team. When V&V program manager, I took the perspective that *every* discovery of a code bug or algorithm deficiency was cause for celebration! Otherwise, they had the potential to distort simulation results in unknown ways.

functionals of the computed acceleration history. Pre/post-processing is typically performed with special programming, sometimes using tools such as Excel or MATLAB, which typically escape the same level of scrutiny and testing that is expected with commercial finite elements codes. Anecdotal stories abound for these types of errors.

User input errors

A user input error is the incorrect implementation of some aspects of the simulation conceptual model. An input error can be as simple as mistyping a material property or assigning the wrong material property to a component.

User-supplied subroutines are another source of user errors. I know of an example where a senior analyst implemented a constitutive model as a user-supplied subroutine. The subroutine was shared with other analysts in the organization. An error in the original implementation distorted many application analyses for many years before discovery.

Some available code features, capabilities, and constitutive models may have many complex options that require both specialized subject matter expertise and experience to make a proper selection for a given application. A novice analyst is more likely to make a bad choice of input options.

Feature abuse is the most serious form of user input error. Feature abuse occurs when an inappropriate feature, capability, or constitutive model is selected or used well outside the range of intended use.

Mathematically ill-posed features

Codes can offer features that are mathematically ill-posed, i.e., results are non-convergent with grid refinement. Examples include:

1. Element death to propagate failure,
2. Contact algorithms for complex interface geometries,
3. “Spot welds” or point contact features to represent fasteners,
4. Switch functions that change the nature of the response if some threshold is exceeded, e.g., failure or discontinuous flow regime maps and closure laws,
5. Grid-dependent sub-grid models, e.g., some early turbulence models or self-shielding in neutron transport.

Restart inconsistencies

Restart is a code feature that is sometimes used (if available) for large and long runtime models. The idea is to periodically stop and save the “state” of the calculation before continuing. If a problem occurs at some subsequent time, it can be addressed, and the calculation “restarted” from the “state” last saved without recalculating the entire transient. With restart, large calculations can be nursed along to completion. My observation with this capability is that you can get different results with different restart histories. This is referred to as restart inconsistency and is code- and application-specific. Restart introduces some small perturbation into the subsequent calculations, presumably because of an unknown but imperfect saving of the system state. The magnitude of restart inconsistency is potentially large if the solution is sensitive to the butterfly effect[13] or if one solution steps over a threshold event (e.g., failure) when the other does not.

Parallel and platform inconsistencies

Simulation results may not be reproducible when the same model is run with a different number of processors on the same computer platform and on a different platform that is currently available. This is referred to as parallel and platform inconsistency. This is a computer science issue that is outside the control of the analyst. Some codes and disciplines are more sensitive to parallel and platform inconsistencies than others, and the effect's magnitude can be application-dependent. Some applications using explicit dynamics or shock physics codes can be sensitive to this problem. The magnitude of parallel and platform inconsistency can be large if one of two solutions steps over a threshold event when the other does not. Current simulation results may not be exactly reproducible when new computers become available.

Roundoff errors

Roundoff error is the consequence of finite precision arithmetic compared to exact arithmetic. It can be a problem for ill-conditioned systems when two nearly identical numbers are differenced or when small roundoff errors are accumulated over many time steps or iterations.

Non-physical algorithm knobs

Codes might offer ad hoc (non-physical) algorithm knobs that make solutions more practical but at the expense of introducing error. Examples include:

1. Hourglass parameter – adds artificial stiffness to hex elements so they do not invert,
2. Mass scaling – add mass to small mass elements, enabling larger time steps in explicit codes,
3. Artificial viscosity – adds dissipation to suppress unphysical oscillations and enhance numerical stability,
4. Remeshing – when mesh quality deteriorates significantly.

Simulation results should be bias-corrected for errors introduced by non-physical algorithm knobs.

Iterative solver errors

When the governing physics equations are discretized and solved implicitly, a large system of coupled linear (or linearized) algebraic equations results. Iterative methods are commonly employed to solve this exceptionally large system of linearized algebraic equations. Iterative error is the difference between the current approximate solution and the exact solution to the system of algebraic equations (not to be confused with the solution to the physics equations).

[13] The butterfly effect is the sensitive dependence on initial conditions in which a small change in one state of a deterministic nonlinear system can result in large differences in a later state. The metaphorical example is a butterfly flapping its wings, thus initiating a significant change in the weather.

Discretization errors

Discretization error occurs when continuous functions and their derivatives are represented only at discrete points in space, time, angle, energy, or any other independent variable. The spacing of discrete points is termed discretization. This results in a large system of loosely coupled linear algebraic equations that can be solved explicitly in time or implicitly, requiring the iterative solution of a large system of coupled linear (or linearized) algebraic equations. Solutions to the governing physics equations are expected to converge to a discretization-insensitive solution when the discretization is sufficiently small; however, the expected convergence will not be realized if mathematically ill-posed features are used in the computational model.

Different code releases

Codes are not stationary in time; different code releases can give different results when running the same computational model. Input formats may be refactored. Commonly, old input files may not run on newer code versions after some period. Algorithms may change. New bugs can be introduced with new capabilities that can also impact old capabilities. Old bugs or algorithm deficiencies can be resolved in new code releases[14]. Consequently, current simulation results may not exactly reproduce prior results produced with an earlier code release.

4.3.3 Processes for the management of simulation solution errors

The left side of Figure 4.12 shows the four main processes for identifying and managing simulation solution errors. Each of the four processes is described in more detail below, with a discussion of best practices. Sometimes, best practices lead to quantifying error and uncertainty in its assessment. In other cases, the errors and uncertainties are intangible (i.e., not quantifiable). Still, the exercise of best practices lends credibility to simulation results by acknowledging potential errors and showing due diligence in their management.

Demonstrate software quality assurance (SQA)

Software quality assurance (SQA) practices are the foundation for code development and managing simulation errors. When code development follows accepted SQA practices, users expect and can derive some level of confidence that code capabilities are correctly implemented and robust.

In general, users will not have the expertise or access to supporting evidence to assess SQA for codes. Ideally, users should rely on independent certification to some accepted SQA standard such as NQA-1, IEEE, or ISO-9000. Independent certification is more likely to exist for codes

[14] I have over 25 years of experience with the @Risk software. During that period, I discovered and reported 3 code bugs that were acknowledged and subsequently resolved by the code developers. Each bug correction was implemented in a new code release.

developed by government agencies[15] because commercial codes consider their SQA practices and testing strategy to be proprietary.

In the absence of independent certification, users should seek answers to the following questions:

1. Is a user manual and training available with sample problems relevant to the current application?
2. Is there a theory manual describing methods, algorithms, and technical basis for capabilities in the code?
3. Is technical support available, and does the code maintain an issue-tracking system?
4. Are the features tested that are used in my application?

Commercial codes generally have a good user manual and training that helps minimize user input errors. Sample inputs for application-relevant problems can benefit novice users but may not be available for specific applications. Some available code features and constitutive models may have many complex options that require specialized subject matter expertise and experience for proper selection. A theory manual describing the technical basis for algorithms and methods will help minimize feature abuse and, more importantly, help the user understand if the code has the features and capabilities needed for a given application. A theory manual is essential for ePIRT development. A theory manual is more likely to exist for codes developed by government agencies because commercial codes consider their algorithms and methods proprietary.

Issue tracking is a key practice in all accepted SQA standards. It allows users to submit questions and evidence of potential bugs to the code development team. The team tracks user submittals and addresses them on a priority basis. Issues submitted to the code team often result from user errors or requests for new capabilities. Code corrections, if necessary, appear in new code releases. Commercial codes have technical support available to the user; if required, the need for bug fixes will be addressed.

Unit and regression testing is critical and necessary in code development activities. Unit tests are usually simple tests internal to the code and inaccessible through code inputs. Regression tests are simple tests that are accessed through code inputs. Regression tests are fast-running, and the regression test suite (RTS) is typically run nightly to ensure the stability and correctness of the code during development and with each code release.

Unit and regression testing have limitations worth noting. Unit and regression testing requires a substantial inferential leap to say that simulations will converge to a correct answer for an intended application because unit and regression testing typically (1) do not quantify convergence behavior, (2) do not systematically address interactions of features and capabilities within the code, (3) do not detect deficient algorithms, which can be coded correctly, but which can produce wrong answers, and (4) do not detect mathematically ill-posed features that lead to non-

[15] Codes developed by the DOE as part of the CASL Program (Consortium for Advanced Simulation of Light Water Reactors) are NQA-1 certified. I helped develop the code verification requirements and test coverage strategies for these codes.

convergent solutions. Unit and regression testing primarily serve code developers. The user community cannot access unit tests or the regression test suite; consequently, users cannot assess test coverage or test quality for the features used in a specific application.

Completeness is one question that comes to mind when considering what is necessary and sufficient in code testing. It is common to hear that a given code has thousands or even tens of thousands of regression tests. However impressive, such numbers leave open the question as to whether this is insufficient (needing even more resources) or overkill (a waste of resources), or thoughtfully “just right.” So, how much testing is enough?

Code testing is enough from a code development perspective, (1) when there are not many bugs that escape into the user community, and (2) when there is no fear of needing catastrophic changes to the code (i.e., small changes should not propagate into major unexpected faults throughout the code). Code testing is enough from a user perspective when simulation results are stable with new code releases and the features and capabilities used in a specific application are tested and work.

There is no way to automate the reporting of unit test coverage, so even the code developer may not know. Tools are commonly available to code development teams to assess line, function, or path coverage. Coverage metrics of this sort are always evaluated against the regression test suite.

Test coverage metrics potentially serve two purposes: (1) to identify what in the code has not been tested and (2) to communicate a measure of completeness and quality. Both are elusive to the user and regulatory communities.

A clear understanding of gaps can focus priorities for new testing activities. This is commonly thought to be the responsibility of code developers, but users can and should play a key role in setting priorities for testing. This requires that gaps in testing be related to applications and exposed to the user community in a way that they can appreciate in the context of their application. This is not the case with commercial codes. In general, there is no way for the user community to assess the completeness or quality of code testing as it relates to their specific applications. Uncertainty regarding the potential for undetected code bugs or algorithm deficiencies should be identified as intangibles.

There are two notable exceptions regarding the ability of the user community to assess the completeness or quality of code testing as it relates to their specific applications. First, (Porter et al., 2020b) assessed unit test coverage of constitutive models available for analyzing pressurized water reactors (PWRs) using the CTF code. Unit tests are embedded in the code and not available for assessment by the user or regulatory communities. Table 4.8 shows that most constitutive models were already tested (81% coverage). Still, there were two notable gaps, the Dittus–Boelter equation for wall heat transfer and the Chen equation for subcooled nucleate boiling, so new tests were developed (increasing coverage to 88%). The Dittus–Boelter equation is used in all PWR analyses using the CTF code, so the gap was significant. Depending on what constitutive models are used in a specific application, the user (and the regulatory) agency can easily assess if there are any residual gaps in constitutive model testing and compute a meaningful application-specific coverage metric. The user can then ask that new tests be developed to fill gaps in test coverage. In this way, users can share the responsibility of identifying the need for new unit tests.

The feature coverage tool (FCT) (Sandia, 2014) , developed for the Sierra suite of codes as part of the ASC Program at Sandia National Laboratories, is the second notable exception. The FCT is application-specific. The FCT looks at every command line in an application model and looks for an identical command line in the RTS[16]. When a match is found, the command line (i.e., feature) is flagged as being tested with a regression test. Figure 4.13 shows a snippet of a two-way coverage report generated by the FCT for a Sierra input file. The left column lists “features” in command line format. The matrix to the right shows which features have been tested by the RTS. Blue shading on the diagonal means that this feature is at least tested on its own. Red means the feature has not been tested. Off-diagonal elements represent the two-way interaction between features and whether they have been tested. I have seen examples where features were successfully tested individually but failed when their interactions were tested.

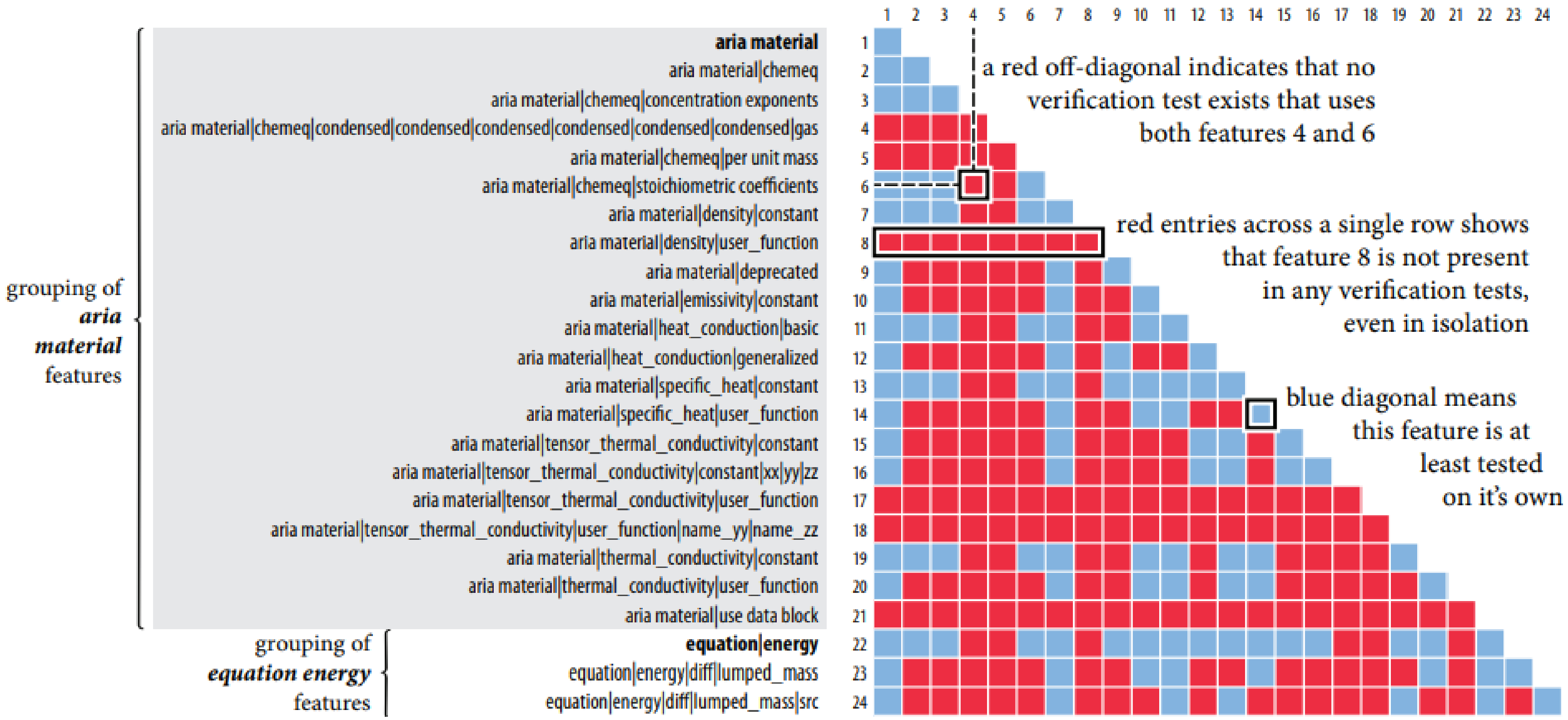
In practice, a user would run the FCT when a large application model is first developed and prioritize the gaps for the code development team to address. Meaningful and application-specific coverage metrics are readily calculated from the matrix. There are two drawbacks to the FCT. First, the quality of the tests is not easily assessed. Second, FCT is command-line-centric and not physics-centric[17], and the features (command lines) are only meaningful to the users and code developers. In addition to the RTS, the FCT can be run independently against the verification test suite (VERTS, order of accuracy tests) if one exists.

[16] All the options must match if a command line has multiple options.

[17] Regulatory agencies are more likely to understand physics than command lines.

Table 4.8: Unit testing of constitutive models in the CTF code (Porter et al., 2020b)

		Unit Tests	
	Model	Existing	Updated
Energy Conservation in Solids			
Fuel Material Properties	rUO2: constant	X	X
	kUO2: MATPRO-11	X	X
	kUO2: Modified NFI	X	X
	kUO2: Halden	X	X
	kUO2: Density correction	X	X
	Cp,UO2: MATPRO-11	X	X
Zircaloy Material Properties	rzirc: constant	X	X
	kzirc: MATPRO-11	X	X
	Cp,zirc: MATPRO-11	X	X
Dynamic Gap	Hgas	X	X
	Hconstant	X	X
	Hrad	X	X
Single Phase Hydraulics			
Equation of State	h: IAPWS	X	X
	k: IAPWS	X	X
	Cp: IAPWS	X	X
	m: IAPWS	X	X
	s: IAPWS	X	X
	Tsat: IAPWS	X	X
Axial/Lateral Wall Friction	f: CTF	X	X
	f: McAdams	X	X
	f: Zigrang–Sylvester	X	X
	f: Churchill	X	X
	f: User-defined	X	X
Turbulent Mixing	b: Rogers and Rosehart		
	b: Beus		
Grid TKE Enhancement	K/K0: Yao, Hochreiter, Leech		
Coolant Energy Conservation			
Wall Heat Transfer	Dittus–Boelter		X
Subcooled Nucleate Boiling	Thom	X	X
	Chen		X
Near Wall Condensation	Ahmed	X	X
	Hancox–Nicoll		
Grid HT Enhancement	Nu/Nu0: Yao, Hochreiter, Leech	X	X



a red off-diagonal indicates that no verification test exists that uses both features 4 and 6

red entries across a single row shows that feature 8 is not present in any verification tests, even in isolation

blue diagonal means this feature is at least tested on it's own

Figure 4.13: Example of two-way coverage report generated with the FCT (Sandia, 2014)

More generally, test coverage metrics (line, path, function), even if reported to the user community, might not communicate a meaningful measure of completeness and quality. High coverage metrics are considered particularly good and imply high quality in code testing. Low coverage metrics might appear bad, indicating a lack of completeness or low quality. However, such metrics can be deceiving because there is always code that is never referenced in a specific application model. This untouched code might exist because the models and options are irrelevant to the current application, because development coding was created and abandoned, or because code was created for debugging purposes and is no longer used. Line and related coverage metrics say nothing about the quality, effectiveness, or application relevance of the tests, and they say nothing about missing code, missing error handling, or missing requirements. In addition, unit and regression testing say little or nothing about integration (i.e., interaction of critical features and capabilities) or the adequacy of properly coded algorithms. Furthermore, unit and regression tests may not provide uniform coverage across all code elements. Consequently, some applications may be well tested and others poorly tested. A critical new capability, or its interactions with existing capabilities, may be poorly tested, and the user would never know. It is easy to generate meaningless “tests,” even touching the whole code, if pursuing code coverage metrics takes precedence over trying to find bugs!

Acknowledging these qualifiers can diminish the value of these coverage metrics as measures of completeness and quality for the user community if they were available. This is less of a problem for the code developers because they can provide context and priority to coverage gaps. Many organizations would say that less than 50% line coverage is clearly a red flag indicating inadequate testing, that 80% is about the sweet spot, and that achieving 100% may not be cost-effective or even desirable.

Table 4.9 summarizes the evidence of SQA for the codes used in the demonstration. For the user community, MS Excel and @Risk are black boxes with unknown SQA and code testing. Accepting these codes for regulatory purposes should not be an act of faith.

Table 4.9: Evidence of software quality assurance (SQA) for codes used in demonstration

	MS Excel	@Risk
Independent SQA certification?	Unk	No
User manual and training?	Yes	Yes
Theory manual describing algorithms and methods?	No	No
Technical support and issue tracking system?	Yes	Yes
Testing: line, path, or function coverage > 80%?	Unk	Unk

The best practice is to develop a suite of application-relevant acceptance tests for key application classes. The acceptance tests can be regression, verification, or sustainability tests, as defined below. Table 4.10 shows the matrix of acceptance tests developed for the demonstration. Appendix D describes the tests in detail and their acceptance criteria.

Table 4.10: Summary of acceptance tests for demonstration

			Requirements Tests	Foam Model	Static Compression	Boundary Condition	Compression Acceleration	Lumbar Load
1	Regression Test Suite		100% Coverage					
	Test	1.1	Foam constitutive model	X				
	Test	1.2	Decompression	X				
	Test	1.3	Initial static compression		X			
	Test	1.4	Seat pan $\ddot{x}$ 0<t*<1			X		
	Test	1.5	Seat pan $\ddot{x}$ 1<t*<2			X		
	Test	1.6	Cushion $\ddot{\varphi}$				X	
	Test	1.7	Lumbar load no cushion					X
2	Verification Test Suite		100% Coverage					
	Test	2.1	Linear spring/mass system	X	X	X	X	X
3	Sustainability Test Suite		100% Coverage					
	Test	3.1	Baseline seat design (test A15008)	X	X	X	X	X

Demonstrate Code Verification

The VVUQ10 standard (ASME, 2019) provides a formal definition of verification, which is the process of determining that a computational model accurately represents the underlying mathematical model and its solution. There is an expectation that simulation results will converge to the correct answer for the intended application. Satisfaction of this goal can only be *inferred* from evidence of code verification (CVER) and solution verification (SVER). Code verification addresses convergence to the correct answer for benchmarks that are not the intended application. Solution verification addresses convergence for the intended application, but we do not know if the answer is correct. CVER is addressed first.

(Oberkampf & Roy, 2010) state that "the order-of-accuracy test is the most difficult test to satisfy; therefore, it is the most rigorous of the code verification criteria. It is extremely sensitive to even small mistakes in the code and deficiencies in the numerical algorithm." The primary metric for CVER is the observed order of accuracy. Deficient algorithms can be coded correctly and still produce wrong answers!

(Roy, 2005) provides an excellent review of code verification procedures for computational simulation. The discrepancy between a computed solution (L) and a known benchmark solution (L_E) is given by a power law expansion in the asymptotic regime,

$$\mathrm{D} = \mathrm{L} - \mathrm{L_E} = \mathrm{g}\left(\frac{1}{\mathrm{N}}\right)^{\mathrm{p}} + \mathrm{HOT}, \qquad \textbf{0-57}$$

where N is the number of elements in the numerical solution, p is the order of accuracy, and g is a fitting parameter. An asymptotic regime means that higher-order terms (HOTs) are negligible and that other sources of numerical error (e.g., roundoff errors and iterative solver errors) are insignificant. The additional assumption is that the benchmark solution is smooth and does not exhibit discontinuities or singularities.

Two unknown parameters, g and p, can be determined if we have solutions on two grids. The first grid will be called the coarse grid, with N_C elements and solution L_C. The second grid is obtained by a single isotropic refinement and is referred to as the medium grid with N_M elements and solution L_M. The order of accuracy (p) is then given by,

$$\mathrm{p} = \frac{\ln \dfrac{\mathrm{L_C} - \mathrm{L_E}}{\mathrm{L_M} - \mathrm{L_E}}}{\ln \mathrm{r}}, \qquad \textbf{0-58}$$

where r is the refinement ratio,

$$r = \frac{N_M}{N_C}. \quad \text{0-59}$$

The numerical solution of the benchmark test is assumed error-free if

$$\left|E_p\right| = \left|\frac{\hat{p} - p_f}{p_f}\right| \leq 0.10\,, \quad \text{0-60}$$

where $\hat{p}$ is the observed order of accuracy, and p_f is the formal order of accuracy for the numerical scheme.

Only two grids are required when you are in the asymptotic regime, but three (or more) grids are required to provide *evidence* that you are in the asymptotic regime. The asymptotic regime is evidenced by stability in the observed order of accuracy with additional refinement.

(Roy, 2005) discusses two methods for obtaining benchmarks for code verification. The first is the method of exact solutions. Here, the benchmark is an exact solution to the governing equations with specified initial and boundary conditions. The main drawback to this method is that only a limited number of exact solutions are available when nonlinear physics or coupled multi-physics is involved.

The second method discussed by (Roy, 2005) for obtaining verification benchmarks is the method of manufactured solutions (MMS). The method of manufactured solutions, or MMS, is a general and powerful approach to code verification. Rather than trying to find an exact solution to a system of partial differential equations, the goal is to “manufacture” an exact solution to a slightly modified set of equations. The general concept behind MMS is to choose the solution a priori and then operate the governing partial differential equations onto the selected solution, thereby generating analytical source terms. The chosen (manufactured) solution is then the exact solution to the modified governing equations comprising the original equations plus the analytical source terms. Thus, MMS involves solving the backward problem: given an original set of equations and a chosen solution, find a modified set of equations that the chosen solution will satisfy. The initial and boundary conditions are then determined from the solution. An essential requirement is that codes have “hooks” for accepting the source terms generated by MMS.

A single verification test seldom tests all the features available in a code or needed in a specific application; consequently, a suite of verification tests is expected. The verification test suite (VERTS) should be documented, maintained under configuration control, and rerun on demand, e.g., before major code releases. Previous comments on test coverage are equally applicable to the VERTS.

Verification testing is normally thought to be the responsibility of code developers, but users can and should play a key role in setting priorities for verification testing. This requires that gaps in verification testing be related to applications and exposed to the user community in a way that they can appreciate in the context of their application. This is not the case with commercial codes. In general, there is no way for the user community to assess the completeness or quality of verification testing as it relates to their specific applications. The potential for undetected code bugs or algorithm deficiencies should be identified as intangibles.

There are notable exceptions regarding the user community's ability to assess the completeness or quality of verification testing related to their specific applications. Table 4.11 shows the VERTS for the CTF code when used for PWR applications. Note that the VERTS is organized by the controlling conservative equations and the corresponding physics (left column). Test names are shown vertically across the top. Some verification tests already existed at the time of the assessment, but there were many gaps (62% coverage). (Porter et al., 2020b) expanded the VERTS with test problems from the literature and newly developed test problems to fill the gaps. If successful, the coverage would have increased to 83%; however, two of the added tests failed! This is significant because CTF was mature code with extensive unit and regression testing and an experienced user base that has used the code for test design, interpretation of test results, and application predictions for many years.

(Toptan, Porter, Hales, Williamson, et al., 2020) provides a second example of a verification test matrix for the BISON code used for nuclear reactor fuel performance analysis. The code has three governing equations: transient heat conduction, transient species inventory, and stress.

Table 4.12 shows the verification matrix for the heat conduction equation. The matrix is organized by physics across the top, and tests are numbered in the left-hand column. This assessment did not discover any code bugs or algorithm deficiencies.

The VERTS for CTF and BISON have some things in common. Both are government-funded codes; their VERTSs are documented and available to the user and regulatory community. Both are organized around physics, so it is easy for the user and regulatory communities to assess the completeness and quality of verification testing for specific applications.

This is not the case with commercial codes, but there is one example where commercial codes were independently assessed. (Abanto et al., 2005) assessed, without attribution, three commercial CFD codes against test cases with known exact solutions as benchmarks. Non-monotonic grid convergence was observed in all test cases, indicating numerical and model errors. One code did particularly poorly, exhibiting non-convergent behavior in some cases. Anecdotally, the code developer cried “user error”, only to discover a code bug in implementing a boundary condition when he ran the test cases himself.

My personal experience at SANDIA is similar. On two occasions, I insisted on a verification test (order of convergence) for capabilities that passed regression testing only to find that the capability was non-convergent. Code bugs were subsequently found and corrected, resulting in the expected convergence behavior.

The messages are clear: (1) verification testing that uses the observed order of accuracy as a metric is extremely effective at finding undetected code bugs and algorithm deficiencies, and (2) do not assume codes are bug-free. The VERTS should be documented and organized around physics, allowing users and regulatory agencies to easily identify gaps and assess quality in verification coverage.

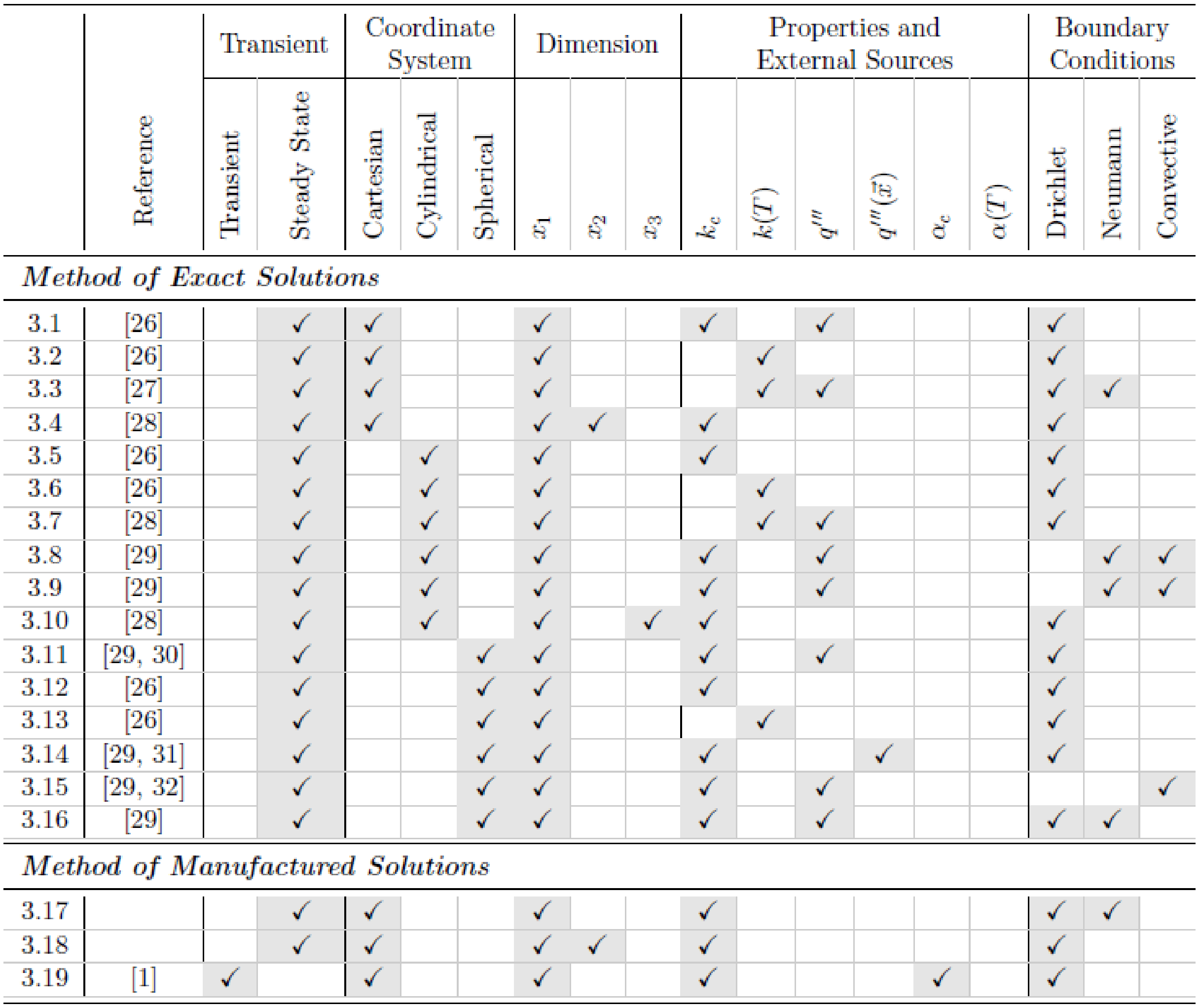
The best practice is to work with the code developers to identify application-relevant verification tests in code documentation, but if none are available, develop at least one application-relevant order of accuracy verification test. Ideally, all verification tests should be run on the applicant's computers and operating systems. A verification test was developed for this project (Appendix D) in the form of a linear spring-mass system subjected to the FAA-prescribed boundary conditions and is included in Table 4.10 as part of the acceptance test suite.

(AIAA, 2024) has taken the strongest stand on this issue of any organization: "The users of a CFD code should be prepared to conduct their own code verification for their specific application, or to at least audit, check, analyze, or reproduce some of the developer's verification results or to confirm the adequacy of coverage of these results for their intended applications."

Table 4.11: CTF VERTS for PWR applications (Porter et al., 2020b)

	Existing VERTS					From Literature			Newly Developed			
Physics	Heat Exchanger	Turbulent Mixing	Flow Split	Grid Enhancement	Grid Spacer	Isokinetic Advection	Linear Conduction	Water Faucet*	Friction and Gravity	Convection	Nonlinear conduction*	Pipe Boiling
Fluid Mass												
Transient						X		X				
Axial Advection						X		X				
Lateral Advection												
Mass Transfer												X
Turbulent Mixing												
Fluid Energy												
Transient						X				X		X
Axial Advection	X	X		X		X						X
Lateral Advection												
Interfacial transfer												X
Convection	X			X						X		
Grid Enhancement				X								
Turbulent Mixing		X										
Fluid Momentum												
Transient								X				
Axial Advection								X				
Gravity			X					X	X			
Axial Pressure			X		X				X			
Lateral Pressure			X									
Shear			X						X			
Grid Enhancement				X								
Form Loss					X							
Interfacial Shear												
Turbulent Mixing												
Solid Energy												
Transient										X		
Linear Conduction							X					
Nonlinear Conduction											X	
Energy Generation							X				X	
Convection										X		
Two-Phase								X				X
Equation of State					X	X		X				X

Table 4.12: Verification matrix for the BISON heat conduction equation



		Transient		Coordinate System			Dimension			Properties and External Sources						Boundary Conditions		
	Reference	Transient	Steady State	Cartesian	Cylindrical	Spherical	x_1	x_2	x_3	k_c	$k(T)$	q''	$q'''(\vec{x})$	α_c	$\alpha(T)$	Drichlet	Neumann	Convective
Method of Exact Solutions																		
3.1	[26]		✓	✓			✓			✓		✓				✓		
3.2	[26]		✓	✓			✓				✓					✓		
3.3	[27]		✓	✓			✓				✓	✓				✓	✓	
3.4	[28]		✓	✓			✓	✓		✓						✓		
3.5	[26]		✓		✓		✓			✓						✓		
3.6	[26]		✓		✓		✓				✓					✓		
3.7	[28]		✓		✓		✓				✓	✓				✓		
3.8	[29]		✓		✓		✓			✓		✓					✓	✓
3.9	[29]		✓		✓		✓			✓		✓					✓	✓
3.10	[28]		✓		✓		✓		✓	✓						✓		
3.11	[29, 30]		✓			✓	✓			✓		✓				✓		
3.12	[26]		✓			✓	✓			✓						✓		
3.13	[26]		✓			✓	✓				✓					✓		
3.14	[29, 31]		✓			✓	✓			✓			✓			✓		
3.15	[29, 32]		✓			✓	✓			✓		✓						✓
3.16	[29]		✓			✓	✓			✓		✓				✓	✓	
Method of Manufactured Solutions																		
3.17			✓	✓			✓			✓						✓	✓	
3.18			✓	✓			✓	✓		✓						✓		
3.19	[1]	✓		✓			✓			✓				✓		✓		

Demonstrate Solution Verification

There is an expectation that simulation results will converge to the correct answer for the intended application. Satisfaction of this goal can only be *inferred* from evidence of code verification (CVER) and solution verification (SVER). Code verification addresses convergence to the correct answer for benchmarks that are not the intended application. CVER was previously discussed. Solution verification addresses convergence for the intended application, but we do not know if the answer is correct. The primary goal of SVER is to quantify numerical errors for a given discretization. A secondary metric is the observed order of accuracy, which is sensitive to code bugs and algorithm deficiencies.

Quantification of numerical errors (SVER) is required for all applications of the computational model. As a minimum, SVER is required for simulation of the baseline seat design and the nearby design. SVER is also required for the simulation of every test in a validation suite. The exception is when numerical errors are shown to be negligible for the most numerically demanding of the environments or designs that need to be assessed. In this case, numerical errors can be assumed negligible for the less numerically challenging simulations.

The estimation of numerical errors is a function of both discretization and mesh quality. Poor mesh quality, as evidenced by highly skewed elements, can significantly impact numerical errors. (SAE, 2021) provides guidance on mesh quality which should be addressed in conjunction with attempts to quantify numerical errors.

(Roy, 2005) provides an excellent review of solution verification procedures for computational simulation. SVER addresses convergence for the intended application, so there is no benchmark (L_E) against which to compare the numerical solution (L). A benchmark solution can be replaced by an extrapolated solution (Le) in the power law expansion,

$$\mathrm{L} = \mathrm{L_e} + \mathrm{g}\left(\frac{1}{\mathrm{N}}\right)^{\mathrm{p}} + \mathrm{HOT}, \qquad \textbf{0-61}$$

where N is the number of elements in the numerical solution, p is the order of accuracy, and g is a fitting parameter. This is referred to as Richardson Extrapolation. An asymptotic regime means that higher-order terms (HOTs) are negligible and that other sources of numerical error are negligible. Sources of solution errors that can undermine the process will be addressed later in this section. The additional assumption is that the solution is smooth and does not exhibit discontinuities or singularities.

Equation 0-61 has three unknown parameters (L_e, g, p) that can be determined if we have three grids (N_C, N_M, N_F) with corresponding solutions (L_C, L_M, L_F), where $N_C < N_M < N_F$. The main constraint of refinement is that it is isotropic. Note, it is possible to solve for L_e and g on two girds if you make the risky (and not recommended) assumption that the order of accuracy (p) would be the same as the formal order of convergence (p_f) for the numerical scheme.

In general, the three constants can be determined by a least square fitting procedure: but if the refinement ratio,

$$\mathrm{r} = \frac{\mathrm{N_M}}{\mathrm{N_C}} = \frac{\mathrm{N_F}}{\mathrm{N_M}}, \tag{0-62}$$

is constant, an analytic solution is possible,

$$\mathrm{L_e} = \mathrm{L_F} + \frac{\mathrm{L_F} - \mathrm{L_M}}{\mathrm{r^p} - 1} \tag{0-63}$$

where

$$\mathrm{p} = \frac{\ln \frac{\mathrm{L_C} - \mathrm{L_M}}{\mathrm{L_M} - \mathrm{L_F}}}{\ln \mathrm{r}}. \tag{0-64}$$

Grid doubling, i.e., r = 2, is common, but this is not a requirement. The relative error in the solution L_N on any grid of size N is easily computed:

$$|\mathrm{E_N}| = \left|\frac{\mathrm{L_N} - \mathrm{L_e}}{\mathrm{L_e}}\right|. \tag{0-65}$$

Three grids are required when you are in the asymptotic regime, but four (or more) grids are required to provide *evidence* that you are in the asymptotic regime. The asymptotic regime is evidenced by stability in the observed order of convergence for additional refinement.

The error estimate, $|E_N|$, is epistemically uncertain. Different values will be realized with different starting grids, grid triplets, mesh quality, etc. (Roache, 2009) quantified a factor of safety (FS) based on many case studies where the benchmark was either an analytic solution or a high-quality hyper-refined numerical solution. Roache defined a grid convergence index (GCI) as

$$\mathrm{GCI} = \mathrm{FS}|\mathrm{E_N}| \tag{0-66}$$

such that GCI bounds the actual value of $|E_N|$ with 95% confidence.

Recommendations for the implementation of the GCI for solutions on three or more systematically refined grids are given by (Oberkampf & Roy, 2010) and shown in Table 4.13.

Table 4.13: Recommended implementation of the GCI

$\left\|E_p\right\| = \left\|\frac{\hat{p} - p_f}{p_f}\right\|$	**FS**	**p**
< 0.1	1.25	p_f
> 0.1	3.0	$\min(\max(0.5, \hat{p}), p_f)$

Other sources of solution error can impact estimates of numerical errors using Richardson extrapolation. These are addressed next. In many cases, they can only be managed through best practices and should be identified as intangibles.

Undetected code bugs and algorithm deficiencies

Undetected code bugs and algorithm deficiencies can impact the results of Richardson Extrapolation; consequently, SQA, unit and regression testing, and code verification (order of accuracy tests) are essential prerequisites. Additional confidence that the code is free of bugs and algorithm deficiencies is provided when the observed order of accuracy reasonably agrees with the formal order of accuracy. The factor of safety (FS) in Table 4.12 is larger when this is not the case. Note that the order of accuracy to use in the implementation is constrained to $0.5< p < p_f$. Outside this range, the simulation results are suspect, and the error bounds afforded by the GCI are unreliable.

The GCI was computed for both validation and certification simulations. GCI results are shown in Figure 4.14. Errors are about an order of magnitude larger for certification simulation compared to validation simulations. The horizontal line represents an error of 1 lb_f in 1500 lb_f. Errors smaller than this are considered negligible and can be ignored for seat certification applications. The order of accuracy was computed on the grid triplet, 2000-, 4000-, and 8000-time steps and $|E_p| < 0.1$ in both cases, lending additional evidence that no code bugs or algorithm deficiencies are polluting the solution. Validation and certification simulations will use 8000-time steps since I have the results already; consequently, the discretization errors are considered negligible.

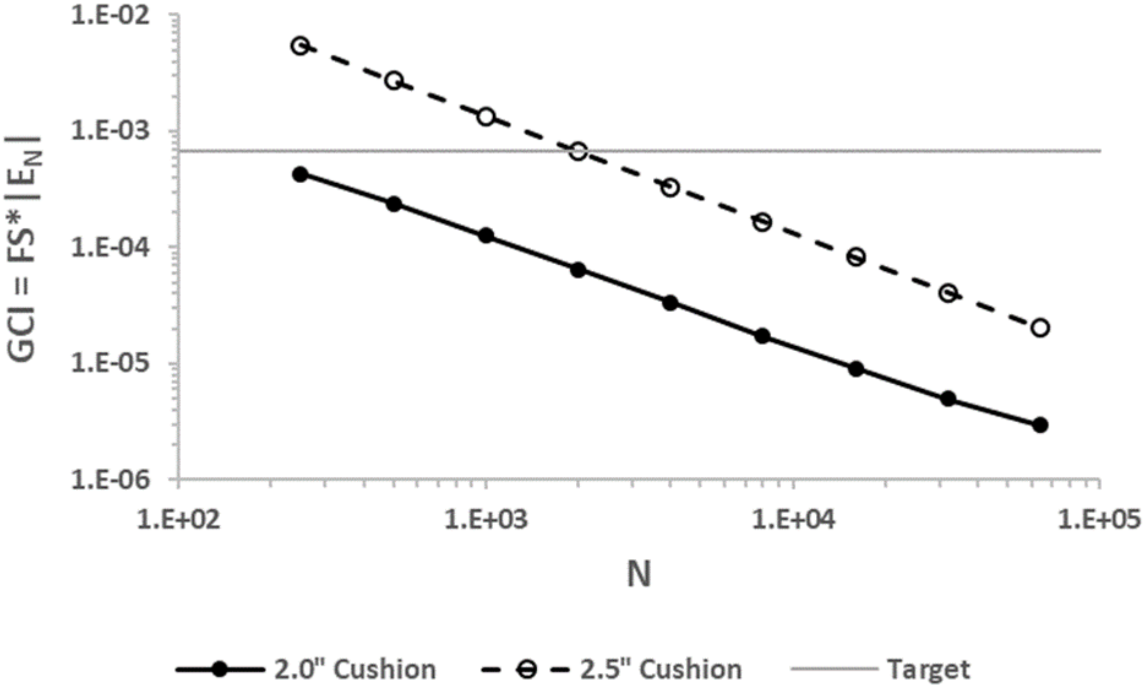


Figure 4.14: GCI for validation and certification simulations

Iterative solver errors

Iterative solver errors are irrelevant to explicit dynamics codes typically used for aircraft seat certification. In other applications and disciplines where implicit solutions are involved, it is essential to show that iterative solver errors are negligible (<1%) compared to the discretization errors you are trying to estimate.

Non-physical algorithm knobs

Non-physical algorithm knobs will bias simulation results. Be forthright about their existence and limitations and document a strategy for use. Knob selection will be dependent on discretization.

In some cases, the bias error and uncertainty can be estimated. For instance, the impact of mass scaling can be estimated by comparison to solutions without mass scaling, and in some cases, the assessment can be performed for a simpler nearby problem. I once observed a formal assessment of the interactions of discretization and an hourglass parameter such that an extrapolated solution could be estimated in the limit of infinite discretization and no hourglass stiffness. Numerical errors could then be estimated for finite discretization and hourglass stiffness.

Quantify sensitivity to knob selection when estimation of bias errors is not possible. This is a subjective process that is heavily based on the experience and judgment of the analyst. I typically observe one-at-a-time sensitivity studies, but I’ve seen interactions assessed in a more formal matrix of parameter selections. Judgment in the ranging and selection of parameters should be documented, and sensitivities should be included in the uncertainty rollup for simulation solution errors.

(SAE, 2021) recommends that mass scaling or hourglass energies be limited to 5% for critical components and 10% for non-critical components. This best practice is a useful guidance, but the impact of parameter selection on simulation results needs to be quantified.

There are no non-physical algorithm knobs in the demonstration.

Roundoff errors

Roundoff errors will bias simulation results. Good coding practices can minimize roundoff errors, but the analyst cannot control them unless they are writing user routines. Roundoff errors can be estimated by comparing simulation results with single and double precision. For the demonstration, roundoff errors were assessed to be less than 1.7×10^{-8}.

Parallel and platform inconsistencies

Users and regulators should know that explicit dynamics codes are sensitive to parallel and platform inconsistencies. Code developers test cases on different platforms, and possibly with different processor counts before new code is released, so this is an excellent place to start when looking for evidence of parallel and platform inconsistencies. I’ve observed proprietary data showing that even simple tests can produce different answers on different platforms.

Parallel and processor inconsistency is also application-specific, so user application models should be assessed by running an application model with different processor counts on other available platforms. I know of one sensitivity study that used an explicit dynamics code for a crash-type scenario. The results are given in Table 4.14. The most significant effect occurred when simulation results varied enough across platforms or with different processor counts that the solution changed critical load paths.

Parallel and processor inconsistencies can also be estimated by maintaining a mission-relevant test under configuration control and rerunning the model in the future when new platforms become available, which will also enable simulations on more processors.

Sensitivities should be included in the uncertainty rollup for simulation solution errors.
The best practice is to run validation simulations and certification predictions on the same platform with the same processor count. Recognize, however, that this does not eliminate the potential error; it only ensures that the error is consistent in the two sets of simulations. Platform inconsistencies could not be estimated for the demonstration because I only have one platform, my laptop. Potential parallel inconsistencies were assessed by turning off multithreading in Excel and performing Monte Carlo simulations on one to eight processors. In all cases, simulation results were identical to machine precision.

Table 4.14: Parallel and platform inconsistencies for a "crash" scenario

Characteristics of the Application	Sensitivity
Limited plasticity, no contact, no failure, or failure propagation	0%
Large plasticity, pervasive contact, with failure and failure propagation QOI: stress or strain	+/-3%
Large plasticity, pervasive contact, with failure and failure propagation Derived QOI: SRS or jerk	+/-8%
Different load paths with different processor counts or platforms	+/-50%

Restart inconsistencies

Avoid using restart capabilities; otherwise, assess sensitivity to different restart histories. Of course, restart is typically used for exceptionally long run-time problems prone to crashing, so sensitivity studies with application models may be impractical. Consider sensitivity studies with smaller surrogate models. Sensitivities should be included in the uncertainty rollup for simulation solution errors.

Mathematically ill-posed features

Be forthright, acknowledge features known not to be convergent, and document a strategy for their use. The strategy could involve ensuring global convergence without the feature and then adding the feature back into the model. Calibration is always part of the strategy. Calibration to separate effects tests (SETs) is preferable, but it is critically important that the stress state when calibrating to a SET is the same as when the feature is used in an application.

I have also seen cases where a non-convergent feature was calibrated with system-level test data. This led to an uncomfortable situation where the same feature, used in separate locations of the same component (material), required different calibration parameters because the stress states were different.

Sensitivity studies to calibration parameters should be performed and included in the uncertainty rollup for simulation solution errors.

User input errors

User input errors are exceedingly difficult to detect, especially for large input files, unless the code bombs or the results are egregiously wrong. The potential impact of undetected input errors is difficult to quantify. Simulation errors resulting from input errors are often masked by calibration activities or attributed to other sources of uncertainty. The potential for user input errors should be acknowledged and treated as an intangible. That said, the potential for user input errors can be minimized through best practices, lending credibility to simulation results if documented.

Code developers can play a role in reducing user input errors. Documented sample problems and training materials are helpful. Inappropriate combinations of input parameters should be flagged to the user. Code developers can help by putting “guardrails” on constitutive models.

Guardrails flag for the user (without terminating the calculation) when a constitutive model is used outside its limits of applicability. Another code feature might be mirroring back of code inputs in a more easily understandable format, e.g., tabular inputs might be charted.

The user's credentials are important. Management’s responsibility is to match the skill sets (education and experience) of staff with the application's complexity and consequences. Continuing education and professional certifications (e.g., as offered by NAFEMS) should be encouraged. Organizations typically have “promotion” ladders, reflecting to some degree the technical staff's education, experience, and performance. Formal mentoring by senior staff is good practice.

Self-inspection of inputs is expected but not adequate. Peer inspection of large models is mind-numbing and not effective. Reuse of large blocks of input that were developed by senior staff for nearby applications is common[18] and potentially useful. Some organizations have simulation governance to ensure consistency in applying simulation to classes of applications. Internal peer review is good practice.

Pre/post-processing errors

Pre/post-processing errors can occur when information needs to be processed before it is useful for code input or to compute relevant QOI from code outputs. Pre/post-processing is typically performed with special programming, sometimes using tools such as MS Excel or MATLAB. Pre/post-processing tools should be standardized, documented, subjected to some degree of acceptance testing, and placed under organizational configuration control.

Demonstrate Model Sustainment

Codes and the computing environment are not static. An organization can expect new code releases and new platforms over time. Bigger computers allow problems to be run on more processors. The best practice is to maintain an application-relevant acceptance test that can be rerun when these events occur. Historical sensitivity (if any) of simulation results to these events should be included in the uncertainty rollup for simulation solution errors.

Model sustainability is evaluated in Table 4.15 for the demonstration. The sustainability test is the validation model for A15008, run with 8000 time-steps. The baseline model was frozen on 3/27/24. Earlier definitions of a sustainment test showed no changes in results (within machine precision) over time despite multiple software updates in the OS and EXCEL. The final definition of the sustainment test came late in the project, so there is limited history supporting assessment. However, simulation results surprisingly changed with the first reassessment on 5/5/24. The change was smaller than 1 lb_f in 1500 lb_f, so it does not impact validation or certification predictions. The change is color-coded yellow as a flag.

[18] There is a potential downside to reuse. I know of an example where a senior analyst developed a user routine for a class of analyses. The input routine was reused for many years by other analysts through generations of nearby applications, sometimes tweaked or enhanced. A bug was discovered in the original user routine when I pushed for formal verification. The bug was corrected, and the user routine brought under configuration control by the code team.

EXCEL technical support was consulted on the issue and responded. “There are many reasons roundoff errors can propagate differently when a model is rerun. While Excel is known for its stability and consistency, it is not immune to bugs or differences between builds. However, significant differences in results between Excel builds are rare.”

Table 4.15: Model sustainability

Date	Platform	OS	EXCEL	$\lvert E_S \rvert$
3/27/2024				0.00E+00
5/5/2024		Changed	Changed	1.67E-04

4.3.4 Summary assessment of solution errors for demonstration problem

Simulation results should be bias-corrected for known sources of error unless they are demonstrated to be negligible in the application context. Uncertainties in the assessment of simulation solution error will cloud validation and certification simulations. I recommend that the applicant provide a reference to specific application-relevant acceptance tests when seeking CbA for a new seat design. The applicant should work with the code developer to find relevant tests in the code documentation; otherwise, they should develop and document their own set of acceptance tests. Even if relevant tests can be found in code documentation, it is preferable if the applicant reruns the tests on their own hardware and operating systems. The acceptance tests can be regression tests, verification tests (order of accuracy), and a sustainability test. Appendix D documents the acceptance tests that were developed for this demonstration. Simulation solution errors/uncertainties are assessed to be negligible (< 1 lb_f) or intangible for this demonstration's validation and certification simulations. Table 4.16 summarizes the evidence.

Demonstrating negligible solution errors/uncertainties may be difficult when explicit dynamics codes are used for realistic seat designs. This is partly because the required level of grid refinement may not be practical and, most assuredly, if mathematically ill-posed features (fasteners, contact, etc.) are used. The use of non-physical knobs (hourglass parameters and mass scaling) is common. In addition, explicit dynamics codes can be sensitive to parallel and platform inconsistencies, depending on the application.

For regulatory purposes, reference to best practice or simulation governance is useful, but not sufficient. Regulators should also look for statements of sensitivity or estimates of errors and uncertainties. If not negligible, their impact on regulatory decision metrics should be quantified.

Table 4.16: Simulation solution errors/uncertainties are negligible or intangible

Source	Errors/Unc	Comments and Technical Approach
Undetected code bugs and algorithm deficiencies	0	• A RTS was developed to assess application-specific coverage of physics and constitutive models. The RTS covers 100% of the physics capabilities individually. Acceptance is taken as machine precision. • A VERTS was developed for a nearby problem involving a linear spring-mass-damper system with a known analytic solution. The verification test covers 100% of interactions between the ICs, BC, constitutive model, and conservation equations. The observed order of accuracy differs from the formal order of accuracy by less than 10%. • Simulation of the validation test (A15008) is placed under configuration control and used as an acceptance test. The acceptance test covers 100% of the interactions between the ICs, BC, constitutive model, and conservation equations. The observed order of accuracy differs from the formal order of accuracy by less than 10%
Pre/post-processing errors	0	• No pre/post-processing required. Charting performed within Excel.
User input errors	Intangible	• Inputs are self-verified at various times, but there is no independent check that correct values are passed into the subroutines.
Mathematically ill-posed features and capabilities	0	• Computational model does not involve mathematically ill-posed features and capabilities.
Restart inconsistencies	0	• Not applicable. Restart capability not available or used.
Parallel inconsistency	0	• The model does not involve potential triggers for sensitivity. i.e., pervasive contact, failure, and failure propagation. A sensitivity study was performed by limiting the number of processors available for simulation on my Dell laptop. No change in simulation results down to machine precision was observed
Platform inconsistency	Intangible	• My Dell laptop is the only platform available to run the simulations.
Roundoff	0	• Sensitivity to roundoff is assessed by comparing results computed with double precision (assumed to be the truth) to results of the acceptance test computed with single precision. Roundoff error < 1.69×10^{-8} for single precision (and much less for double precision), which does not impact reported lumbar loads.
Non-physical solution knobs	0	• Non-physical solution knobs are not used.

Iterative solver tolerances	0	• Not applicable for explicit solution scheme used.
Discretization	0	• Discretization errors were estimated by performing Richardson Extrapolation and modified by Roache's factor of safety (FS=1.25). With 95% confidence, the numerical errors are less than 1 lbf in 1500 for N=8000 time-steps for both baseline seat design and the nearby design.
New code release	0	• Potential changes in the sustainability test are correlated over time with updates in all relevant software. |ES|=1.67e-4 is correlated with changes in the OS and EXCEL. This is smaller than 1 lb_f out of 1500 lb_f and can be ignored.

4.4 Assess accuracy of simulation conceptual model

The primary product of this element is a risk-informed decision to accept or reject the model for regulatory predictions. Validation is the model assessment process, from the perspective of intended use, by comparing simulation results with relevant experiment results when both simulation and test results are clouded by uncertainty. Model acceptance is distinguished from model validation and is judged based on the magnitude of the discrepancy between predictions and relevant test data. Model calibration is an empirical adjustment of the model to improve agreement with data and reduce prediction errors and uncertainties. Confidence in the model is derived by applying the process to a hierarchy of validation tests of increasing complexity and application relevance.

4.4.1 Assess model accuracy

Validation hierarchy

Confidence in the model is derived by applying an assessment process to a hierarchy of validation tests of increasing complexity and application relevance. Many organizations honor the spirit of a validation hierarchy, but (Luckring et al., 2023) offers a valuable perspective and demonstrates the process for a very complex system.

Figure 4.15 shows Luckring's conceptualization of the validation hierarchy. Physics taxonomy is at the bottom of the hierarchy and focuses on a complexity decomposition that follows physics. System Architecture is at the top of the hierarchy and focuses on complexity decomposition that follows the system's functionality. A Transition Tier bridges the gap between these two perspectives. A validation hierarchy is application-specific; every level can be decomposed into multiple validation activities as appropriate. The propagation of errors and uncertainties up the hierarchy to system-level predictions is a highly challenging research topic; consequently, each validation activity in the hierarchy is usually treated as a pass/fail gate or an opportunity to calibrate the model if necessary.

Typically, there are more data to assess models at the bottom of the hierarchy and little and possibly no data at the top. The application relevance of the data must be documented in all validation activities. You want to know if the application parameter space lies within the validation data (interpolation), if it lies well outside the application parameter space (extrapolation), and whether validation data can test a model's ability to predict trend behaviors with respect to design or environment parameters.

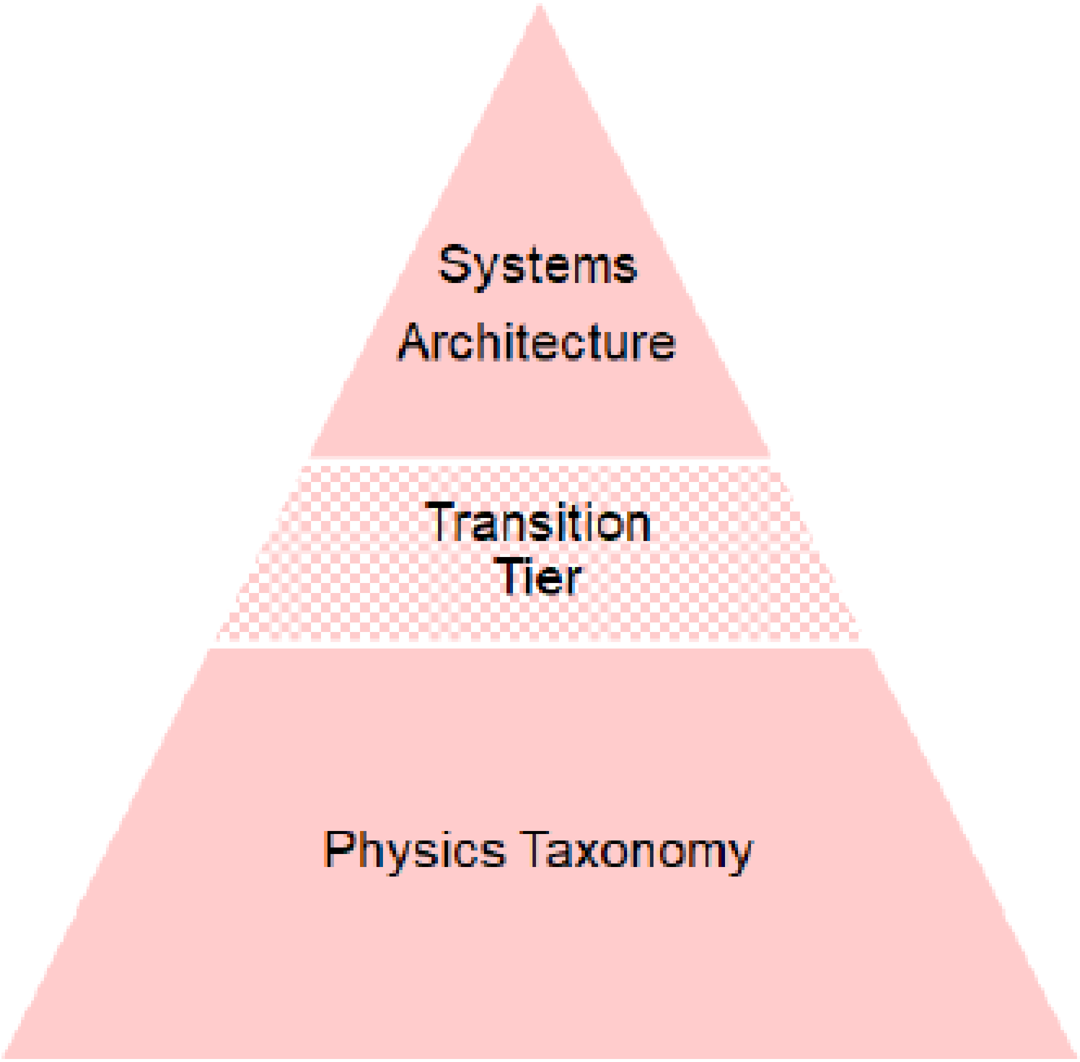


Figure 4.15: Conceptualization of a validation hierarchy

Validation metric

Validation is the model assessment process, from the perspective of intended use, by comparing predictions (P) with relevant experiment measurements (M). A validation metric in the form of a discrepancy measure is defined by

$$E = \ln \frac{M}{P} \, , \qquad \textbf{0-67}$$

which limits to the relative error when predictions are close to measurements. This validation metric honors that the QOI in this application can never be negative and would be inappropriate when M and P can have different signs.

The validation metric is generally epistemically uncertain because of epistemic uncertainties in predictions (P) and test measurements (M). Uncertainties in predictions fall into two categories: simulation solution error/uncertainties, Section 0, and alternate plausible models if the conceptual model is not fixed. This report focuses on the former. Uncertainties in test results will be addressed in Section 0.

Validation against a single test is the least informative form of validation; see Figure 4.16. The circle represents the discrepancy measure for a single test, and the vertical green line represents the application parameter space i.e., design point (DP).

Extrapolation is always required when only a single test is available. The horizontal dashed lines represent validation acceptance limits, which are discussed in Section 0. The best practice is always to show the relationship of test data to the design point so that regulators can judge the degree of extrapolation (or interpolation) involved in certification predictions.

It is risky, but common, to judge model acceptance based on a single value of discrepancy and to interpret the single value as an estimate of model form error (bias),

$$\mathrm{E_{mf}(DP) = E}\,, \tag{0-68}$$

that is evaluated at the validation point and applied at the design point (DP). This estimate of E_{mf} can be either conservative or non-conservative, as will be discussed later.

As a minimum, a database of replicate tests is required to estimate model form error and uncertainty. Figure 4.17 shows the original test plus additional replicate tests as solid black symbols. Seemingly identical replicate tests are subject to sources of precision uncertainty (see Section 0) and will produce different results. Model bias is estimated as the median of the discrepancy measures and represented by the short horizontal line in the data cluster. The scatter of data about the median is an estimate of model uncertainty. Model form error (E_{mf}) is estimated with validation data,

$$\mathrm{E_{mf}(DP) = E_{mf}(bias, unc)}\,. \tag{0-69}$$

The median, $E_{mf,50}$, is used to judge model acceptance. In addition, E_{mf} is applied at the design point, even though the replicate test provides no evidence that the model can predict trends with design and environment parameters. Risks associated with this assumption will be addressed later. Limits of applicability of the model still cannot be assessed.

Figure 4.17 illustrates the risk of interpreting the discrepancy measure of a single test as an estimate of model bias. The discrepancy measure is as likely to be conservative as not conservative, depending on where the single test lies relative to the median of replicate tests if they were available. Using the FAA prescription for bias correction (FAA, 2018) could lead to bias correction when it is not needed and no bias correction when it is required.

A single validation test or replicate tests at a single point in the design space say nothing about the model's ability to extrapolate to other points in the design space, nor do they give insight into the model's limits of applicability.

Replicate tests are an inefficient use of testing resources. The same number of tests can yield more information by exploring the sensitivity of model predictions to design or environment parameters.

Figure 4.18 depicts the case where the design point interpolates on the database (dashed green line). The black up-to-the-right line is an empirical correlation line that passes through the data. The line is a measure of trend bias. Ideally, the discrepancy measures would show no sensitivity to design or environment parameters (trend bias is a constant), indicating that the model fully predicts any observed trends in the data. In this case, the model form error given by Equation 0-69.

More generally, trend bias is an empirical function of design and environment parameters,

$$trend\ bias = f(design\ paramters, environment\ paramters)\ , \qquad \textbf{0-70}$$

and model form error is given by,

$$\mathrm{E_{mf}(trend\ bias, unc)}\ , \qquad \textbf{0-71}$$

where the uncertainty is characterized relative to the trend line. The median model form error evaluated at the design point, $E_{mf,50}$(DP), is used to judge model acceptance, which would be the case in Figure 0-18. The limits of model applicability are shown as red lines in Figure 4.18.

Tests that span a range in parameter space give insight into the model's ability to extrapolate to other points in the design space and into its limits of applicability, i.e., where the trend line passes outside the acceptance bounds. Figure 4.19 shows that acceptable errors where data exist might leverage into very large errors when extrapolating to the design point.

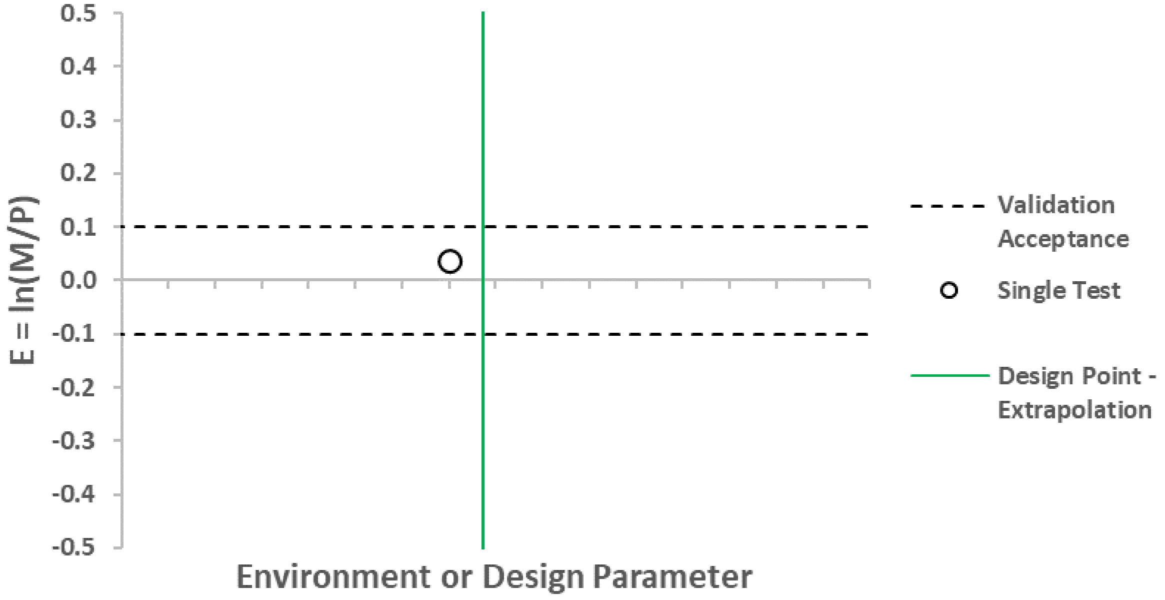


Figure 4.16: Model validation with a single test

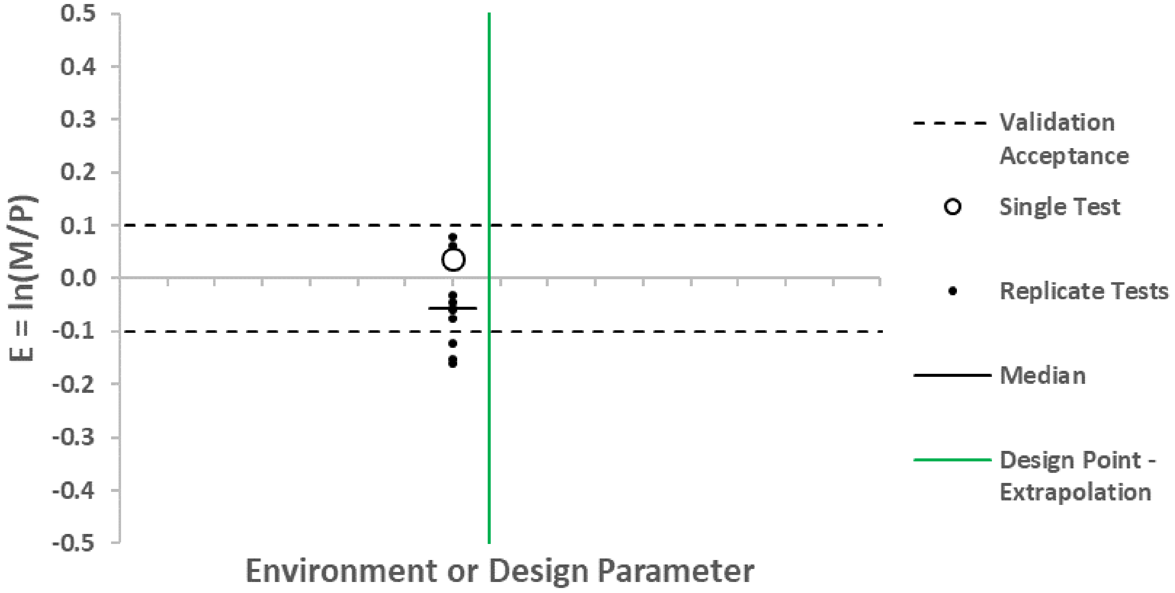


Figure 4.17: Model validation with a set of replicate tests

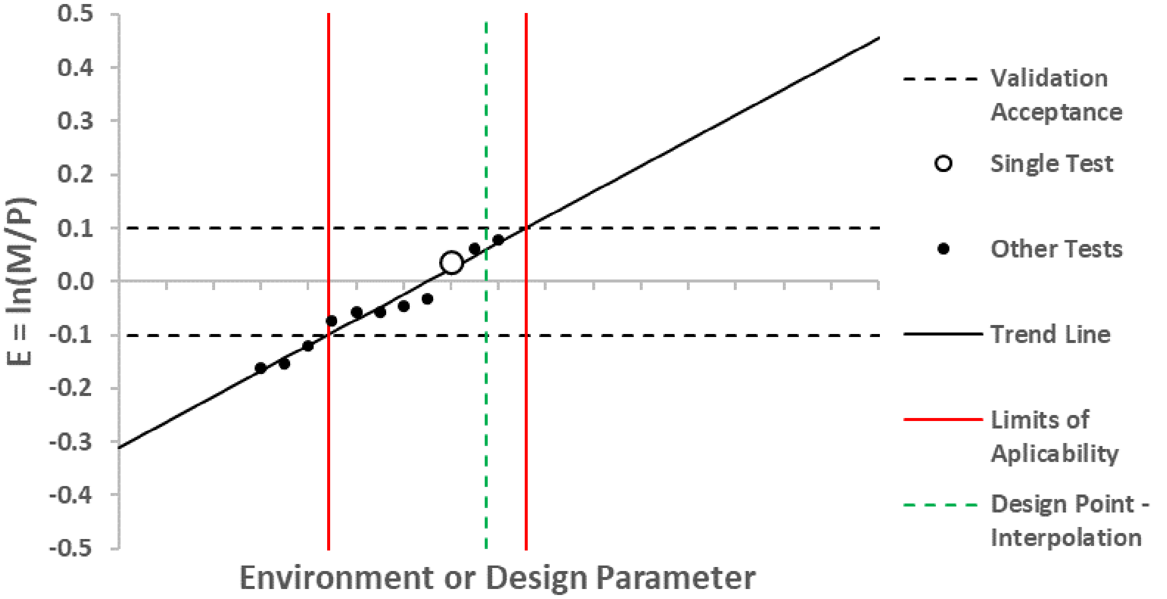


Figure 4.18: Model validation when the design point interpolates

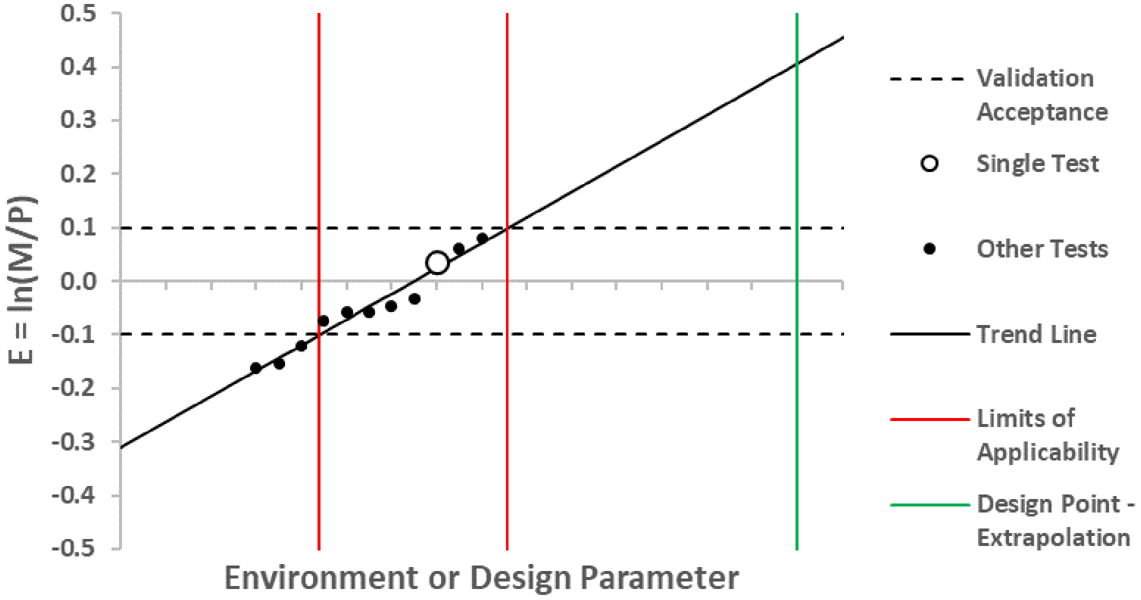


Figure 4.19: Model validation when the design point requires extrapolation

4.4.2 Testing errors and uncertainties

Testing errors and uncertainties distort and cloud validation assessments and CbT. Figure 4.20 organizes the sources of errors and uncertainties in testing. The overarching goal is to bias correct for known sources of testing errors before making validation assessments and to reflect the rollup of testing uncertainties in discrepancy measures. Sources of testing errors and uncertainties are briefly discussed next.

Scaling distortions in the test conceptual model

Test results are application-relevant only if the test conceptual model matches the application conceptual model. Otherwise, test results will be distorted relative to what would be expected in an application. The equivalency of test and application conceptual models is sometimes accomplished with full-scale testing replicating relevant environments with the same system state and the same controlling physics. Sled tests for aircraft seat certification fall into this category. Tests are conducted with an actual seat and the FAA-prescribed ATD. The sled test closely replicates the prescribed regulatory environment. Lumbar load measurements can then directly inform regulatory decisions if other sources of testing error are negligible.

This is not always possible. For example, it isn't easy to test the performance of a full-scale aircraft wing. Typically, testing is performed at a small physical scale in a wind tunnel. If there is equivalency in physics scaling, test results can be applied directly to the full-scale application. This is usually accomplished by demonstrating the equivalency of all the controlling dimensionless groups, such as the Reynolds number.

Accuracy errors

Accuracy errors are bias errors that should be recognized, managed, and minimized. Test results should be bias-corrected for known sources of bias error.

Measurement distortion

The act of measurement often distorts the thing you are trying to measure. For instance, you can mount a transducer on a component to measure the vibration environment it experiences. Unless the transducer mass is negligible, the transducer distorts the component's response, and you must consider the transducer and component as a coupled system. “Wall effects” in wind tunnels are another example where the act of measurement distorts what you are measuring. Long communication lines to data recorders can distort measurements. Measurement distortion must be recognized, managed, and minimized, and the results bias-corrected. The latter typically involves some form of supplemental analysis.

Alternate measurement techniques

Alternate measurement techniques will produce different results. Often, there are multiple ways to measure the same QOI in a test. For instance, temperatures can be measured with thermocouples or pyrometers. Type C and Type K thermocouples are within the class of thermocouples. There are single-color and two-color pyrometers within the broader class of pyrometers. The modulus of elasticity can be measured with traditional dog-bone tests or beam vibrations. Both the FAA (emergency landing conditions) and the Airforce (pilot ejection from the aircraft) are interested in lumbar injuries. The FAA creates its environments in a sled test, while the Airforce uses a drop tower. Different ATDs are used as well.

User errors

To err is human, and I have witnessed multiple examples in my career. Inappropriate gauges can be mounted in the wrong place, in the wrong way, or with an incorrect orientation. Measurement channels can be crossed to data recording systems. My friend in graduate school spent over a year trying to reconcile significant differences between his model and the test data he took, only to discover that he misread "the decade dial" by an order of magnitude. User errors can be managed with well-defined processes and checklists and, if possible, testing the instrumentation before the test. User errors are sometimes detected after the test; otherwise, the potential for undetected user errors should be treated as intangible.

Post-processing errors

Post-processing errors can occur when information needs to be processed before it is useful. As examples of post-processing, the head injury criterion (HIC, used in crash assessments) and shock response spectrum (SRS, used in structural dynamics) are QOIs that are functionals of the measured acceleration history. Post-processing is typically performed with special programming, sometimes using tools such as Excel or MATLAB. This programming usually escapes the same scrutiny and quality assurance expected with commercial finite element codes. Pre/post-processing tools should be standardized, documented, subjected to acceptance testing, and placed under organizational configuration control.

Precision uncertainty

Organizations like the American Society for Testing and Materials (ASTM) often provide formal estimates of repeatability and reproducibility in their testing reports. For instance, (ASTM, 2010) reports that the repeatability and reproducibility for testing of flexible cellular materials are 1.5% and 3.8%, respectively.

Often, we pool data from different references found in the literature to increase the amount of data available for analysis. This type of meta-analysis violates the formalism of repeatability and reproducibility; however, pooled data can be desirable. Pooled data is more likely to encompass truth when the sources of precision uncertainty are represented more completely.

Precision uncertainty is a known source of uncertainty for all test data, but sources of precision uncertainty may not be sufficiently represented in the database to support reliable quantification. One-of-a-kind tests are also subject to precision errors, although there is no basis for direct quantification. Precision uncertainty can sometimes be inferred (sometimes subjectively) from other sources.

Calibration is risky when precision uncertainty is not recognized or acknowledged. Calibration to a single test hardwires a bias of unknown magnitude and sign. Different results will be produced if replicated tests are performed, if other organizations try to reproduce the results, if different samples or units are tested, if different gauges are used, or if a different measurement technique is used. Which result is the most “correct”?

Precision uncertainty is an important intangible to note when there is no basis for quantification.

Repeatability uncertainty

Replicate testing within the same organization will yield different results. Replicate testing is intended to quantify the uncertainty associated with test procedures and testing equipment. Ideally, all replicate tests should be performed by the same test operator, on the same test equipment, using the same sample or unit, in as close a period as possible.

The same sample (or unit) cannot be reused if testing is destructive. To minimize confounding with sample-to-sample uncertainty, samples (or units) should be prepared to the same specification, by the same procedures, by the same person, at the same time, and with materials obtained from the same stock.

Reproducibility uncertainty

Reproducibility uncertainty is associated with the ability of different organizations to reproduce test results reported by others. Reproducibility testing is intended to quantify the uncertainty of different operators using different units of the same test equipment. Each organization follows the same test procedures. Ideally, the same sample (or unit) is passed from organization to organization for testing. If the tests are destructive, all the samples (or units) should be prepared by a single organization, as noted above.

Sample-to-sample (S2S) uncertainty

Testing on nominally identical materials (or units) will produce different results when materials are obtained from the same supplier or other suppliers at various times. S2S uncertainty results from variability in manufacturing, source material, and time. Testing within an organization (e.g., material characterization tests) often tries to minimize the S2S uncertainty by testing on materials (units) obtained from the same stock.

Recognize that S2S uncertainty, as observed in testing, is *never* representative of S2S variability of the population of fielded units after certification. The variability of fielded units can only be obtained by a posteriori testing of fielded units. This is because the dominant sources of variability in fielded units (from manufacturing and materials) will never be known and quantified to the degree necessary to ensure that they are replicated in testing.

Gauge-to-gauge (G2G) Uncertainty

If a gauge does not measure “truth,” it is biased. What is truth? Gauge manufacturers typically specify that “the gauge is accurate to +/-X% full range.” G2G uncertainty is a statement about the population of gauges the manufacturer produces and is a consequence of variability in manufacturing, materials, and time. G2G uncertainty also exists when gauges of a similar type are obtained from different manufacturers.

When repeatably exposed to the same environment, the uncertainty of any specific gauge is much less than the manufacturer's specification for G2G uncertainty. For analysis purposes, a particular gauge produces a fixed but unknown result somewhere within the manufacturer’s specification. The best practice is calibrating each gauge individually, virtually eliminating gauge bias and uncertainty.

Sampling uncertainty

Data sets are finite, generally small, and sometimes consist of only one test. Any statistic (e.g., min/max, mean, standard deviation, percentile, correlation coefficient, etc.) computed from a dataset of finite size will differ from the statistic calculated from a different dataset of the same size randomly drawn from the same population. This is referred to as sampling uncertainty and is an epistemic uncertainty. More data is better, but how much is enough? Appendix F.2 provides some guidelines for random sampling. Approximately 10 to 100 samples are needed to estimate the median, and 190 to 1900 samples are needed to estimate the 95th percentile. Datasets this size are highly unusual; consequently, the sample size should be reported, and sampling uncertainties should be acknowledged as important intangibles.

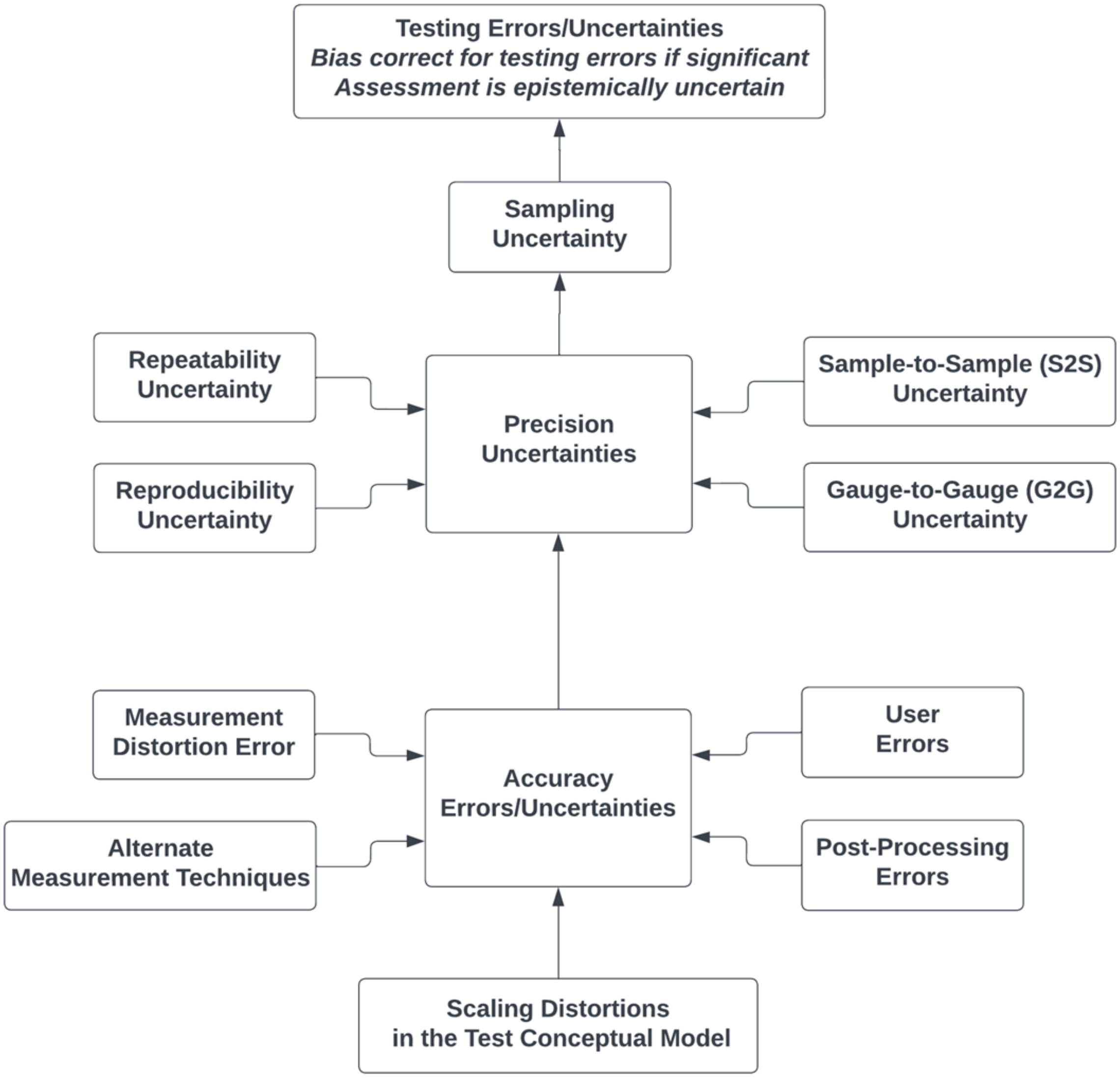

Testing Errors/Uncertainties
Bias correct for testing errors if significant
Assessment is epistemically uncertain
Sampling Uncertainty
Repeatability Uncertainty
Precision Uncertainties
Sample-to-Sample (S2S) Uncertainty
Reproducibility Uncertainty
Gauge-to-Gauge (G2G) Uncertainty
Measurement Distortion Error
Accuracy Errors/Uncertainties
User Errors
Alternate Measurement Techniques
Post-Processing Errors
Scaling Distortions in the Test Conceptual Model

Figure 4.20: Sources of errors and uncertainties in testing

4.4.3 Model acceptance

Model acceptance is a risk-informed decision, separate from assessment, which is informed by,

1. Risk tolerance,
2. Historical precedent for similar applications,
3. Completeness and relevance of the validation hierarchy,
4. The degree of model calibration,
5. The maturity and quality of testing, and the quantity of test data,
6. Acknowledgement and potential consequences of specific intangibles,
7. The degree to which the design point interpolates or extrapolates on the validation database, and

8. How well the model represents system level validation data (bias and ability to predict trend behavior).

Items 7 and 8 are the only quantitative elements informing the decision. In general, the assessment of model form error is epistemically uncertain because of uncertainties in prediction and testing. I recommend that model acceptance be based on the median evaluated at the design point without consideration of uncertainty,

$$\left|\mathrm{E}_{\mathrm{mf,50}}(\mathrm{DP})\right| < 0.10 \,, \qquad \textbf{0-72}$$

which is an expansion equivalent to the FAA's validation acceptance criteria. It would be extraordinarily difficult to ensure that 95 percent of the distribution of E_{mf} falls within these acceptance limits, especially if the model is biased to near the acceptance limits. The “penalty” for large uncertainty in E_{mf} will be addressed next.

The decision to accept the model must be accompanied by guidance on how to make predictions at the design point that account for uncertainties,

$$\mathrm{P}(\mathrm{DP}) = \mathrm{S}(\mathrm{DP})\mathrm{e}^{\mathrm{E}_{\mathrm{mf}}(\mathrm{DP})} \,, \qquad \textbf{0-73}$$

where S(DP) is a single computer simulation at the design point.

Note that this use of the model form error, with a trend bias, is a form of empirical model calibration. There is no physics in E_{mf}. As the (NaRC, 2012) commented that using simulation to interpolate can be risky, and I might add that this is especially true if interpolating in a high-dimensional space of design and environment parameters and the test spacing is large. The physics in the model might be deficient in the spaces between tests. For example, the physical world might exhibit resonance behavior that would not be observed in widely spaced tests. Consequently, expert judgment is required in the use of Equation 0-73.

Extrapolation is even more challenging, as depicted in Figure 4.19. Although discrepancies may be acceptable where data exist, even small discrepancies in the model's ability to predict trend behaviors can lead to large errors at the design point, causing the model to be rejected. A NRC staff member told me of an experience where they were reviewing validation results where the code only had turbulent heat transfer correlations, and the test was well within the laminar regime.

Extrapolation can be even more risky than interpolation. The nature of physics can change in a way not represented in the conceptual model. For example, flows can transition from laminar to turbulent, threshold phenomena may occur (failure, phase change, runaway chemical reactions, etc.), or resonance phenomena may occur. Consequently, expert judgment is required in the use of Equation 0-73.

4.4.4 Calibrate the model

Calibration is an empirical adjustment of a model to improve agreement with test data or the purpose of improving predictions where there is no data. The decision to calibrate or accept a calibrated model is risk-informed.

Calibration may be necessary, but the indiscriminate use of calibration sweeps many problems under the rug, thus undermining credibility. Calibration compensates for the following:

1. Missing or unknown physics that requires a temporary surrogate until research can resolve the gap,
2. Unknown model parameters not accessible through separate effects testing,
3. Simulation solution errors,
4. Testing errors, and
5. Errors in the conceptual model.

The first reason is the most compelling. There are high-consequence applications dependent on simulation where calibrated “knobs” allow important work to proceed, while long-term research activities are slowly replacing the knobs with validated physics.

The second reason is rare. I know of only one example. Ideally, parameters in constitutive models are calibrated to separate effect test data at lower levels in the validation hierarchy. This is expected and not the kind of calibration we are concerned about here. The concern is calibrating parameters in constitutive models based on system-level test results, but this is common.

We usually calibrate because models are demonstrably wrong, i.e., we do not like the agreement between model predictions and the data we have. The likely suspects are numerical errors or a deficiency in some element of the conceptual model. It is walking on thin ice when calibrated model parameters have nothing to do with why the discrepancy exists. Then, the model is used to extrapolate to untested conditions, expecting more accurate predictions. Calibration is not a substitute for a formal assessment of simulation solution errors (verification). Calibrating a material parameter to compensate for numerical errors (or code bugs) will not lead to improved predictions. I once observed an analyst systematically vary the discretization in his computational model until simulation results agreed with a single test result. This does not improve prediction.

Calibration is not a substitute for formal model accuracy assessment (validation). The recommended use of Equation 0-73 is the product of formal validation and involves an empirical correction to the conceptual model, S(DP), through the trend bias term in E_{mf}. This is a form of model calibration and is limited to small corrections through the acceptance criteria given by Equation 0-72.

Calibrated parameters should be physically reasonable. I observed a presentation where gravitational acceleration (a constant of nature, $g=32.2\ ft/s^2$) was calibrated along with a host of other parameters in the model to improve agreement with data.

I observed an effort where a material property was calibrated to improve comparison with observed component responses. Calibration produced material property values that were well outside limited available data for the specific material and more abundant data for similar materials; however, calibration was not necessary when geometric fidelity was improved in the conceptual model.

Calibrated parameters may not be unique, thus undermining credible extrapolation to untested conditions. Because of testing uncertainties, calibration to a single system test will hardwire in a bias, which is equally likely to be conservative as not. Calibrating to different QOI for a single system test often leads to different values of calibrated parameters. Calibration to multiple system tests is preferable; but commonly, wildly different values of a calibrated parameter result.

A healthy skepticism of calibration is justified. Some best practices for responsible calibration include:

1. Formal verification and validation should precede calibration.
2. Explicitly document what is being calibrated, why, and a strategy for calibration.
3. Calibration should not make substantial changes in simulation results.
4. Calibrated parameters must be physically reasonable. Do not calibrate physical constants or material properties you know.
5. Do not calibrate to a single test because you will hardwire in a model bias, which is equally likely to be conservative as not.
6. Calibrate only a few select parameters to an ensemble of data. Otherwise, the calibration is not unique.

4.4.5 Assessment of model accuracy for the demonstration

Validation hierarchy

Figure 4.21 shows the validation hierarchy for the demonstration. The assessment of model accuracy will proceed from the bottom to the top of the hierarchy.

Physics taxonomy

Physics taxonomy deals with calibrating constitutive models to separate effects tests. Appendix C.3 discusses the calibration of the CF42 (AC) constitutive model, which is found acceptable for the demonstration. As part of the simulation conceptual model, the CF42 (AC) constitutive model, is frozen with nominal fitting parameters without uncertainty.

In general, calibration will be required for other constitutive models for a more general finite element model (FEM) assessment of the system. One example is the FEM for the ATD, which involves mass distributions and stiffness/moment parameters for all the joints. The necessary masses of ATD elements for the demonstration are summarized in Table 4.4 and treated without uncertainty as part of the simulation conceptual model.

Other constitutive models requiring calibration that might appear in a complete FEM analysis are failure criteria for fasteners, material models for frame materials, and constitutive models and failure criteria for belts and harnesses.

Transition tier

The transition tier involves two validation activities for the demonstration. The first addresses validating the lumbar load model for seats without cushions. In this case, Equation 0-24 is an analytic expression for the maximum lumbar load. Twenty-one relevant tests can be found in (Adams et al., 2003) (Olivares, 2013), and (Pellettiere et al., 2011). Figure 4.22 shows the validation assessment for 15 tests with a Hybrid II ATD. Figure 4.23 shows the validation assessment for 6 tests with an FAA-Hybrid III ATD. The dashed lines represent the validation acceptance limits. The black line shows the trend line through the data. The red lines represent the limits of applicability of the model, and the green line represents the environment of the application (DP) we are trying to predict.

The FAA-Hybrid III ATD appears to be more sensitive to environments compared to the Hybrid II ATD. To not confound differences in the ATDs, the focus should be on the FAA-Hybrid III ATD. The median error at the design point is E_{50} (FAA-Hybrid III,14G) = 0.048, which is within the validation acceptance bounds; consequently, the model is accepted. The observed environmental sensitivity severely limits the model's applicability to 13.1G to 16.4G environments. The model significantly underpredicts measured lumbar loads at 19G. This sensitivity is potentially associated with the assumptions of LT and UT rigidity. A sensitivity study showed that the predicted lumbar load is enhanced if a 1.0” rate-insensitive cushion is added to the model as a surrogate for pelvic “cushion”. The FAA also suggests that flexibility of a rubber column above the load cell could also lead to dynamic amplification of loads at higher G environments.

The scatter around the regression line should be interpreted as precision uncertainty, and more sources are represented in the Hybrid II data.

Validation of the model for initial static compression is the second activity in the transition tier of the validation hierarchy. Equation 0-6 is an analytic expression for the initial static compression. The initial static compression is measured with the seat back vertical in the lab frame of reference, i.e., aligned with g. Compression responds to the ATD UB weight plus pretension in the lap belt. Because of the pretension, the initial static compression remains the same when the seat is rotated 60^0 in preparation for the sled test.

Four relevant tests can be found in the literature, (Taylor, DeWeese, et al., 2017). Model assessment is shown in Figure 4.24 and the error at the design point is given by

$$|\mathrm{E}_{50}(\mathrm{FAA-Hybrid\ III, CF42\ (AC), 2.5"})| = 0.26\ , \qquad \textbf{0-74}$$

which is well outside the acceptance bounds; consequently, the model is rejected.

Mis-interpretation of the data is a possible cause, but the discrepancy is not resolvable in the available time, so calibration of the model is the only recourse. The modified equation for the initial static compression is given by

$$\frac{\varphi_0}{\varphi_c} = 1 - \left(\frac{F_0}{kF\, W_0}\right)^{1/a}, \qquad \textbf{0-75}$$

where kF = 0.4364 is a calibration factor obtained by minimizing the RMS error between data and the model using EXCEL's Solver functionality. I chose to modify W_0 because I have more confidence in F_0 and f_c, which are properties of the constitutive model, Appendix C.3. The risk is that W_0 is not the cause of the discrepancy, so Equation 0-75 should not be used for any other materials or conditions without additional assessment. The best solution is to resolve the cause of the discrepancy.

Assessment of the calibrated model for initial static compression is shown in Figure 4.25 and the error at the design point is now

$$|E_{50}(\text{FAA} - \text{Hybrid III}, \text{CF42 (AC)}, 2.5")| = 0.036\,, \qquad \textbf{0-76}$$

which is within the acceptance bounds; consequently, the calibrated model is accepted.

Figure 4.25 shows that the initial static compression exhibits a potential sensitivity to cushion thickness (not represented in the model), which limits the applicability of the calibrated model to cushion thicknesses of 1.6” to 4.4”. The need for calibration may be due to misinterpretation of the data, and the sensitivity to cushion thickness may be associated with the assumption of LT rigidity or a small sample size.

System architecture

There is only one test at the system architecture level of the validation hierarchy, CAMI A15008, for the baseline seat design. Table 4.17 summarizes the validation assessment. Only the single discrepancy measure can be computed, which is within the validation acceptance limits; consequently, the model is accepted.

Table 4.17: Model validation with baseline seat design

M (lb_f)	P (lb_f)	$\mathbf{E = \ln\frac{M}{P}}$
1048	1058.3	-0.00981

Validation against a single test provides no evidence that the model can predict trends with respect to design or environment variations, and extrapolation is always necessary with a single validation test.

Model form error based on a single validation test is computed from

$$E_{mf} = \ln\frac{P}{M}, \qquad \textbf{0-77}$$

which is epistemically uncertain because of simulation solution uncertainties (P) and testing uncertainties (M). Only one simulation is required for P, and Section 0 showed that simulation solution errors are negligible for the baseline seat design.

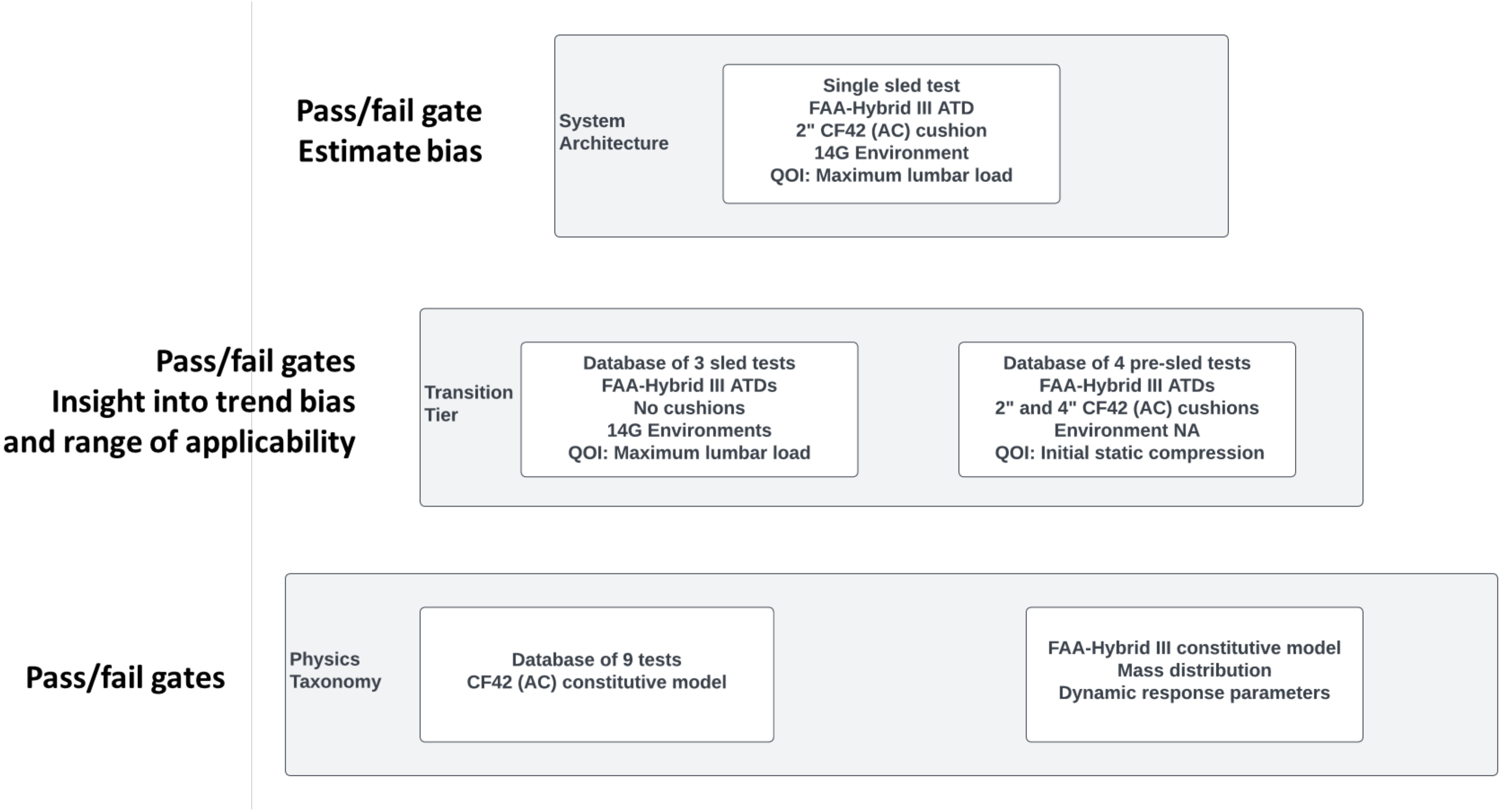
There is only one test, so the impact of testing precision uncertainties cannot be assessed directly. Still, it can be inferred from a database of related replicate tests conducted over a wide range of seat design and test environments. Appendix E shows that if many replicate tests were available, then the distribution of measured lumbar loads would be given by

$$\mathrm{M} = \ L_{50}e^{Laplace(0,unc)} \quad \textbf{0-78}$$

where unc = 0.058197. The best we can do is equate L_{50} with the measured lumbar load for the baseline seat design. This can lead to either conservative or non-conservative predictions depending on whether the single test lies above or below the median of a population of replicate tests if they were available. This leads to

$$\mathrm{E_{mf}} = \ln\frac{\mathrm{P}}{\mathrm{M}} = \mathrm{Laplace(bias, unc)}, \quad \textbf{0-79}$$

where bias equals the discrepancy measure based on the single test, bias = -0.00981, and the unc = 0.058197 is inferred from a database of relevant replicates tests and assumed applicable to the baseline seat design.



Pass/fail gate
Estimate bias

System Architecture

Single sled test
FAA-Hybrid III ATD
2" CF42 (AC) cushion
14G Environment
QOI: Maximum lumbar load

Pass/fail gates
Insight into trend bias
and range of applicability

Transition Tier

Database of 3 sled tests
FAA-Hybrid III ATDs
No cushions
14G Environments
QOI: Maximum lumbar load

Database of 4 pre-sled tests
FAA-Hybrid III ATDs
2" and 4" CF42 (AC) cushions
Environment NA
QOI: Initial static compression

Pass/fail gates

Physics Taxonomy

Database of 9 tests
CF42 (AC) constitutive model

FAA-Hybrid III constitutive model
Mass distribution
Dynamic response parameters

Figure 4.21: Validation hierarchy for the demonstration

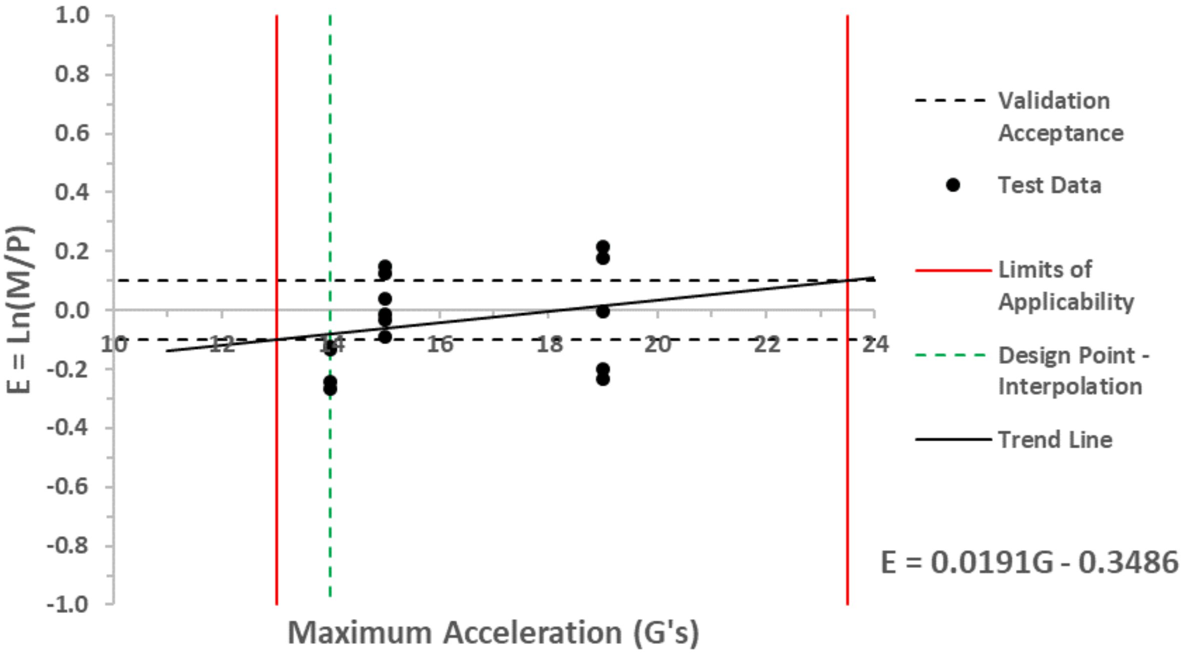


Figure 4.22: Validation of lumbar load model for seats without cushions – Hybrid II

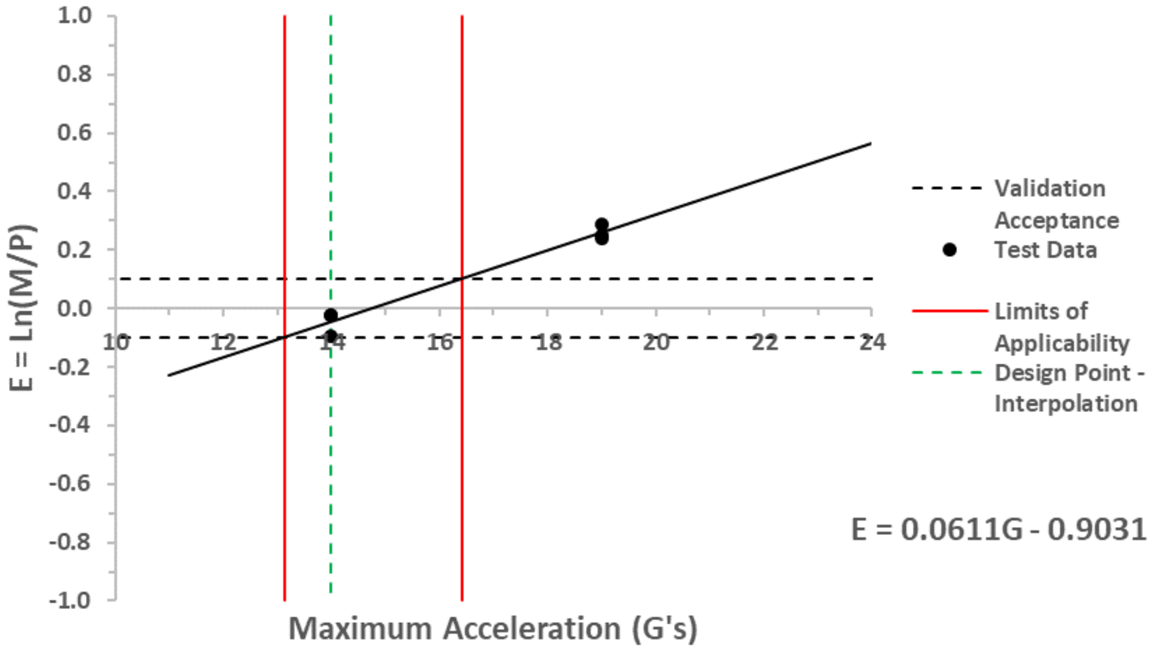


Figure 4.23: Validation of lumbar load model for seats without cushions - FAA-Hybrid III

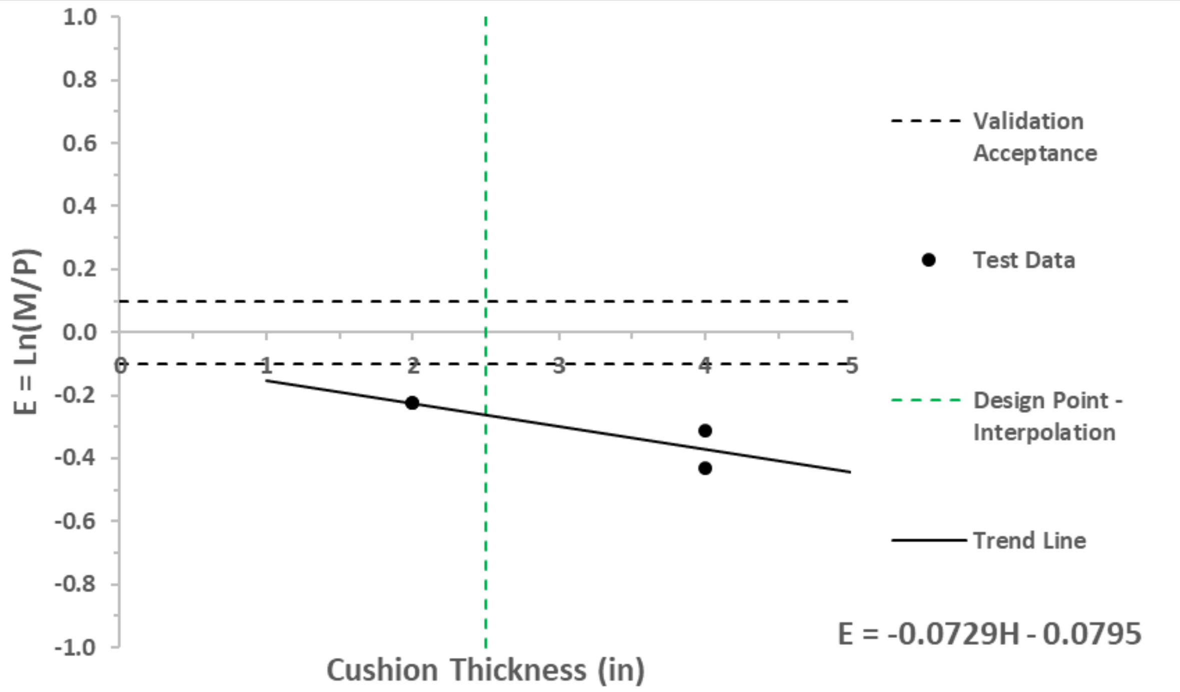


Figure 4.24: Validation of model for initial static compression

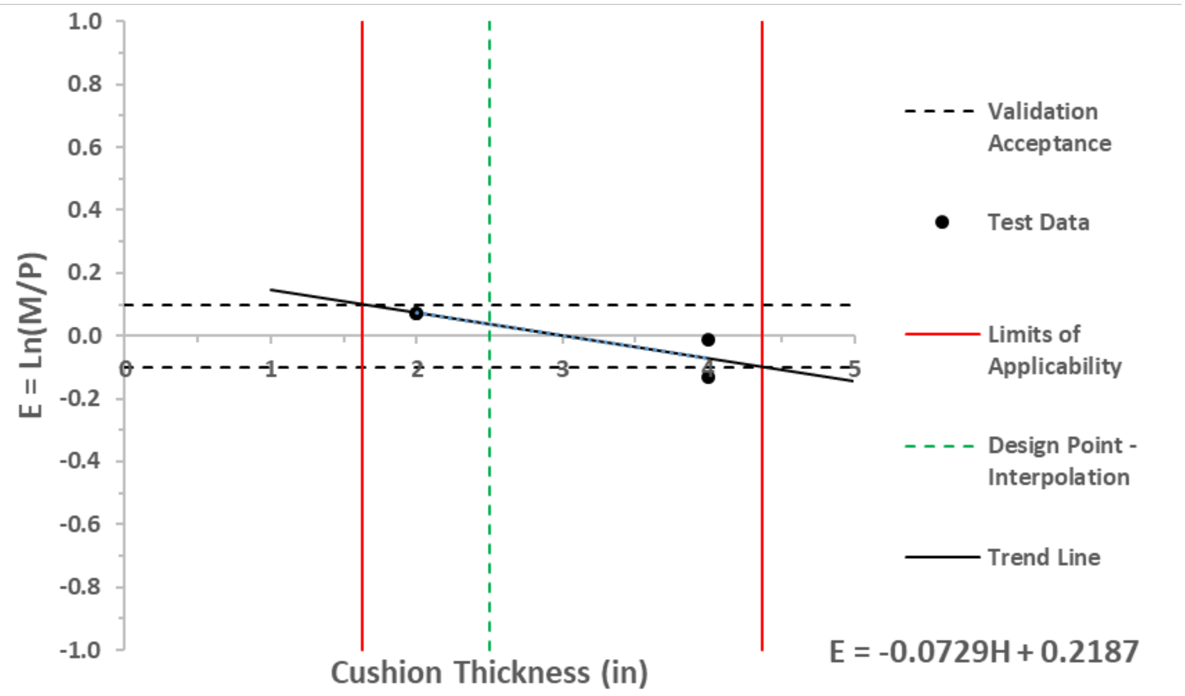


Figure 4.25: Assessment of the calibrated model for initial static compression

4.5 Integrate risk

This step's product is quantifying the required regulatory metrics and identifying dominant sources of uncertainty. Predicted results for the certification design are bias-corrected (positive or negative) for known sources of errors and their uncertainties. Simulation solution errors and their uncertainties are assessed specifically for predictions of the certification design. Model form errors and uncertainties are quantified in Section 0. An acceptable lumbar load threshold of 1500 $\mathrm{lb_f}$ is prescribed without uncertainty in 14 CFR Part 25.562.

4.5.1 Quantify consequences of "loads"

14 CFR Part 25.562 prescribes without uncertainty an acceptable lumbar load threshold of 1500 $\mathrm{lb_f}$. The technical basis is briefly reviewed here for completeness and perspective.

Figure 4.26 reproduces two key figures from (DeWeese et al., 2021). The top figure shows lumbar injury rates as a function of the dynamic response index (DRI). A straight line on this graphic indicates a normal distribution. The data are specific to the ejection of military pilots from several aircraft types in the 1960s.

Tests were performed specifically for the purpose of mapping DRI to lumbar loads; see the bottom graphic in
Figure 4.26. The mapping is specific to a 170 $\mathrm{lb_f}$ Hybrid II or FAA-Hybrid III ATD, seated upright, and conditional on emergency landing conditions prescribed by 14 CFR Part 25.562.

The mapping of lumbar injury rates to lumbar loads is given in Table 1 of (DeWeese et al., 2021) and reproduced here as Table 4.18. The regulatory threshold of 1500 $\mathrm{lb_f}$ corresponds to a DRI=19 and an injury rate R=0.09. The injury rate (R) can be represented by a normal distribution,

$$\mathrm{R} = \mathrm{Normal}(1718.91{,}166.05)\,, \qquad \textbf{0-80}$$

fit to values in Table 1 from (DeWeese et al., 2021). 14 CFR Part 25.562 prescribes a derived regulatory threshold of 1500 $\mathrm{lb_f}$ corresponding to a 9% injury rate.
The injury rate is a frequency statement about the population of military pilots ejecting from aircraft and should not be applied to the population of airline passengers without additional assessment. The pilots were fit males and 28 years old on average; consequently, the population of military pilots is not representative of the population of airline passengers. The mapping of DRI to lumbar load also should not be applied to the population of airline passengers. The mapping is specific to a 170 $\mathrm{lb_f}$ Hybrid II or FAA-Hybrid III ATD, seated upright, and conditional on emergency landing conditions.

Acknowledging these limitations, 14 CFR Part 25.562 prescribes without uncertainty,

1. Risk-informed lumbar load limit of 1500 $\mathrm{lb_f}$, conditional on
2. A 170 $\mathrm{lb_f}$ Hybrid II or FAA-Hybrid III ATD, seated upright, subject to the
3. Environments shown in Table 4.2.

Excluding seat design, which is specified by the applicant, items two and three specify the reality of interest that forms the basis for testing and simulation.

These requirements were implemented in 14 CFR Part 25 by the FAA in 1988. The adequacy is judged by operational experience. (Poland et al., 2016) showed that spinal injuries were low and about what should be expected for two crashes: 15.5% for the Asiana crash in 2013 and 7.5% for the Turkish Airlines crash in 2009. (NTSB, 2020) reviewed a wide range of crashes that occurred between 1983 and 2017. Among serious accidents that were determined to be survivable, 80.8% of occupants (passengers and flight crew) survived. Of the 19.2% fatalities, two-thirds were impact-related, and the remainder were related to fire, smoke, and other causes. When low operational injury rates are combined with the low occurrence of commercial aviation crashes, there is little benefit to implementing more rigorous regulatory requirements.

Table 4.18: Mapping of lumbar injury rates to lumbar loads

Table 1 from (DeWeese et al., 2021)		
Injury Rate (IR)	**DRI Operational Trendline**	**Lumbar Load (lb_f)**
0.01	16	1330
0.05	18	1450
0.09	19	1500
0.20	20	1580
0.40	22	1670
0.50	23	1710

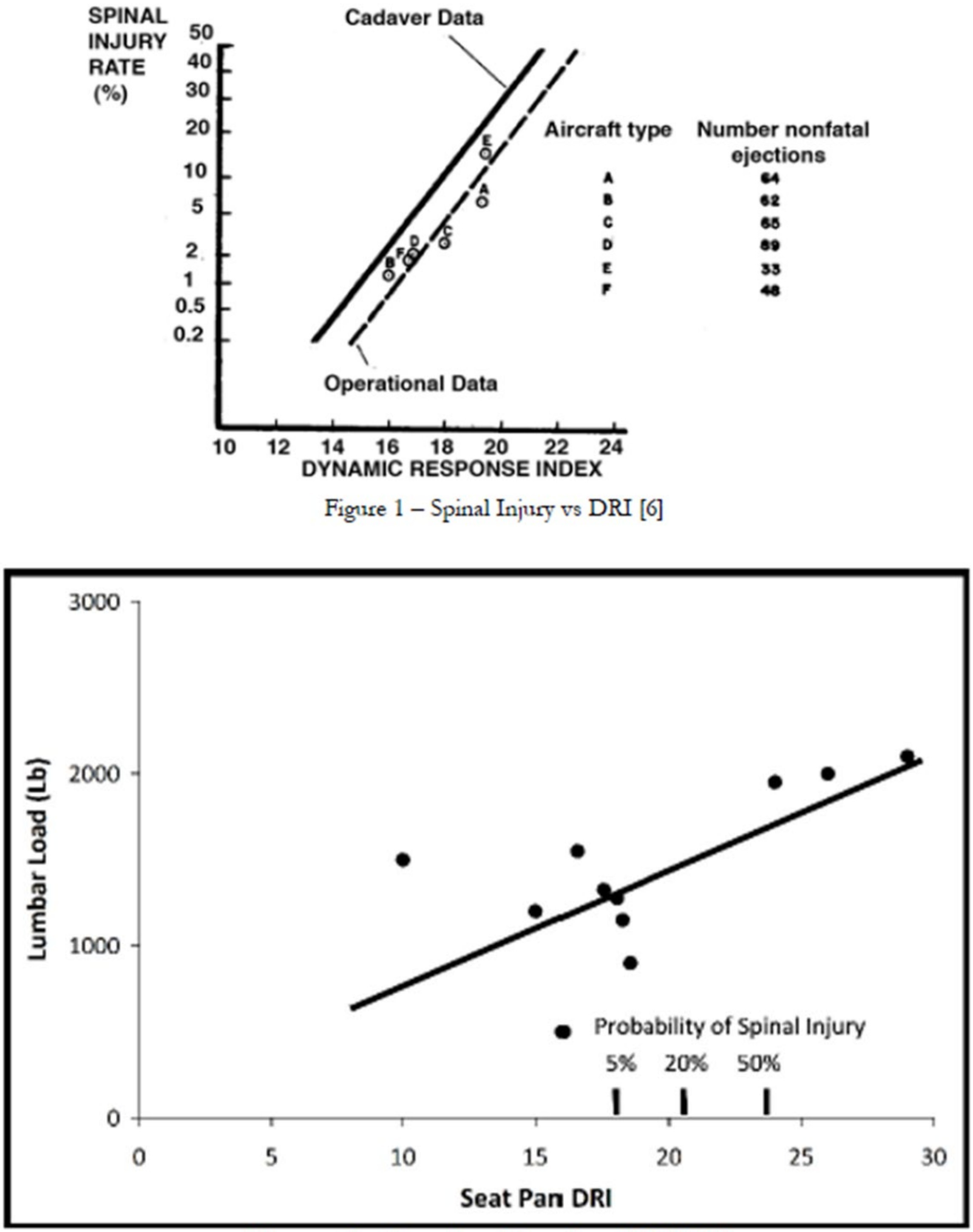


Figure 1 – Spinal Injury vs DRI [6]

Figure 2 - Lumbar Load vs DRI [8]

Figure 4.26: Technical basis regulated lumbar injury threshold (DeWeese et al., 2021)

4.5.2 Predict loads for untested design

The prediction of lumbar loads for the untested nearby design seeking CbA must bias-correct for two potential sources of error and uncertainties. In general, simulations at the design point should be bias-corrected for simulation solution errors and uncertainties. Eleven potential sources of simulation solution errors and uncertainties were assessed to be either negligible or intangible at the design point (Section 0.

Simulations at the design point, S(2.5”), should also be bias-corrected for model form errors and uncertainties. The predicted lumbar load at the design point is given by

$$\mathrm{L}(2.5") = \mathrm{S}(2.5")\mathrm{e}^{\mathrm{E_{mf}}}, \qquad \textbf{0-81}$$

where

$$\mathrm{E_{mf}} = \mathrm{Laplace}(\mathrm{bias}, \mathrm{unc}), \qquad \textbf{0-82}$$

and where the bias = -0.00981 is estimated by validation with the single test of the baseline seat design, and unc = 0.058197 is the model form uncertainty inferred from a relevant database of 101 tests. Although assessed for the baseline seat design, E_{mf} is assumed applicable for the nearby design seeking CbA.

4.5.3 Quantify decision metrics and perform sensitivity analysis

Figure1.1 provides the information necessary to integrate the risk. The green line is the FAA’s acceptance threshold for lumbar loads, prescribed without uncertainty. The red line is the predicted lumbar load for the nearby design (Equation 0-82), which is bias-corrected and accounts for uncertainties. The blue line is the median load for the distribution, and the black line is the 95th percentile of the load distribution. The decision metrics are quantified in Table 04.19. The FoS based on L_{50} is the best estimate, and the FoS based on L_{95} is the Best Estimate Plus Uncertainty, i.e., BEPU.

Table 04.19: Quantified decision metrics for the nearby design seeking CbA

Lumbar Load (lb_f)		Factor of Safety: FoS		
L_{50}	1102	L_{req}/L_{50}	1.36	BE
L_{95}	1211	L_{req}/L_{95}	1.24	BEPU

The goal of sensitivity analysis is to identify the dominant contributors to uncertainty to reduce uncertainty if needed and inform the regulatory decision efficiently. In general, there are three broad sources of uncertainty in the evaluation of the FoS: simulation solution uncertainties, model form uncertainties, and uncertainty in the lumbar injury criteria. The relative contribution of each to the total uncertainty would be an essential product of sensitivity analysis. They would steer potential follow-on efforts to where they would have the most significant impact. Each of the three primary sources of uncertainty will have multiple contributors, e.g., there are eleven potential sources of simulation solution errors. Testing uncertainties and simulation solution uncertainties for validation contribute to model form uncertainty. Here, the lumbar injury criteria are prescribed in 14 CFR Part 25.562 without uncertainty.

Sensitivity analysis for the demonstration is degenerate. Only model-form uncertainties exist, and they are dominated by precision uncertainties. Simulation solution uncertainties for the nearby design were assessed to be negligible. Uncertainty in the lumbar injury criteria was regulated away.

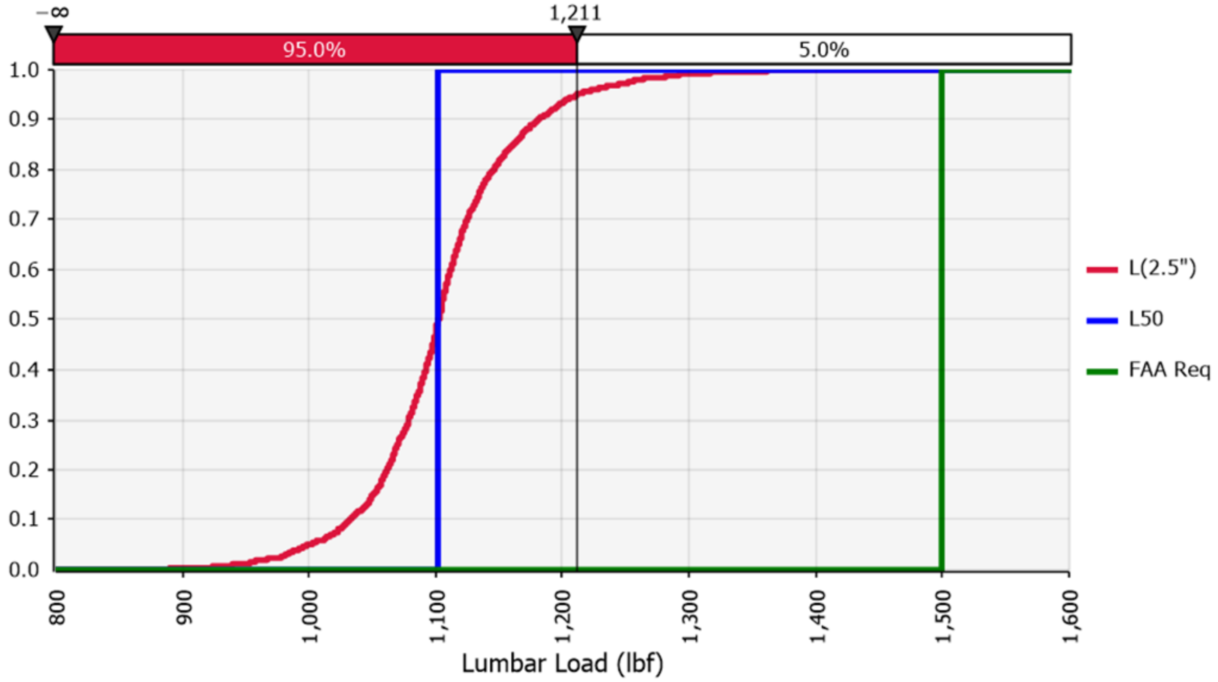


Figure 4.27: Risk integration for design seeking CbA (single test validation)

4.6 Make regulatory decision

The product of this element is a risk-informed decision to accept or reject a proposed design based on CbA. Quantitative inputs to risk-informed decisions are the decision metrics (e.g., factor of safety, FoS) and sensitivity of results to uncertainties. Other subjective factors that could inform the regulatory decision are:

1. Environmental factors of the regulatory agency: congressional oversight and mandates, risk tolerance, the complexity of the decision, and experience with this type of decision.
2. Corroborating evidence: Is there a balance between testing and computational simulation? Does the process complement the FAA's current approach to CbT and CbA?
3. Credibility of the assessments: evidence of completeness and correctness (VVUQ), communicated in a forthright and understandable manner, and documented for the record with sufficient detail that test results and simulation results can be recreated.
4. Findings of regulatory review and independent peer review as appropriate.

4.6.1 Regulatory decision when validation is based on a single test

Quantitative inputs

Decision metrics for the design seeking CbA were computed in Section 0 and repeated in Table 4.20. Regardless of best estimate or intended high confidence, the FoS's are greater than 1.0, and consequently color-coded green. Uncertainties are dominated by model form uncertainty, with precision errors in testing as the only contributor. Precision errors are dominated by the positioning of the ATD in the seat. Since documented processes exist, it is unlikely that this source of uncertainty can be significantly reduced.

Table 4.20: Quantified decision metrics based on validation with a single test

Lumbar Load (lb_f)		Factor of Safety: FoS		
L_{50}	1102	L_{req}/L_{50}	1.36	BE
L_{95}	1211	L_{req}/L_{95}	1.24	BEPU

Qualitative inputs

Corroborating evidence

The process proposed and demonstrated here is a logical extension to FAA's current approach to CbT based on a single test of the baseline seat design. CbA leverages the single test of the baseline seat design for model validation, and testing uncertainty is inferred from a relevant database.

Credibility of assessments

Eleven sources of simulation solution errors are demonstrably negligible (or intangible) for the baseline seat design and the nearby seat design. An acceptance test suite was developed that includes regression tests with known deterministic solutions, a verification test with a known analytic solution that tests the numerical scheme for order of accuracy, and a sustainability test to ensure solutions are reproducible in the future. Discretization errors are rendered negligible, with sufficient refinement for both validation and certification simulations. Monte Carlo solution uncertainties are negligible with an informed choice for the number of iterations and confirmed using bootstrap methods.

Models were successfully validated (or calibrated) against a hierarchy of relevant tests. Calibration of the CF42 (AC) constitutive model spans the full range of expected compression and compression rates observed in the demonstration. The lumbar load model was validated and accepted for seats without cushions. Model calibration was required for initial static compression before acceptance was possible. The full system model was validated and accepted against a single test of the baseline seat design. Trend bias and limits of applicability are quantified for the lower tiers of the validation hierarchy but not at the system level. System-level simulations for the certification design are corrected for estimated model bias (positive or negative). Uncertainties are inferred from a database of 101 related sled tests.

Summary of key concerns and intangibles

There are three technical concerns:

1. The inability to distinguish model bias from test precision uncertainty when validating to a single system-level test can lead to conservative or non-conservative predictions of the certification design.
2. The lack of evidence that the model can adequately predict the sensitivity of lumbar loads to cushion thickness undermines confidence in extrapolation to the certification design.
3. The assumption of lower and upper torso rigidity limits the applicability of the model.

Consideration of four intangibles identified during assessment should inform the regulatory decision. Intangibles are acknowledged but unquantified (or unquantifiable) sources of potential error and uncertainty. The four intangibles are:

1. Analyst-to-analyst (A2A) uncertainty is managed by applying a common modeling approach to get the correct answer for the right reason; however, there is no expectation that A2A uncertainty can ever be eliminated.
2. User errors and platform inconsistencies could not be quantified as contributors to the simulation solution errors/uncertainties assessment.
3. Sampling uncertainties and accuracy errors/uncertainties could not be quantified as contributors to the assessment of testing errors/uncertainties.
4. Seat-to-seat variability in the fielded seat population cannot be predicted with simulation and can only be estimated by testing a posteriori.

Risk-informed regulatory decision

The regulatory decision is risk-informed. Quantitative inputs in the form of factors of safety (with and without uncertainties) are acceptable, but the first two technical concerns require further attention; consequently, MPilchConsulting conditionally ***rejects*** certification of the nearby design. The resolution of the first two technical concerns is the condition for certification.

4.6.2 Regulatory decision when validation is based on an ensemble of tests

To address the technical concerns with validation against a single system-level test, the system architecture tier of the validation hierarchy is supplemented with additional system-level tests (Figure 4.28).

The sensitivity of lumbar loads for CF42 (AC) cushions to thickness and environments is shown in Figure 4.29, where the data are taken from (Taylor, DeWeese, et al., 2017) and (Pellettiere et al., 2019). The database has five 14G tests and seven 19G tests. Observed lumbar loads increase with cushion thickness and environment. The cushion thickness for the nearby design interpolates on the database. Lumbar loads for 19G environments are 700 lb_f greater than at 14G, depending on the cushion thickness.

The ability of the model to predict lumbar load sensitivity to cushion thickness and environment is assessed in Figure 4.30. The model captures the observed sensitivity to cushion thickness at the design point

$$|E_{50}(\mathrm{CF42\ (AC)}, 14\mathrm{G}, 2.5")| = 0.0446\,, \qquad 0\text{-}83$$

which is within the validation acceptance bounds; consequently, the model is accepted. Although there is no significant dependency of discrepancy measures, E, with cushion thickness, extrapolation beyond 4.0” might require additional evaluation. Based on the 14G data, the model form error can be represented by

$$E_{\mathrm{mf}} = \mathrm{Normal(bias, unc)} \qquad 0\text{-}84$$

where bias = -0.04464 and unc = 0.03607. Here, bias is an estimate without assumption based on model assessment against multiple tests and resolves the ambiguity associated with single-test validation.

The model is limited to 14G environments. Figure 4.30 shows significant discrepancies for 19G environments when seats have no cushions. This was commented upon in Section 0 and is likely related to the assumption of lower and upper torso rigidity in the conceptual model.

Equations 0-81 and 0-84 are combined to predict lumbar loads for the certification design when the model form error is based on ensemble validation. The results are shown in Figure 4.31, and Table 4.21 summarizes the computed decision metrics. Regardless of the best estimate or intended high confidence, the FoSs are greater than 1.0 and consequently color-coded green. The FoSs are larger based on ensemble validation compared to the previous FoSs based on single-test validation.

Table 4.21: Quantified decision metrics based on ensemble validation

Lumbar Load (lb_f)		Factor of Safety: FoS		
L_{50}	1064	L_{req}/L_{50}	1.41	BE
L_{95}	1125	L_{req}/L_{95}	1.33	BEPU

There are two contributors to the larger FoSs when based on ensemble validation. First, L_{50} = 1064 lbf is smaller than L_{50} = 1102 lbf based on single-test validation because the bias correction is larger (bias = -0.04464) when based on ensemble validation when compared to the bias correction (bias = -0.00981) when based on single-test validation. In both cases, the model over-predicted the observed lumbar loads in testing, but the over-prediction was larger with ensemble validation when compared to single-test validation. Recall from Figure1.1 that current FAA guidance intentionally disallows bias correction when the model overpredicts the validation benchmark, but it is allowed in the proposed process. The FAA intended to introduce conservatism at this step, but this is not in the spirit of BEPU and is not always conservative.

The assessed values for precision uncertainty are the second reason for the larger FoSs when based on ensemble validation. The precision uncertainty from ensemble validation (unc = 0.0361) is computed from a limited database of five tests from only two references and is smaller than the precision uncertainty (unc = 0.0582) when derived from 101 datapoints associated with 42 replicate test series.

Precision uncertainty may not be fully representative in small datasets typically associated with ensemble validation. A hybrid approach would compute trend bias from ensemble data and infer precision uncertainty from the much larger database of related replicate testing, but this was not done here.

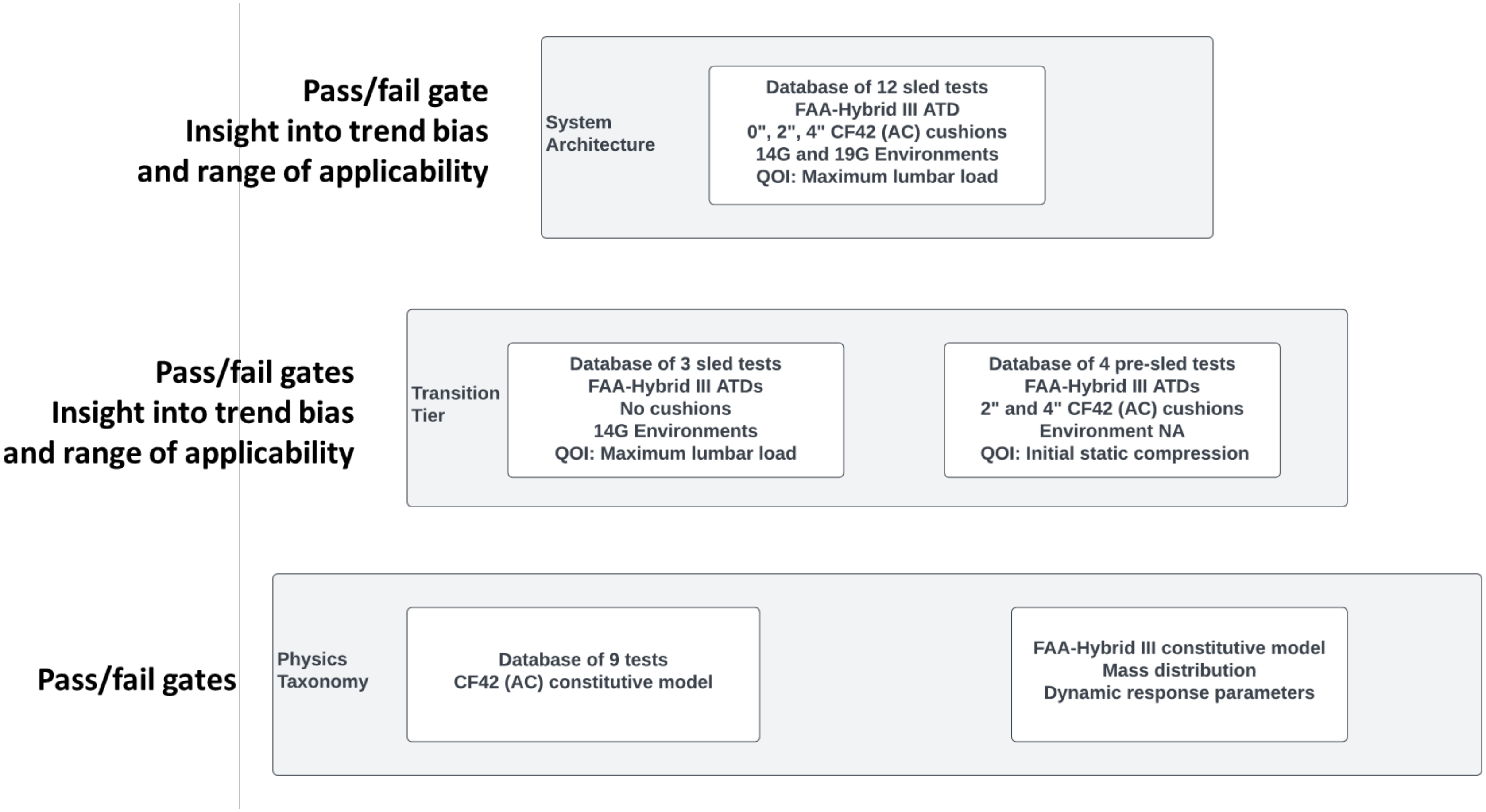
Uncertainties are dominated by model form uncertainty with precision errors in testing the only contributor. Precision errors seem dominated by positioning of the ATD in the seat. Since documented processes exist, it is unlikely that this source of uncertainty can be significantly reduced.

Reassessment of risk-informed regulatory decision

MPilchConsulting now recommends certification acceptance of the nearby design.

Re-evaluated decision metrics for the design seeking CbA are shown in Table 4.21. Regardless of best estimate or intended high confidence, the FoSs are greater than 1.0 and consequently color-coded green. Uncertainties are dominated by model form uncertainty with precision errors in testing the only contributor. The magnitude of precision errors may not be fully representative in small datasets. Precision errors are dominated by the positioning of the ATD in the seat. Since documented processes exist, it seems unlikely that this source of uncertainty can be significantly reduced.

The first two technical concerns were successfully addressed by assessing the model against an ensemble of relevant system-level tests. Model bias is now distinguished from test precision, and evidence now exists that the model can predict the sensitivity of lumbar loads to cushion thickness, lending credibility to the model. The model is limited to 14G environments until sensitivity to environment can be resolved, which would expand the range of applicability and increase confidence in the model. Qualitative supporting evidence remains unchanged, and so does the list of intangibles.



Pass/fail gate
Insight into trend bias
and range of applicability

System Architecture

Database of 12 sled tests
FAA-Hybrid III ATD
0", 2", 4" CF42 (AC) cushions
14G and 19G Environments
QOI: Maximum lumbar load

Pass/fail gates
Insight into trend bias
and range of applicability

Transition Tier

Database of 3 sled tests
FAA-Hybrid III ATDs
No cushions
14G Environments
QOI: Maximum lumbar load

Database of 4 pre-sled tests
FAA-Hybrid III ATDs
2" and 4" CF42 (AC) cushions
Environment NA
QOI: Initial static compression

Pass/fail gates

Physics Taxonomy

Database of 9 tests
CF42 (AC) constitutive model

FAA-Hybrid III constitutive model
Mass distribution
Dynamic response parameters

Figure 4.28: Valuation hierarchy supplemented with existing system level- tests

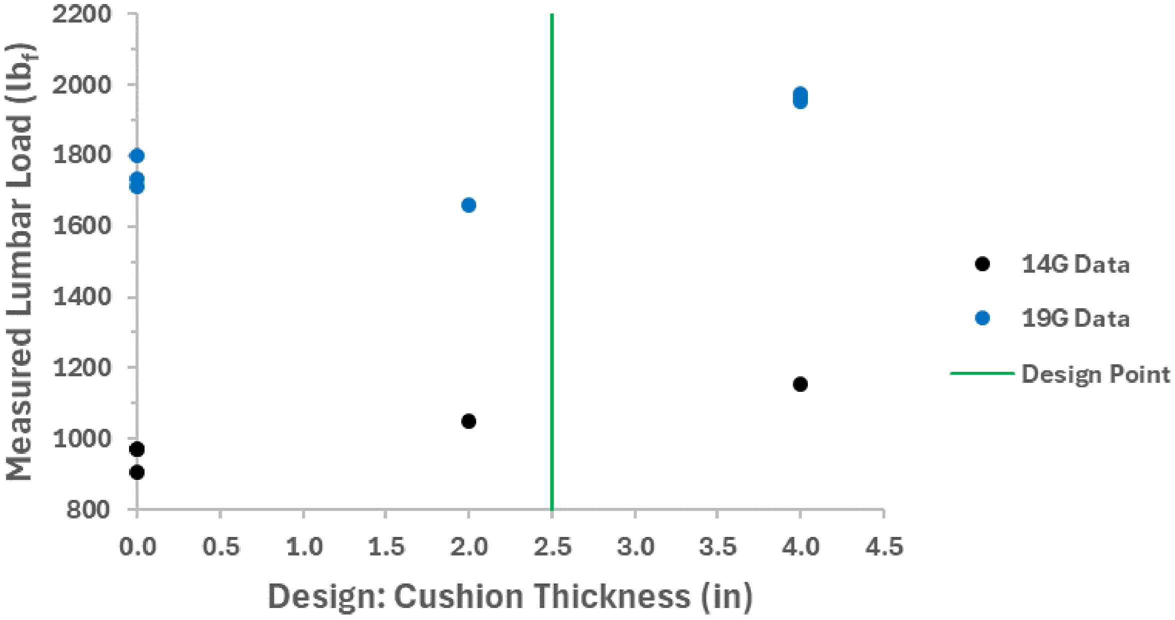


Figure 4.29: Sensitivity of lumbar loads to cushion thickness and environment

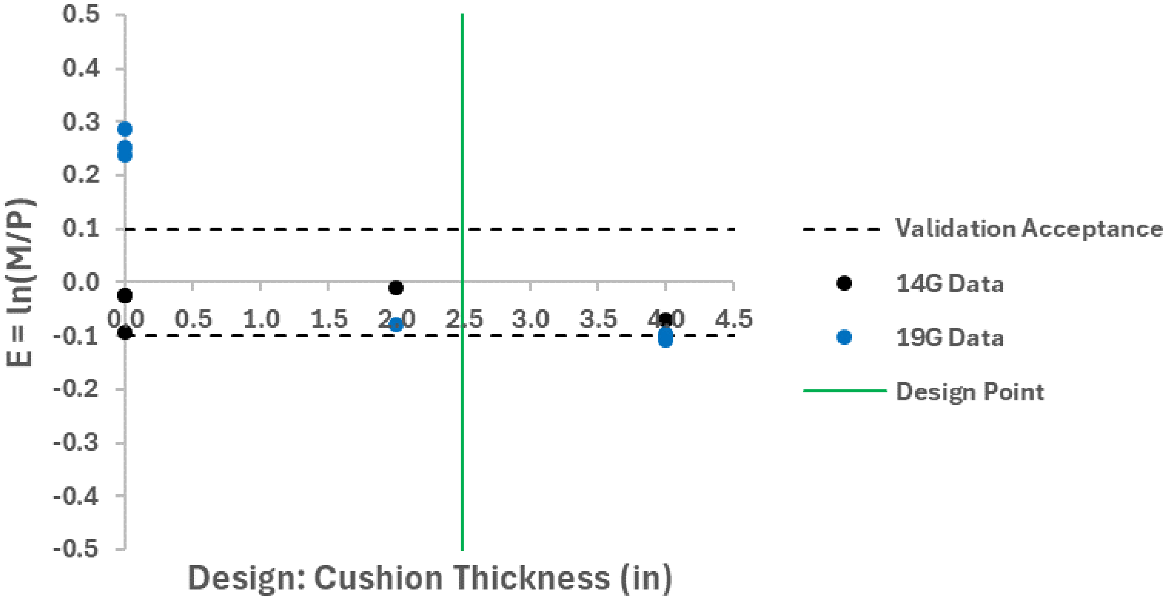


Figure 4.30: Model validation against an ensemble of tests

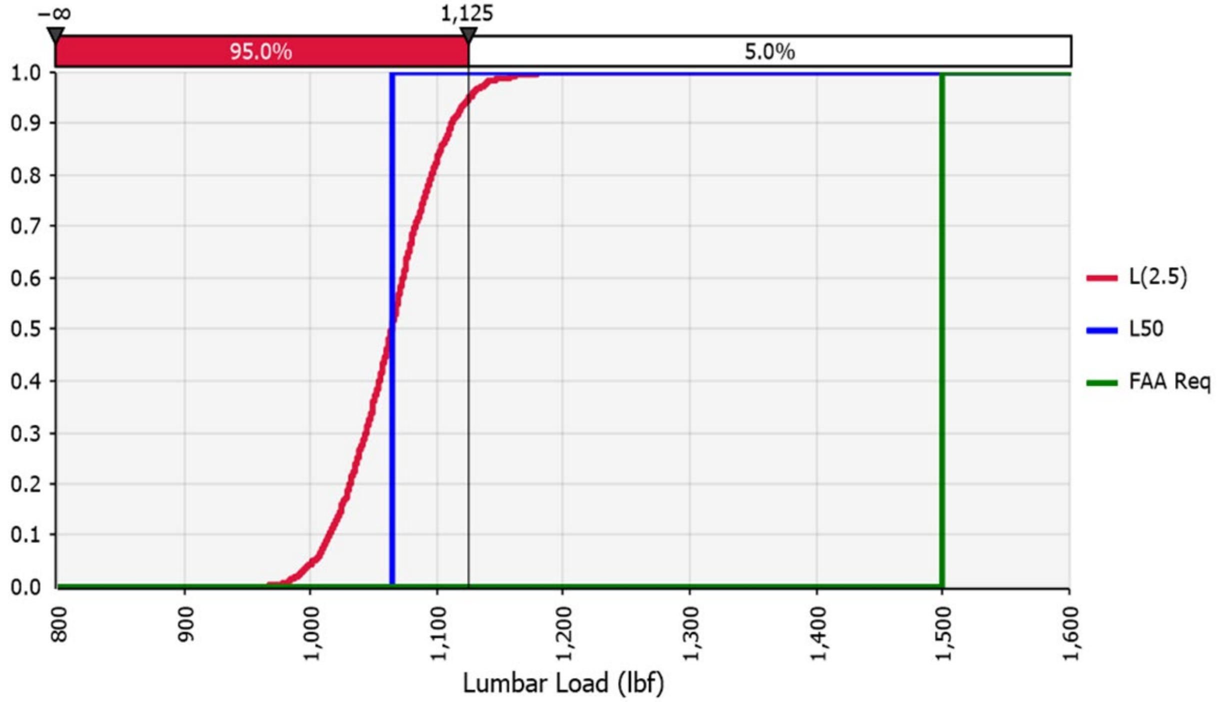


Figure 4.31: Risk integration for design seeking CbA (ensemble validation)

5. Comments and Recommendation on the Implementation of BEPU

Embrace BEPU, i.e., an unbiased assessment with an objective accounting of all dominant sources of uncertainty, but be cautious of putting too much confidence in distribution tails. Numerically *accurate* estimates of the 95 percentile loads are possible, but support (data or judgment) for the underlying input distributions is generally weak, especially near the tails. Small datasets are pervasive in this demonstration and unavoidable in general. You will never have sufficient test data to accurately resolve the 95 percentiles of a distribution and rarely enough test data to resolve the 50 percentiles accurately. Fitting an assumed parametric distribution with a small data sample and extrapolating into the tails is risky, and this was done twice in the demonstration. Support for subjective (belief) distributions is weak, especially when defining tails. This is most likely to occur when assessing simulation solution errors and uncertainties associated with explicit dynamics codes.

There are four regulatory options for dealing with uncertainties that balance risk tolerance and the maturity of assessment capabilities.

1. Accept the uncertainty without formal quantification. The decision metric is FoS = $L_{req}/L_{nominal}$, where $L_{nominal}$ is the assessed lumbar load from a single test or simulation. Test simulation results must be biased-corrected for known sources of error for validation and prediction. Formal uncertainty quantification is not expected, but evidence that best practices are followed might be expected. This option aligns with current FAA guidance for CbT (AC No. 25.562-1B).
2. Quantify uncertainty, regulate to the median, and learn from the uncertainty. The decision metric is FoS = L_{req}/L_{50}, where L_{50} is the 50 percent of a formally quantified uncertainty distribution of lumbar loads. This is evolutionary to Option 1 and is the best estimate (BE). There is no expectation of $L_{50} \sim L_{nominal}$ so that the decision metric could differ. The formalism of uncertainty quantification lends increased credibility to assessing decision metrics, but quantified uncertainties do not further restrict the decision metric. Sensitivity analysis allows for identifying and reducing, if necessary, dominant contributors to uncertainties in predicted lumbar loads.
3. Quantify uncertainty and regulate with the intent of high confidence. The decision metric is FoS = L_{req}/L_{95}, where L_{95} is the 95 percentile of a formally quantified uncertainty distribution of lumbar loads. This is revolutionary to Option 1 and is the Best Estimate Plus Uncertainty (BEPU). Explicit consideration of uncertainties in decision metrics will always be more restrictive than current practice for CbT. Still, it aligns with the spirit of CbA, AC No.20-146A, but the processes in this report add more formalism to the identification and management of uncertainties. Sensitivity analysis allows for identifying and reducing, if necessary, dominant contributors to uncertainties in predicted lumbar loads.
4. Apply conservative acceptance criteria to any of the first three options. This means that the FoS = 1.5 or 2.0 or wherever judgment leads instead of FoS = 1.0. This is the only place that conservatism is encouraged in the spirit of BEPU. This adds robustness to decisions already made when new information becomes available, i.e., surprise when formerly unknown/unknowns become revealed. This adds robustness to intangibles, i.e., acknowledged sources of errors and uncertainties that are not quantifiable or quantified. This also safeguards against putting too much confidence in the 95th percentile of computed distributions.

The dominant sources of uncertainty are trivial to identify in this demonstration. Model form uncertainty dominated, with precision uncertainties being the only contributor because that was the only uncertainty left. Simulation solution errors were assessed as negligible for validation and certification simulations, but this would not be the case when explicit dynamics codes are used.

Be intentional about the potential disparity in rigor between CbA and CbT, but recognize that there are parallels in quality standards for CbA and CbT. For instance, CbT has well-accepted processes and procedures for managing testing errors and uncertainties, including routine instrumentation calibration. Likewise, CbA has processes and procedures to manage simulation solution errors/uncertainties.

The concepts of BEPU are equally applicable to CbT as they are to CbA. Currently, CbA is held accountable for test uncertainties, while CbT is not. This is not in the spirit of BEPU if it were applied to CbT. The application of BEPU to single-test certification is straightforward. There is only one test, so the impact of testing precision uncertainties cannot be assessed directly. Still, it can be inferred from a database of related replicate tests conducted over a wide range of seat design and test environments. Appendix E.2 shows that if many replicate tests were available, then the distribution of measured lumbar loads would be given by

$$\mathrm{M} = L_{50} e^{Laplace(0,unc)} \qquad \textbf{0-1}$$

where unc = 0.058197. The best we can do is equate L_{50} with the measured lumbar load for the baseline seat design, L = 1048 lbf.

Figure 5.1 shows risk integration for applying BEPU to CbT, which parallels a similar assessment for CbA with testing uncertainties. The red curve shows the best estimate distribution of lumbar loads as if replicates were available. The quantified decision metrics are shown in Table 5.1 **Error! Reference source not found.**. The best estimate factor of safety remains unchanged, but as expected, the FoS is more restrictive if the intent is to regulate with high confidence.

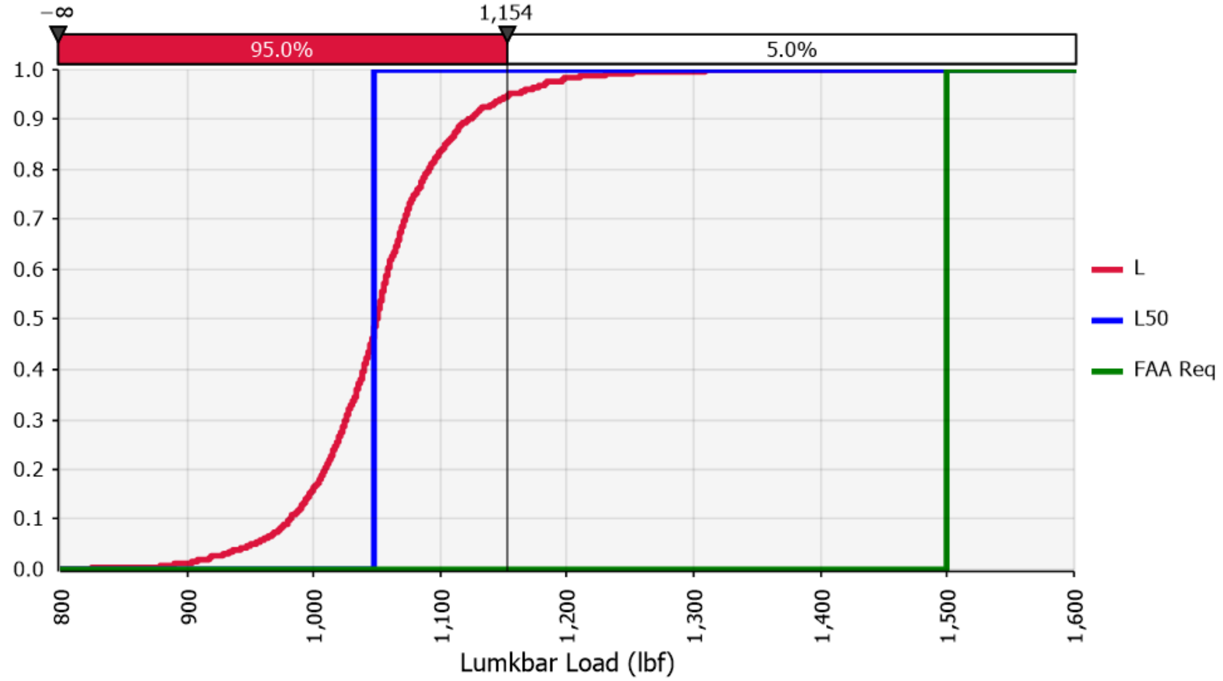


Figure 5.1: Risk integration for applying BEPU to CbT

Table 5.1: Quantified decision metrics for application of BEPU to CbT

Lumbar Load (lb_f)		**Factor of Safety: FoS**		
L_{50}	1048	L_{req}/L_{50}	1.43	BE
L_{95}	1154	L_{req}/L_{95}	1.30	BEPU

CbA necessarily involves interpolation or extrapolation, and the (NaRC, 2012) concluded that both can be risky, but extrapolation arguably involves greater risk. CbT has its own risk but does not share the risk of interpolation or extrapolation. There are no technical barriers to testing a 2.5” cushion design. CbA has the unique burden of providing evidence that it can predict a wide range of designs and environments using a common model approach. There are two approaches:

1. Establish and document a history of successful blind predictions, but this might be difficult to implement in a regulatory environment.
2. Validate the model against a predefined database of tests suitable for assessing sensitivity to design variations and environments. If the applicant has been applying consistent simulation governance for a period, this may be as simple as selecting and organizing examples from their own experience. The case studies offered by the applicant could be supplemented with research tests documented in Appendix E.1.

Benefits of the second option are twofold. First, the limits of applicability are better defined, providing a basis for relaxing what “nearby” means, thus allowing for greater interpolation and extrapolation without additional testing. Second, the model form error can be quantified in terms of bias and uncertainty.

The number of computationally expensive simulations for the proposed process is reasonable and consistent with the current process. Three or more computationally expensive simulations are required to assess simulation solution errors for both the baseline seat design and the nearby design seeking certification. There is already the expectation that simulation solution errors will be assessed. The process proposed here requires one full system simulation for each validation test and one full system simulation of the nearby design. This is consistent with current expectations.

References

Abanto, J., Pelletier, D., Garon, A., Trepanier, J.-Y., & Reggio, M. (2005). *Verification of Some Commercial CFD Codes on Atypical CFD Problems* 43rd AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV.

Adams, A., Lankarani, H. M., & Safai, N. M. (2003). *Aircraft Seat Cushion Performance Evaluation and Replacement Implementaion* Proceedings of the 2003 American Sociaty for Engineering Education Annual Conference & Exposition,

AIAA. (1998). *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*.

AIAA. (2024). Standard for Code Verification in Computational Fluid Dynamics. In (Vol. S-141-2024).

Ang, A. H. S., & Tang, W. H. (1975). *Probability Concepts in Engineering Planning and Design: Vol. I Basic Principles* (1st ed.). John Wiley.

ASME. (2006). *Guide for Verification and Validation in Computational Solid Mechanics* (ASME Standard V&V 10-2006).

ASME. (2012). *An Illustration of the Concepts of Verification and Validation in Computational Solid Mechanics* (ASME Standard V&V 10.1-2012).

ASME. (2018). Assessing Credibility of Computational Modeling Through Verification and Validation: Application to Medical Devices. In. New York, NY: American Society of Mechanical Engineers.

ASME. (2019). *Standard for Verification and Validation in Computational Solid Mechanics* (ASME Standard V&V 10-2019).

ASTM. (2010). Standard Test Methods for Flexible Cellular Materials - Slab, Bonded, and Molded Urethane Foams. In *D 3574-05*.

Bhonge, P., Deweese, R. L., Moorcroft, D., & Islam, R. (2019, Oct 28 - Oct 31, 2019). *Comparison of Dynamic Responses of 50th percentile Hybrid II and FAA Hybrid III Anthropometric Test Devices (ATD) during Aircraft Seat Tests* 9th Triennial International Aircraft Fired and Cabin Safety Research Conference, Atlantic City, NJ.

Blattnig, S. R., Luckring, J. M., Morrison, J. H., Sylvester, A. J., Tripathi, R. K., & Zang, T. A. (2013). NASA Standard for Models and Simulations: Philosophy and Requirements Overview. *Journal of Aircraft*, *50*(1), 20-28.

Boyack, B. E., Motta, A. T., Peddicord, K. L., Alexander, C. A., Deveney, R. C., Dunn, B. M., Fuketa, T., Higar, K. E., Hochreiter, L. E., Langenbuch, S., & etc. (2001). *Phenomenon Identification and Ranking Tables (PIRTs) for Rod Ejection Accidents in Pressurized Water Reactors Containing High Burnup Fuel*.

Breeding, R. J., Helton, J. C., Gorham, E. D., & Harper, F. T. (1992). Summary Description of the Methods used in the Probabilistic Risk Assessments for NUREG-1150. *Nuclear Engineering and Design*, *135*, 1-27.

Clay, R. L., Marburger, S. J., Shneider, M. S., & Trucano, T. G. (2007). *Modeling and Simulation Technology Readiness Levels* (SAND2007-0570).

Croop, B., & Lobo, H. (2009). *Selecting Material Models for the Simulation of Foams in LS-Dyna* 7th European LS-Dyna Conference, Salzburg Austria.

DeWeese, R., & Gowdy, V. (2002). *Human Factors Associated With the Certification of Airplane Passenger Seats: Seat Belt Adjustment and Release*. (DOT/FAA/AM-02/11).

DeWeese, R. L., Moorcroft, D. M., & Taylor, A. M. (2021). Lumbar Load Variability in Dynamic Testing of Transport Category Aircraft Seat Cushions [Tech Report]. https://rosap.ntl.bts.gov/view/dot/57228

Diegert, K., Klenke, S., Novotny, G., Paulsen, R., Pilch, M., & Trucano, T. (2007). *Toward a More Rigorous Application of Margins and Uncertainties within the Nuclear Weapons Life Cycle - A Sandia Perspective* (SAND2007-6219).

FAA. (2018). *Advisory Circular: Methodology for Dynamic Seat Certification by Analysis for Use in Parts 23, 25, 27, and 29 Airplanes and Rotorcraft*.

FDA. (2023). *Assessing the Credibility of Computational Modeling and Simulation in Medical Device Submissions: Guidance for Industry and Food and Drug Administration Staff*.

Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D. S., & Sentz, K. (2002). *Constructing Probability Boxes and Dempster-Shafer Structures* (SAND2002-4015).

Garrick, B. J. (2008). *Quantifying and Controlling Catastrophic Risks*. Academic Press.

Gowdy, V., Deweese, R. L., Beebe, M. S., Wade, B., Duncan, J., Kelly, R., & Blaker, J. L. (1999). A Lumbar Spine Modification to the Hybrid III ATD For Aircraft Seat Tests. *SAE transactions*, *108*, 367-379.

Harmon, S. Y., & Youngblood, S. M. (2005). A Proposed Model for Simulation Validation Process Maturity. *The Journal of Defense Modeling and Simulation*, *2*(4), 179-190.

Helton, J. C. (1994). Treatment of Uncertainty in Performance Assessments for Complex Systems. *Risk Analysis*, *14*(4), 483-511.

Helton, J. C. (2003). Mathematical and numerical approaches in performance assessment for radioactive waste disposal: dealing with uncertainty. In E. M. Scott (Ed.), *Modelling Radioactivity in the Environment* (pp. 353-390). Elsevier Science.

Helton, J. C. (2011). Quantification of Margins and Uncertainties: Conceptual and Computational Basis. *Reliability Engineering and System Safety*, *96*(9), 976-1013.

Helton, J. C., Johnson, J. D., & Oberkampf, W. L. (2004). *Probability of Loss of Assured Safety in Temperature Dependent Systems with Multiple Weak and Strong Links* (SAND2004-5216).

Helton, J. C., Johnson, J. D., & Sallaberry, C. J. (2011). Quantification of Margins and Uncertainties: Example Analyses from Reactor Safety and Radioactive Waste Disposal Involving the Separation of Aleatory and Epistemic Uncertainty. *Reliability Engineering and System Safety*, *96*(9), 1014-1033.

Helton, J. C., & Pilch, M. (2011). Guest Editorial: Quantification of Margins and Uncertainties. *Reliability Engineering and System Safety*, *96*(9), 959-964.

Helton, J. C., & Sallaberry, C. J. (2009). Computational Implementation of Sampling-Based Approaches to the Calculation of Expected Dose in Performance Assessments for the Proposed High-Level Radioactive Waste Repository at Yucca Mountain, Nevada. *Reliability Engineering and System Safety*, *94*, 699-721.

Hills, R. G., Pilch, M., Dowding, K. J., Red-Horse, J., Paez, T. L., Babuska, I., & Tempone, R. (2008). Validation Challenge Workshop. *Computer Methods in Applied Mechanics and Engineering*, *197*(29-32), 2375-2380.

Hills, R. G., Witkowski, W. R., Urbina, A., Rider, W. J., & Trucano, T. G. (2013). *Development of a Fourth Generation Predictive Capability Maturity Model* (SAND2013-8051).

Hooper, S. J., & Henderson, M. J. (2005). *Development and Validation of an Aircraft Seat Cushion Component Test - Volume 1*. (DOT/FAA/AR-05/5,I).

Jeong, K. Y., Cheon, S. S., & Munshi, M. B. (2012). A constitutive model for polyurethane foam with strain rate sensitivity. *Journal of Mechanical Science and Technology*, *26*(7), 2033-2038. https://doi.org/10.1007/s12206-012-0509-1

Johnson, G. R., & Cook, W. H. (1985). Fracture Characteristics of Three Metals Subjected to Various Strains, Strain Rates, Temperatures and Pressures. *Engineering Fracture Mechanics*, *21*, 31-48.

Kelly, J., Corradini, M., Budnitz, R., & Pilch, M. (2011). Perspectives on Advanced Simulation for Nuclear Reactor Safety Applications. *Nuclear Science and Engineering*, *168*(2), 128-137.

Lee, S. W., Chung, B. D., Bang, Y. S., & Bae, S. W. (2014). Analysis of Uncertainty Quantification Method by Comparing Monte-Carlo Method and Wilks' Formula. *Nuclear Engineering and Technology*, *46*, 481-488.

Luckring, J. M., Shaw, S., Oberkampf, W. L., & Graves, R. (2023). Development of a Model Validation Hierarchy for Modeling and Simulation. In J. H. Morrison, E. I. Walker, & M. Nikbay (Eds.), *Approaches to Model Validation*. NATO - Science and Technology Organization.

Manteufel, R. (2000). Evaluating the convergence of Latin Hypercube Sampling. In *41st Structures, Structural Dynamics, and Materials Conference and Exhibit*. https://doi.org/10.2514/6.2000-1636

Mousseau, V. A., & Williams, B. J. (2017). Uncertainty Quantification in a Regulatory Environment. In R. Ghanem, D. Higdon, & H. Owhadi (Eds.), *Handbook of Uncertainty Quantification* (pp. 1613-1648). Springer.

NaRC. (2009). *Evaluation of Quantification of Margins and Uncertainties: Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile* (ISBN-13: 978-0-309-12853-7). T. N. A. Press.

NaRC. (2012). *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification* (ISBN-13: 978-0-309-25634-6). T. N. A. Press.

Neilsen, M., Lu, W.-Y., Kraynik, A., Scherzinger, B., Hinnerichs, T., Jin, H., Bauer, S., & Hong, S. (2007). *Constitutive Models for Polyurethane Foams* TriLaboratory Engineering Conference, Albuquerque NM.

NRC. (1989). *Quantifying Reactor Safety Margins* (NUREG/CR-5249).

NRC. (1990). *Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants* (NUREG-1150).

NTSB. (2020). *Survivability of Accidents Involving Part 121 US Air Carrier Oprations: 2020 Update*. https://www.ntsb.gov/safety/data/Pages/Part121AccidentSurvivability.aspx

Oberkampf, W. L., Pilch, M., & Trucano, T. G. (2007). *Predictive Capability Maturity Model for Computational Modeling and Simulation* (SAND2007-5948).

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and Validation in Scientific Computing*. Cambridge University Press.

Olivares, G. (2013). *Hybrid II and Federal Aviation Administration Hybrid III Anthropomorphic Test Dummy Dynamic Evaluation Test Series*. (DOT/FAA/AR-11/24).

Olivares, G., Pellettiere, J., & Moorcroft, D. (2018). *Anthropomorphic Test Dummy Lumbar Load Variation*.

Pellettiere, J. A., Moorcroft, D. M., & Olivares, G. (2011). Anthropomorphic Test Dummy Lumbar Load Variation.

Pellettiere, J. M., Moorcroft, D. M., & Olivares, G. (2019, August 2019). *Anthropomorphic Test Dummy Lumbar Load Variations* International Technical Conference on the Enhanced Saftey of VCehicles, Washington DC.

Pilch, M. (2019). *Cautionary Tale When Using Computational Simulations to Support Regulatory Decisions* ASME Verification and Validation Symposium, Los Vegas, NV.

Pilch, M., Trucano, T. G., & Helton, J. C. (2011). Ideas Underlying the Quantification of Margins and Uncertainties. *Reliability Engineering and System Safety*, *96*(9), 965-975.

Pilch, M. M., & Allen, M. D. (1996). Closure of the direct containment heating issue for Zion. *Nuclear Engineering and Design*, *164*(1), 37-60. https://doi.org/https://doi.org/10.1016/0029-5493(96)01229-0

Pilch, M. M., Yan, H., & Theofanous, T. G. (1996). The probability of containment failure by direct containment heating in Zion. *Nuclear Engineering and Design*, *164*(1), 1-36. https://doi.org/https://doi.org/10.1016/0029-5493(96)01227-7

Poland, K., McKay, M. P., & Taylor, A. (2016). *Passenger Spinal Injuries in the 2013 Asiana and 2009 Turkish Airline Crashes* FAA Triennial Fire and Cabin Safety Conference,

Porter, N. W. (2019). Wilks' formula applied to computational tools: A practical discussion and verification. *Annals of Nuclear Energy*, *133*, 129-137. https://doi.org/https://doi.org/10.1016/j.anucene.2019.05.012

Porter, N. W., Salko, R. K., & Pilch, M. (2020a). Development and implementation of a CTF code verification suite. *Nuclear Engineering and Design*, *370*, 110879. https://doi.org/https://doi.org/10.1016/j.nucengdes.2020.110879

Porter, N. W., Salko, R. K., & Pilch, M. (2020b). *FY20 Improvements to CTF Code Verification and Unit Testing*. (CASL-U-2020-1938-000).

QMU. *Quantification of margins and uncertainties*. https://en.wikipedia.org/wiki/Quantification_of_margins_and_uncertainties#:~:text=QMU%20focuses%20on%20the%20identification,using%20computational%20modeling%20and%20simulation.

Roache, P. J. (2009). *Fundamentals of Verification and Validation*. Hermosa Publishers.

Roy, C. J. (2005). Review of Code and Solution Verification Procedures for Computational Simulation. *Journal of Computational Physics*, *205*(1), 131-156.

SAE. (2021). Analytical Methods for Aircraft Seat Design and Evaluation. In: SAE International.

Sandia. (2014). *How to Use the Feature Coverage Tool*. (SAND2014-19758).

Soltis, S. J., & Forest, K. E. (1999, 1999). *Comparison of Results from Dynamic Tests of an Airplane Seat at Different Facilities*

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House.

Taylor, A. M., DeWeese, R. L., & Moorcroft, D. M. (2017). *Comparison of AC and Original Formulation Confor Foam Performance in Civil Aircraft Vertical Impact Tests*. (DOT/FAA/AM-17/1).

Taylor, A. M., Moorcroft, D. M., & DeWeese, R. L. (2017). Comparison of the Hybrid II, FAA Hybrid III, and THOR-NT in Vertical Impacts.

Toptan, A., Porter, N. W., Hales, J. D., Jiang, W., Spencer, B. W., & Novascone, S. R. (2022). Verification of MOOSE/Bison's Heat Conduction Solver Using Combined Spatiotemporal Convergence Analysis. *Journal of Verification, Validation and Uncertainty Quantification*, *7*(2), 021006-021001,021006-021018.

Toptan, A., Porter, N. W., Hales, J. D., Spencer, B. W., Pilch, M., & Williamson, R. (2020). Construction of a Code Verification Matrix for Heat Conduction With Finite Element Code Applications.

Toptan, A., Porter, N. W., Hales, J. D., Williamson, R., & Pilch, M. (2020). *FY20 Verification of BISON Using Analytic and Manufactured Solutions*. https://www.osti.gov/biblio/1614683

Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. The University of Chicago Press.

Wald, A. (1943). An Extension of Wilks' Method for Setting Tolerance Limits. *Annals of Mathematical Statistics*, *14*, 45-55.

Wang, J., Martin, W. R., Downar, T. J., Kochunas, B., Andrews, N. C., Gilkey, L., Walker, E. D., Collins, B. S., & Pilch, M. (2022). Code Verification and Solution Verification framework in pin-resolved neutron transport code MPACT. *Annals of Nuclear Energy*, *178*, 109365. https://doi.org/https://doi.org/10.1016/j.anucene.2022.109365

Wilks, S. S. (1942). Statistical Prediction with Special Reference to the Problem of Tolerance Limits. *Annals of Mathematical Statistics*, *13*, 400-409.

Zuber, N., Wilson, G. E., Ishii, M., Wulff, W., Boyack, B. E., Dukler, A. E., Griffith, P., Healzer, J. M., Henry, R. E., Lehner, J. R., Levy, S., & Moody, F. J. (1998). An Integrated Structure and Scaling Methodology for Severe Accident Technical Issue Resolution: Development of Methodology. *Nuclear Engineering and Design*, *186*(1-2), 1-21.

Appendix A Prediction Uncertainty

Prediction is defined by (ASME, 2006) as using a model to calculate a response where the modeler does not know the experimental outputs. This appendix discusses prediction uncertainty from two perspectives: uncertainty associated with alternate plausible models and quantifying model form error and uncertainty. These perspectives are often not recognized. The distinguishing attribute is the availability of system-level data against which the model can be assessed. The following discussions are intended to clarify the differences but are not prescriptive. Methods will vary depending on the nature of the problem.

A.1Uncertainty associated with alternate plausible models

Predictive uncertainty associated with alternate plausible models involves quantifying uncertainties in all elements of the simulation conceptual model and their computationally expensive propagation through the computational model.

The term “model” is often used interchangeably with the family of alternate plausible models, but this is incorrect in a rigorous sense. Typically, the scope of predictive uncertainty is confined to parameter uncertainty (e.g., uncertain material properties), which is unnecessarily restrictive. This approach to predictive uncertainty could, and should, consider all conceptual model elements, e.g., alternate plausible constitutive models (e.g., failure models), alternate plausible system states (including geometric fidelity), and alternate plausible environments as examples. It is important to cast the net widely when identifying sources of uncertainty and estimating their magnitude. This lends credibility to the assertion that some (non-unique) combinations of alternate plausible models approximate an unknown reality.

There is no basis for restricting the family of models in the extreme case that no relevant data exist to assess the family of alternate plausible models at any level of the validation hierarchy. Nothing can be said about bias in the family of models, and nothing can be said about the ability of the family of models to correctly predict trend behaviors for alternate points in the application parameter space. Subjective judgment is the only basis to accept the family of alternate plausible models from an application perspective, so the judgment of subject matter experts and independent peer review is critical.

Regulatory applications require some validation support for the family of models. Consider the case where relevant data are available to test the family of models through a validation hierarchy but not at the system level. The validation hierarchy offers an opportunity to restrict the family of models, reduce errors, and reduce uncertainties, possibly by calibrating the family of models to test results. Bias in system leve- predictions can be inferred, but not quantified, to be acceptable if the bias is acceptable for all elements in the validation hierarchy. Acceptable results for a validation hierarchy provide objective evidence to support a risk-informed decision to accept a restricted family of models for a regulatory application.

Sometimes, a single test is available at the system level. A single test provides a referent to judge whether the restricted family of models is consistent with relevant system-level responses. This is an accept or reject decision. Figure A.1 illustrates this approach to prediction uncertainty. The red distribution aggregates the results of many alternate plausible models. The blue line represents the results of a single test, if available. The proper interpretation is that some reasonable combinations of alternate plausible models are consistent with the test observation. It makes no sense to talk about model bias because many alternate models are represented with this approach to prediction uncertainty. Comparing predictions from a family of models provides an additional opportunity to calibrate, i.e., further restrict the family of models that is consistent with system response data.

This approach to predicting uncertainty comes with a high computational cost. All combinations of input uncertainties must be propagated through the computational model. Simulation solution errors must be understood and at least bounded for all points in the uncertainty space. Appendix F.2 shows that up to 1900 Monte Carlo iterations are required to estimate the 95th percentile of the distribution accurately. Bayesian calibration comes with a substantially higher computational cost. This approach to prediction uncertainty is not practical for aircraft seat certification.

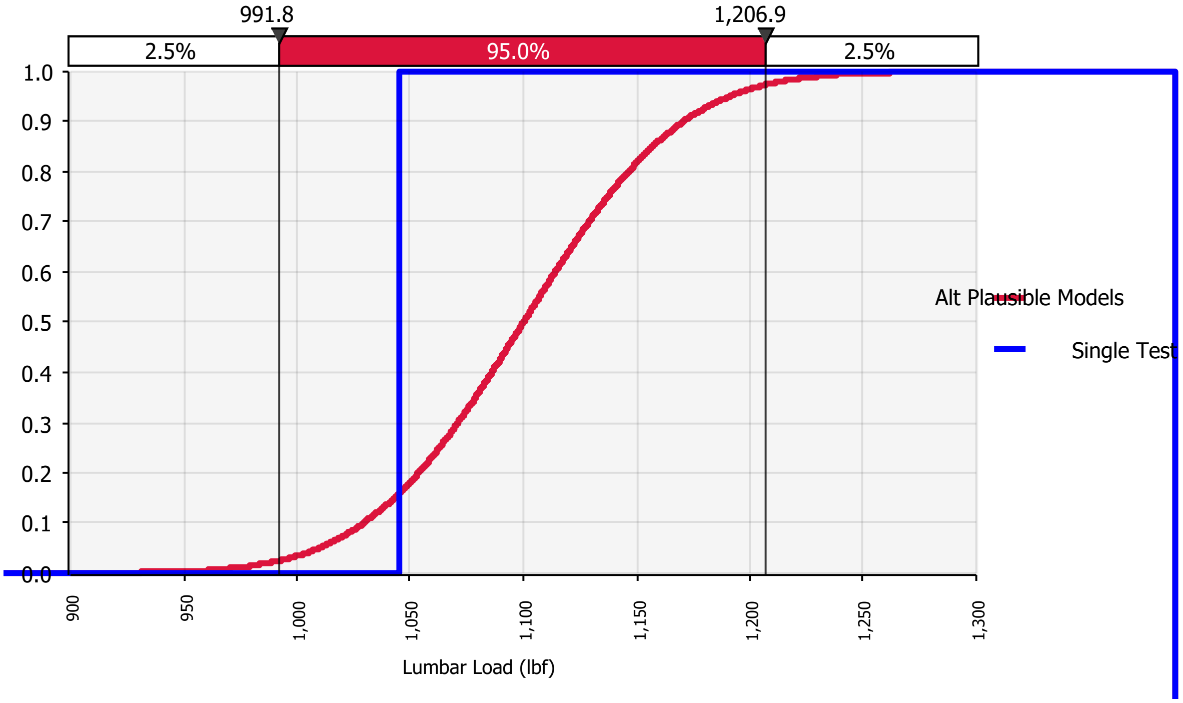


Figure A.1: Predictive uncertainty represented by a family of alternate plausible models

A.2 Model form error and uncertainty

Model form error and uncertainty fundamentally differ from uncertainty associated with alternate plausible models. They are the product of assessing a fully prescribed model (frozen) against test data capable of assessing the sensitivity of predicted QOIs when a design either interpolates or extrapolates on the available data.

This is conceptually illustrated in Figure A.2, where seven available tests (solid circles) span a range of designs and environments. In the first case, we are asked to assess a new design that interpolates on the database (open green circle). In the second case, we are asked to assess a design that extrapolates significantly beyond where there are data in the design space (solid green circle).

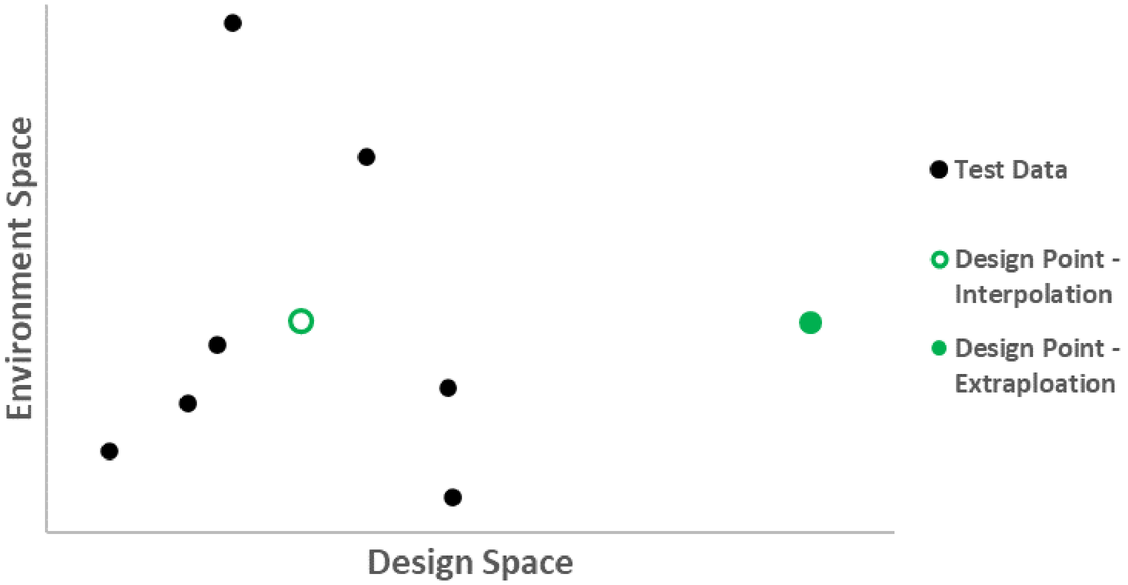


Figure A.2: Design points that either interpolate (open circle) or extrapolate (solid green circle)

Model form error is assessed by comparing predictions to data using the discrepancy metric,

$$E = \ln \frac{M}{P} \;, \qquad \textbf{A-1}$$

which limits to the relative error when predictions (P) and measurements (M) are close. Figure A.3 conceptually depicts model accuracy across the design space. There should be no bias or dependency on position in the design space for a perfectly predictive model. In this illustration, the median error (bias) is -0.028, within typical validation acceptance criteria of $|E_{50}| < 0.10$. The black trendline represents model form error as a function of position in design space, which in this illustration shows that the physics model is not fully predictive over variations in design space. The scatter of data around the model form regression line represents model form uncertainty. Precision errors (sample-to-sample variability, repeatability, and reproducibility) are contributors to model form uncertainty.

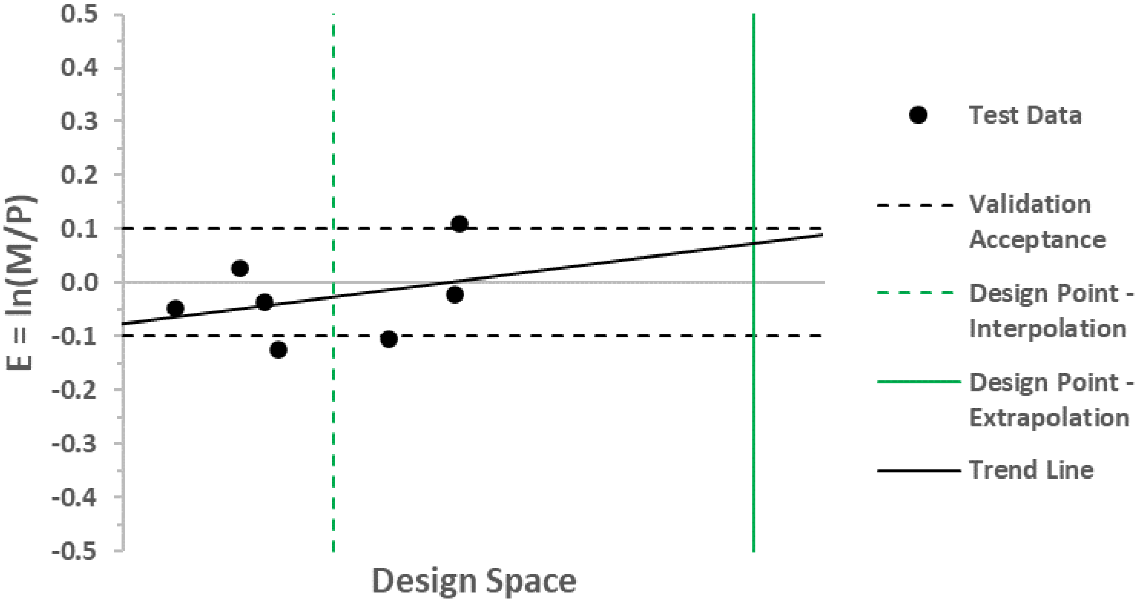


Figure A.3: Assessment of model form errors in design space

Although the model may not be perfect, it may be good enough to assess new designs if |E| < 0.10 at the design point. In this case, the prediction, P(DP), at the design point is expressed as

$$\mathrm{P(DP) = S(DP)\ e^{E(trendline,scatter)}} \quad \textbf{A-2}$$

where S(DP) is the simulation result at the design point. The trendline empirically compensates for deficiencies in the physics model that cannot fully predict trends across the design space. It is a matter of inference that model form errors and uncertainties at the design point are equivalent to what is assessed for the validation database.

(NaRC, 2012) commented that both interpolation and extrapolation can be risky. One risk is illustrated in Figure A.4, where model form errors may be acceptable for interpolation but grow to unacceptable levels when extrapolated outside the range of data. Other sources of risk might occur when Nature exhibits behavior that is not accurately represented in the physics model. Examples are resonance behavior, threshold phenomena such as failure, melting, or runaway chemical reactions, or regime transitions such as laminar flow to turbulent flow. Subject matter expertise is critical in establishing the credibility of models when interpolating and, most significantly, when extrapolating.

It takes more than a few data points to estimate trend lines and characterize the scatter about the trend line. Standard errors are important unless the database is exceptionally large, but they are typically ignored. Standard errors are important intangibles that should be reported with all assessment results.

The computational burden associated with quantifying model form error and uncertainty is negligible compared to exploring alternate plausible models. In both cases, simulations will be required to assess simulation solution errors. While assessing alternate plausible models requires up to 1900 Monte Carlo iterations (each involving a code simulation), quantifying model form error and uncertainty requires only one code simulation for each system-level test, which is only seven code simulations for this illustration. This is because the model is fully prescribed (i.e., fixed, or frozen). This is the approach to prediction uncertainty adopted for this report.

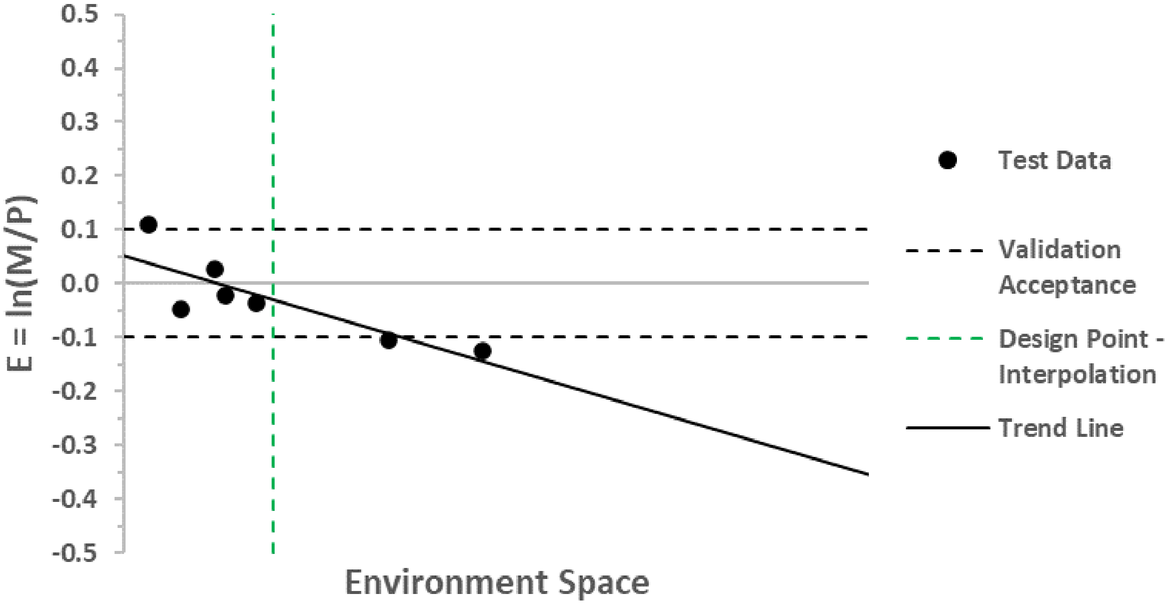


Figure A.4: Assessment of model form errors in environment space

Appendix B Extended Phenomena Identification and Ranking Table

B.1 ePIRT

The US Nuclear Regulatory Commission (NRC) championed the use of phenomenon identification and ranking tables, PIRTs (Boyack et al., 2001; NRC, 1989). Sandia National Laboratories has also had success using PIRTs within its nuclear weapons program and when communicating with external peer review panels. PIRT is an application-specific tool for organizing and communicating:

1. Physics capabilities that are needed and what capabilities are not needed. It is important to identify both and provide evidence or documented rationale justifying the choice.
2. Sufficiency of existing physics capabilities within analysis tools (codes) to meet application needs. PIRT answers the question, do you have the needed capabilities for assessment?
3. What gaps need to be addressed? This drives capability development and research activities.
4. Efficiency of planned activities needed to address gaps in capabilities. Do only what is necessary.

I successfully used PIRTs when managing the ASC V&V program at Sandia National Laboratories. It is a powerful tool for coordinating the user community, capability development community, and research community across multiple organizations and funding streams. PIRT is commonly a subjective process involving key stakeholders as appropriate for the issue's importance. Key stakeholders might include subject matter experts from the analysis community, academia, code developers, industry, and regulatory agencies. The NRC also championed the use of formal scaling as a means of ranking phenomena (Zuber et al., 1998). It wasn't easy to implement for complex models involving multi-physics and disparate time scales. PIRTs are living documents that can change over time as understanding changes or new capabilities are developed and implemented in codes.

The original focus of PIRT was phenomena, i.e., physics and material models. I have extended the scope of PIRT to include all three elements of the simulation conceptual model: environments, system state, and physics. The new tool is called ePIRT for extended PIRT. Table B.1 shows the ePIRT developed for this project. This draft ePIRT was developed at the start of the project by MPC and updated at the conclusion of the project with additional input from the FAA (David Moorcroft, CAMI).

PIRTs are application-specific, so quantities of interest (QOI), regulatory requirements, scenarios, environments, ATD, and seat design must all be specified at the start. Importance and capability assessments in Table B.1 are specific for this demonstration project, but the ePIRT developed here is a framework for expansion by anyone seeking CbA of their real-world seat design.

A comprehensive list of potentially needed elements of the conceptual model is a main feature of an ePIRT. It is essential to cast the net widely in the beginning. Affinity groupings or hierarchies can improve understandability. Importance ranking of the elements is the second key feature of an ePIRT. Simple and complex ranking schemes can be found in the literature, often to aggregate or reconcile conflicting assessments by subject matter experts, e.g., the Analytical Hierarchy Process (AHP) was used in (NRC, 1989). My experience is that simple is better; consequently, a simple high-, medium-, and low-ranking scheme is seen in Table B.1. I subjectively equate medium importance to an element that might have a ~10% impact on the QOI, i.e., large enough that it should not be ignored but not a first-order impact on the QOI. High and medium importance are subjectively judged relative to medium importance. High importance is a first-order term that is essential to include. Low-importance terms can safely be left out of the analysis.

Capability ranking, which assesses existing capabilities, is the next feature of ePIRT. Again, a simple high, medium, and low ranking works best. Medium is assigned to existing capabilities with documented approximations and limitations that might introduce a 10% bias in simulation results. High and low capability are subjectively judged relative to medium capability. Gap identification grades the discrepancy between capability importance and available capabilities. Gap grading is color-coded: green indicates no gap, yellow indicates a one-level gap, and red indicates a two-level gap. A red gap places a high priority on needed research and/or capability development activities.

The assessment summary and modeling approach is the last feature of an ePIRT. The rationale for importance and capability rankings is provided here. This is also an opportunity to summarize the planned or implemented modeling approach. Supplemental evidence should be provided where available. Evidence for some key rankings is provided in Appendix B2 which provides supplemental information regarding the observed sensitivities of lumbar loads to environments and seat design.

Two capability gaps (yellow) are identified in the ePIRT:

1. Compliance of ATD lower and upper torsos.
2. One-dimensional approach to modeling.

It’s expected that these gaps will not introduce dominant first order effects in the demonstration and would not exist in high-fidelity finite element modeling.

Table B.1: Extended phenomena identification and ranking table (ePIRT)

8/17/24	**Date when PIRT was last updated**		
Who Participated	**Affiliation**		
Martin Pilch	MPilchConsulting (MPC), MPilchConsulting@gmail.com		
David Moorcroft	FAA Civil Aerospace Medical Institute (CAMI), David.Moorcroft@faa.gov		
Quantities of Interest (QOI)	**Comments**		
Max Lumbar Load, L_{max}	14CFR Part 25.526		
Regulatory Requirement	**Comments**		
$L_{max} < 1500$ lb_f	14 CFR Part 25.526		
Scenarios	**Comments**		
Emergency landing conditions	14CFR Part 25.526 Transport category aircraft, MTOW>12500 lbs		
Environments	**Comments**		
Triangular pulse	14CFR Part 25.526 (see Figure 4.3)		
14G	Maximum acceleration		
80 ms	Rise time		
30 degrees	Impact angle (see Figure 4.1)		
ATD	**Comments**		
FAA Approved Hybrid II or FAA-Hybrid III Seated upright	14 CFR Part 25.526		
Seat Design	**Comments**		
	Defined by applicant		
Single seat with rigid frame	Hypothetical design for demonstration		
Forward facing			
No arm rests	Cushion thickness in baseline seat design and nearby design seeking CbA		
2.0” or 2.5” cushion	Cushion material		
Monolithic CF42 (AC)			
Environments	**Imp.**	**Cap.**	**Assessment Summary and Model Approach**
Magnitude of acceleration pulse	High	High	. Maximum G is the key environmental factor controlling lumbar loads prescribed by FAA requirements. All else being constant, the peak lumbar load is expected to scale linearly with maximum acceleration. However, the target is rarely achieved exactly in seat testing. As compensation, the FAA typically normalizes the measured lumbar load to the target G. . FAA practice will be followed by modeling tests using target G as input. Predicted loads will be compared to normalized lumbar loads observed in a validation test. The prescribed input environment can be described analytically as input to the seat.

Rise time of acceleration pulse	Med	High	. Rise time is the second key environmental factor controlling lumbar loads. The rise time is correlated with maximum G as prescribed in FAA requirements. The rise time potentially affects how much momentum can be generated in the ATD before cushion materials lock up. This added momentum may be the source of the load enhancement associated with cushions. Sled testing rarely achieves the target rise time exactly. Target rise times are routinely reported. Still, reporting actual rise time is much less common, and there is no accepted way to rescale observed lumber loads to the target rise time. . I will use the target rise time in all validation and application predictions. The prescribed input environment can be described analytically as input to the seat.
Shape of acceleration pulse	Low	Low	. FAA requirements prescribe a triangular acceleration pulse, which is rarely reproduced in a test exactly. The most important characteristics of the pulse are maximum acceleration and rise time. Discrepancies from triangular can be seen in earlier testing, but more recent testing does a much better job (not exact) of reproducing the target triangular pulse. . I will only model the prescribed target triangular pulse in validation tests and prediction for the nearby design. A validation test could be modeled using the observed acceleration history as input. However, this is judged an insignificant effect if maximum G and rise times are right.
Impact angle	Med	High	. The angle of airplane impact (see Figure 4.1) dictates how much of the acceleration pulse is directed along the spine. FAA requirements prescribe a common impact angle of 30^0 for all classes of aircraft, which reduces the maximum possible lumbar load by 13%. The impact angle is well reproduced and documented in all test data. . The prescribed impact angle can be represented exactly in the model.
Seat Design	**Imp.**	**Cap.**	**Assessment Summary and Model Approach**
Number of seats	NA	NA	. Commercial aircraft seats are typically two or three seats supported on a single integrated frame. That means that some seats are bridging while others are cantilevered relative to the support elements of the frame. This could introduce additional compliance in the frame that might impact lumbar loads as a secondary effect. The industry performs tests with full seats as designed. (DeWeese et al., 2021) showed significant sensitivity of lumber loads to cantilevered and bridge positions. Most research data in the open literature are for single-seat configurations, and the hypothetical design for the demonstration prescribes a single seat. . The number of seats is not a factor in this assessment and demonstration.
Seat orientation	Low	NA	. Aircraft seats are forward-facing in transport category aircraft. Data in the open literature is predominantly for forward-facing seats. . I will limit the assessment to forward-facing seats.
Armrests	NA	NA	. As fielded, aircraft seats have armrests that would commonly be in the down position before a landing accident. (DeWeese et al., 2021) showed that armrests could attenuate lumbar loads if arms are supported by the arm rests. Armrests are not represented in the open research literature. . Armrests will not be included in the model developed here, consistent FAA assessment practices.
Seat pan	Low	NA	. The seat pan in commercial seats may exhibit compliance not observed in rigid, single-occupant seats common in the open research literature. . Compliance in the seat pan will not be modeled, meaning that the ATD lower torso will directly experience accelerations without modification by the seat pan.

Seat frame	Low	NA	. Seat frames are metal structures designed not to collapse or fail during emergency landing conditions. Seats in a bridging or cantilevered position (multiple seats carried by a single frame) might experience additional compliance that could affect lumbar loads. (DeWeese et al., 2021) showed sensitivity of lumber loads to cantilevered and bridge positions. Most research data in the open literature are for single-seat configurations with intentionally rigid frames. The hypothetical design for this demonstration prescribes a rigid frame. . The seat frame will be modeled as rigid, which means that the seat pan will directly experience acceleration pulses without modification by the seat frame.
Seat frame attachment To cabin floor	Low	NA	. The seat frame is anchored to the cabin floor by attachments, allowing the seat to be removed or repositioned on demand by aircraft maintenance crews. The attachment is designed to not fail during emergency landing conditions. It is unlikely that attachments could introduce compliance into the system that could alter lumbar loads. . The seat frame attachment to the cabin floor is assumed rigid and not modeled.
Cushion	High	Med	. Cushions compromise passenger comfort with safety. Cushions can enhance lumbar loads by up to 50%, depending on the cushion material and thickness. The contact point between a passenger, or ATD, and the cushion is inherently 3D. . The cushion will be modeled as a separate 1D spring—damper component. The 3D nature of the contact point will be reflected in the 1D model using material characterization data that replicates the 3D contact.
Cushion material	High	High	. Load enhancement due to cushions depends on the cushion material. This study's baseline test and nearby design specify CF42 (AC) foam for the cushion. The Confor family of foams is known to be highly rate-dependent, which means that the seat and the passenger (or ATD) are tightly (nearly) coupled. Confor foams were recently reformulated to comply with restrictions on using certain fire-retardant chemicals in the original formulation. (Taylor, DeWeese, et al., 2017)showed that the reformulation did not introduce a systematic bias in lumbar loads and that observed random differences are what can be expected for replicate test variability. . Material characterization tests (using current generation CF42 (AC) foam) will be conducted for this study.
Cushion thickness	Med	High	. Load enhancement due to cushions depends on the cushion thickness (see Figure B.1). This study’s baseline test and nearby design specify monolithic CF42(AC) foam cushions of 2.0” and 2.5” uniform thicknesses, respectively. Existing sled track data for CF42 (AC) foams suggest that the load enhancement is approximately 10%. DAX26 exhibits similar sensitivity, while AF4050 exhibits greater sensitivity to cushion thickness. Nominal cushion thickness is well reported for each test, but the actual thickness may vary slightly. . Validation and certification predictions will be performed using the nominal cushion thickness, which can be represented exactly in the model. Variations from nominal will be captured in the treatment of precision errors (see Appendix E.2).
Cushion cover	NA	NA	. Commercial aircraft seat cushions have a seat cover (cloth, leather, etc.), but much of the research data in the open literature is for cushions without any cover. The cover stiffens the cushion by providing both lateral support and added resistance to the expulsion of air from the foam cells. Limited data in (DeWeese et al., 2021) suggests that adding a seat cover can reduce lumbar loads by about 12.8%. Unlike reality, this demonstration study prescribes a hypothetical cushion design without a cover consistent with the available research data. . No need to model the cushion cover for the baseline test or the nearby design.
Composite cushions	NA	NA	. Aircraft seats are typically composed of different foam layers and different stack-ups of foam. For instance, seat cushions might involve a comfort layer stacked upon a layer of floatation foam. In the open research literature, only limited data exists for this configuration. In practice, aircraft seat cushions are typically contoured by the buildup of different foams. Unlike reality, this demonstration study proposes a hypothetical monolithic cushion of uniform thickness. . There is no need to model the features of composite cushions for the demonstration.

Passenger Demographics	Imp.	Cap.	Assessment Summary and Model Approach
FAA approach to passenger demographics	NA	NA	. 14 CFR Part 25.526 prescribes that lumbar load assessments be performed for a 170 lb_f approved ATD. This prescription is foundational to the establishment of the 1500 lb_f lumbar load threshold. See Section 0 for more discussion. . Accept regulatory approach and requirements.
ATD model	High	High	. The FAA specifies that either Hybrid II or FAA-Hybrid III can be used for seat certification. The Hybrid II ATD was historically used for aircraft seat certification. The manufacturer replaced the Hybrid II with the Hybrid-III, which had some significant differences (e.g., spine and lumbar load cell orientation) that were problematic for the FAA when comparing results to those generated using the Hybrid II ATD. The FAA then developed the FAA-Hybrid lll ATD to provide compatibility with seats certified with the Hybrid II ATD; however, (Olivares et al., 2018) note that FAA-Hybrid III generally produces higher loads (~12.5% higher at 19G) compared to the Hybrid II ATD. This artifact of the ATD has nothing to do with seat performance. In addition to variations introduced by the ATD model, different units of a given model exist within and across testing organizations, thus introducing unit-to-unit variability. . The baseline seat design uses the FAA-Hybrid III ATD, which will be used in the assessment of the nearby design seeking CbA. Any differences resulting from the ATD model (or unit) will be absorbed into the assessment of precision uncertainty (see Appendix E.2).
ATD upper torso	High	Med	. The ATD upper torso (head, chest, arms, and abdomen) represents the weight that can compress the lower torso and cushion. The upper torso weight controls load cell response. The upper torso is compliant because of a rubber column (spine) above the load cell. Compliance of the upper torso has been offered as one explanation why the model developed here poorly represents 19G environments, but the effect seems less important in the 14G environments used in the demonstration. . I will treat the upper torso as rigid mass and use weights appropriate to the ATD model.
ATD lower torso	High	Med	. The ATD lower torso (hips, buttocks, and thighs) represents the additional weight that can contribute to cushion compression but not directly to lumbar loads. The ATD lower torso (buttock) is soft and compressible and might act as a load-enhancing cushion; however, this region is thin compared to cushions and incredibly soft. Consequently, this area will quickly lock up, making additional responses appear as if the lower torso were rigid. When seat cushions are absent in tests, the database suggests that measured lumbar loads are unbiased compared to predictions assuming a rigid lower torso. In summary, the mass of the lower torso is significant for seat compression but is insignificant as an additional load-enhancing "cushion." . The ATD lower torso will be treated as rigid in the model with masses appropriate to the ATD model.
ATD lower body	Low	NA	. The lower body's weight does not compress the cushion or load the lumbar; consequently, does not contribute to lumbar loads. . The ATD lower body will not be modeled.
Seating position	High	High	. Passengers will be instructed by flight attendants to assume the brace position if there is sufficient warning. The brace position has been shown to reduce crash injuries, primarily for horizontal crashes. However, the FAA states that most crashes will not have sufficient time for passengers to assume the brace position; consequently, 14 CFR Part 25.526 conservatively prescribes that the ATD be seated upright for lumbar load assessments. . The ATD will be in the upright seated position in this study, consistent with FAA assessment practices.

Positioning of ATD	Med	Med	. Positioning of the ATD in a seat is a dominant source of variability in sled tests, even though test procedures seek to minimize this effect. There is sufficient data in the open literature to estimate its magnitude, COV~5%. It may be possible to perform sensitivity studies with high-fidelity finite element models, but not the 1-D modeling approach taken in the demonstration. The observed variability observed in testing cannot be predicted with simulation because the input distributions are not known. Although acknowledged, this source of test variability is not explicitly addressed in the FAA test-based certification process, but it does lead to the possibility of a design being wrongly accepted or rejected. The FAA accounts for this variability source when validating a simulation model. . Variations due to seating will be captured in the treatment of precision errors (see Appendix E.2). The potential impact of ignoring this source of variability in test-based certification will be quantified with a simple scoping study.
Preload of seat belt	Low	Low	. ATDs have been observed to “float” from the seat when positioned for a sled test. This means that the ATD weight resulting in static compression of the lower torso and cushion is ill-characterized, and more importantly, the ATD could be out of position when the acceleration pulse is experienced. The FAA concluded that float undermined the repeatability of test results. Seat belt preload and float are compensating effects that crudely mimic the expected static compression of the cushion by the ATD weight. By itself, the preload compression is less than 1% of the observed loads during the dynamic event, so any impact on lumbar loads would be through an alteration of the initial static compression before the dynamic event. . I will stylize the contribution of preload on initial seat compression. I will ignore the impact of preload during the dynamic event.
Physics and Material Models	**Imp.**	**Cap.**	**Assessment Summary and Model Approach**
Momentum Equation			. The demonstration will use a simple 1D spring-mass-damper representation of the ATD (passenger) and the seat cushion. Emphasis will be upon a more rigorous quantification and interpretation of errors, variabilities, and uncertainties in both the validation test and the computer simulations (predictions) and how they impact both validation and the use of the computer model for certification of a nearby design. A simple model ensures that the process application can be completely transparent to FAA and industry engineers and that the governing equations can be completely solved within MS Excel. Finite Element Modeling (FEM) will not be performed in this proposal.
Momentum Term	High	Med	. The momentum equation is the governing conservation equation for the system. The solution of this equation returns the lumbar force and foam compression as a function of time subject to the initial conditions, seat design, and passenger state. A potential capability limitation is the one-dimensional nature of the system of equations. . The momentum term is modeled with the noted limitations
Inertia term	High	Med	. Inertia characterizes resistance to motion changes resulting from the acceleration impulse. A potential capability limitation is the one-dimensional nature of the system of equations. . The inertia of the ATD upper body and the ATD lower torso will be modeled. The inertia of the cushion is negligible and will be neglected.
Foam constitutive model	High	Med	. The foam constitutive model represents the spring and damper in the system of equations. It outputs load as a function of compression and compression rate. Much of the transient and resistance to compression is expected to occur under dynamic limitations. A potential capability limitation is the one-dimensional nature of the system of equations. . The Jeung constitutive model will be adapted and calibrated to model foam compression.

Friction between ATD lower torso and seat pan	Low	Low	. NIAR (Olivares et al., 2018) started placing Teflon sheets between the ATD and the seat pan (tests without cushions) to make friction more repeatable in tests without cushions. The sensitivity of lumbar load magnitude to the presence of Teflon sheets was not quantified. (Olivares et al., 2018) stated that interface friction increased test variability for the Hybrid II ATD and decreased test variability for the FAA-Hybrid III ATD. Differences in the ATDs and acceleration pulses may confound these different trends. . Interface friction will not be modeled. Any potential effect will be absorbed into the inherent variability of the database.
Initial static compression	Med	Med	. The ATD (passenger) will be seated, partially compressing the cushion before the dynamic event. The amount of pre-compression may impact how much momentum buildup occurs before the system locks up and becomes fully coupled. This additional impact momentum is a possible source for the enhanced loads observed with cushions. A potential capability limitation is the one-dimensional nature of the system of equations. . Static compression will be modeled as an initial condition before the dynamic event
Foam Constitutive Model			. The compression curve is the sole constitutive model used in the analyses of lumbar loads. As assessed under physics phenomena, a quasi-static “spring” term and a dynamic “damper” term are needed. This formulation of the constitutive model is 1D, which is consistent with the 1D nature of the modeling approach. . I will adapt the Jeung constitutive model to model quasi-static and dynamic compression. This formulation of the constitutive model is 1D.
Quasi-static compression of cushion i.e., rate insensitive	Med	Med	. Quasi-static (i.e., rate insensitive) cushion compression occurs when a passenger first sits on the seat and during the early phases of a dynamic event. However, a dynamic event will quickly move into a rate-sensitive regime. The quasi-static compression curve (force as a function of compression in %) comprises three regimes: elastic, plastic plateau, and densification. The elastic regime is limited to small forces and compressions. The plastic plateau occurs over a wide range of compressions until densification becomes dominant. Small increases in force result in substantial changes in compression in the plastic plateau regime. There is a densification limit (typically near 90%) beyond which further compression is impossible. Asymptotically larger forces are required for incrementally larger compressions as the densification limit is approached. . The FAA will conduct material characterization tests specific to CF42 (AC) at quasi-static rates to support calibrating the foam constitutive model.
Dynamic compression of cushion i.e., rate sensitivity	High	High	. Dynamic (i.e., rate-sensitive) cushion compression occurs over much of the event. Hooper demonstrated the importance of rate effects for several common aircraft seat foams. DAX26R as an example, the compressive resistance force at 30 in/s was about 5.3 times higher than the quasi-static compressive force at 0.033 in/s. Compression rates up to 60 in/s may be possible during dynamic events. . The FAA will conduct material characterization tests specific to CF42 (AC) at dynamic rates up to 60 inches/s to support calibrating the foam constitutive model.

Temperature effects on compression	Low	Low	. The sensitivity of lumbar loads to temperature, either environmental or through self-heating during compression, has not been assessed for aircraft seat foams. Sled tests occur at 66-78 degrees F and a relative humidity from 10% to 70% per 49 CFR 572, like cabin environments during emergency landing conditions. Consequently, environmental temperatures are not expected to impact lumbar loads. Self-heating can be significant in the high-rate deformation of metals with temperatures approaching some significant fraction of the melting temperature. At these temperatures, the strength of metals is significantly degraded. Consequently, applying the Johnson-Cook constitutive model to metals often includes a temperature term that softens the response. Foams “decompose” at elevated temperatures rather than melt. Still, it has never been assessed through testing that self-heating can significantly degrade the compressive response of foams used in aircraft seat applications. Softening of the compressive response curve has been observed in other polyurethane foams at rates of 200 inches/s, which is significantly higher than expected in aircraft accidents. . Potential temperature effects will not be factored in the assessments and demonstrations performed in this study.
Aging	Low	Low	. Some aircraft seats may be in service for many years, a decade or more, experiencing many loading and unloading cycles. Although initially certified to limit lumbar loads < 1500 lb$_f$, the foams in aircraft seat cushions could potentially deteriorate over extended periods, altering their safety function. The sensitivity of lumbar loads to aging foams has yet to be studied. Once certified, always certified. The FAA has no process to recertify seats in service for extended periods. This project is about the initial certification of a seat and relies on validation data from a sled test. In both cases, the foam will be “fresh” and not previously stressed through many cycles. . Aging will not be modeled in the assessments and demonstrations performed in this study.
Creep	Low	Low	. Creep presents a rate sensitivity at extremely low rates and is typically characterized through displacement-controlled testing, the kind of testing done for foam characterization. Material testing for polyurethane foams exhibits a clear quasi-static regime where loads are rate incentive, which implies no creep effect. Compression rates during an accident are too high to exhibit creep behavior. . Creep will not be modeled in the assessments and demonstrations performed in this study.
Lumbar Fragility	**Imp.**	**Cap.**	**Assessment Summary and Model Approach**
FAA approach to lumbar injuries	NA	NA	. 14 CFR Part 25.526 prescribes a risk-informed lumbar load threshold of 1500 lb$_f$. The threshold is conditional on assessment with a 170 lb$_f$ approved ATD and the environments prescribed in the same regulation. See Section 0 for more discussion. . Accept regulatory approach and requirements.

B.2 Sensitivity of lumbar loads to seat design and environment

A research database of 105 sled track tests was compiled from the open literature and presented in Appendix E.1. The database is filtered to gain insights into the sensitivity of observed lumbar loads to design and environments. What does data alone have to say by itself without any computer simulations?

Figure B.1 illustrates that lumbar loads have a first-order sensitivity to cushion material. Lumbar loads for AF4050 are about 35% to 65% larger than CF42 (AC) and DAX26, depending on the cushion thickness. Doubling the foam thickness increases loads by less than 10% for CF42 (AC) and DAX26.

Figure B.2 shows the sensitivity of CF42 (AC) foam to the environment. Lumbar loads increase by ~65%, while the environmental G increases by ~36%. A cushion enhances loads by more than double what would be expected for a seat with no cushion.

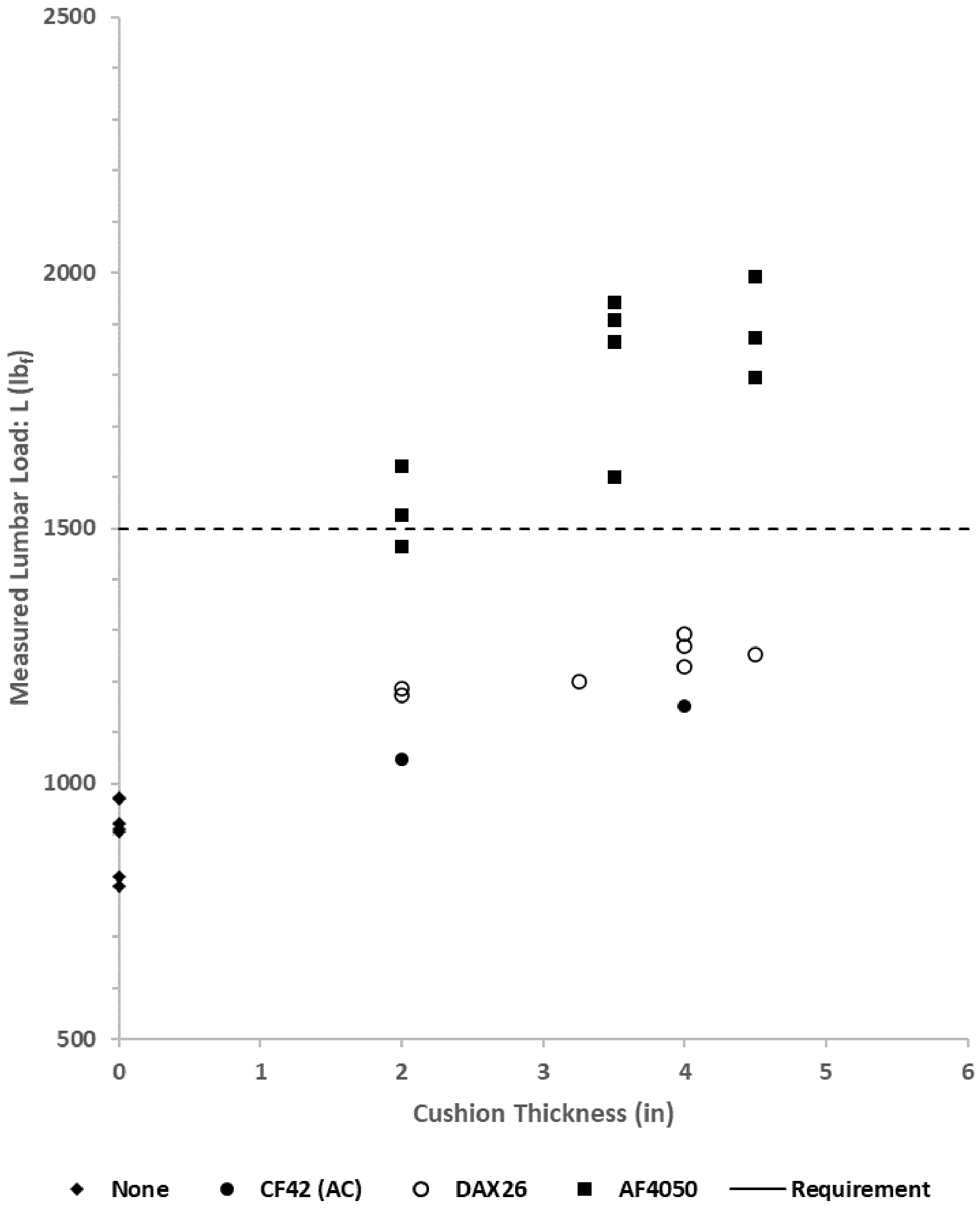


Figure B.1: Sensitivity of lumbar loads to design for a 14G environment

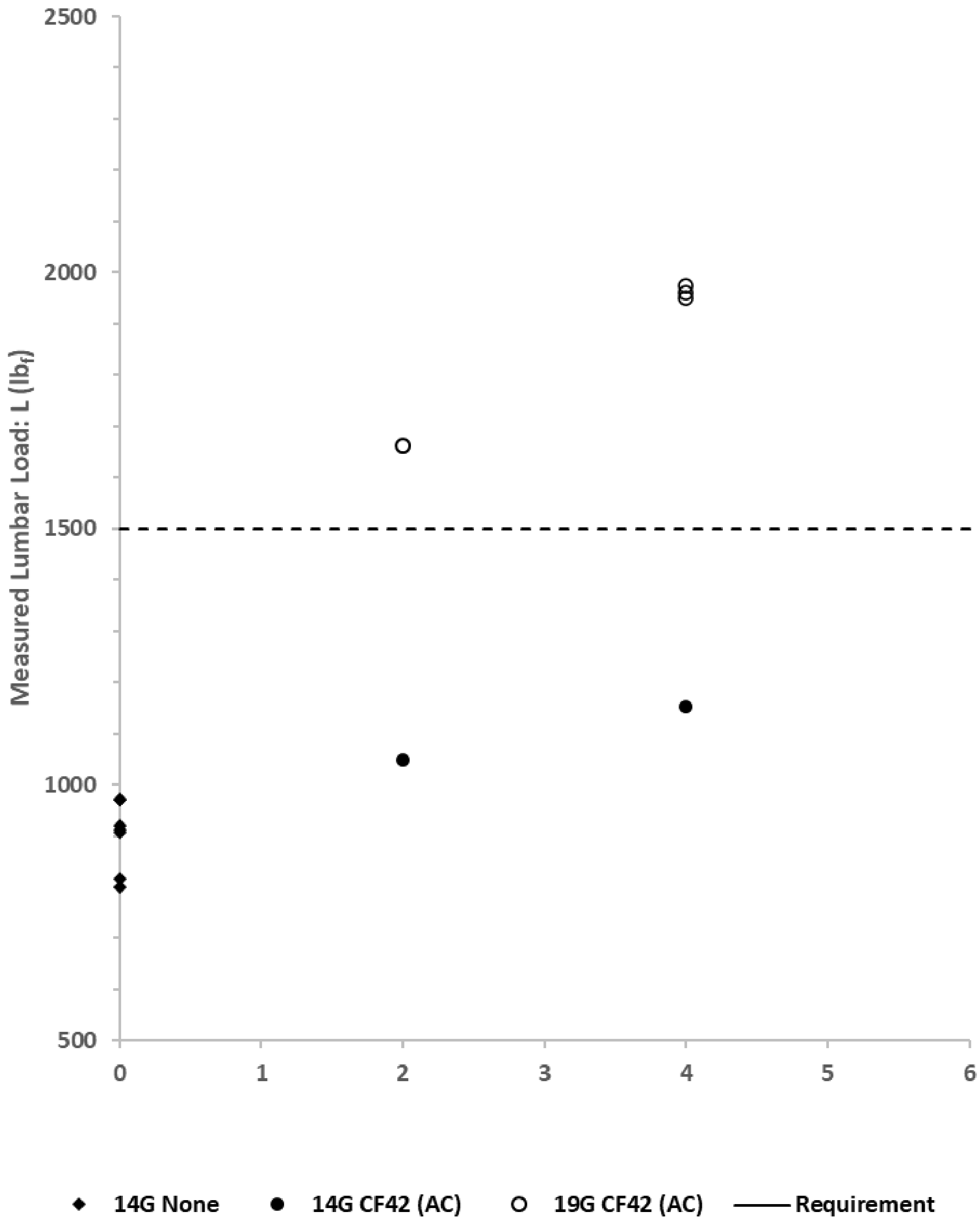


Figure B.2: Sensitivity of lumbar loads to environments for CF42 (AC) foam

`

Appendix C CF42 (AC) Constitutive Model

The open literature did not provide adequate data for calibrating a foam constitutive model for CF42 (AC) foam. This appendix documents material testing performed specially for this project by the FAA and the calibration of a foam constitutive model for CF42 (AC) foam.

C.1 Material characterization tests for CF42 (AC)

The FAA conducted material characterization tests to support the demonstration of BEPU. The tests were conducted by Willian H. Carroll, a member of the Aeromedical Engineering Sciences Section (AAM-632) of the FAA. They were conducted during September 2023, with one follow-up test in October 2023 at the FAA Civil Aerospace Medical Institute (CAMI).

All tests were conducted on an MTS Landmark Hydraulic Load Frame calibrated before material testing. The protocol followed ASTM D 3574 (ASTM, 2010) for Type B (IFD) testing. The setup for a Type B test is shown in Figure C.1. The platen (bottom) has a contact area of 50 in^2 with the foam (above), which is representative of passenger (and ATD) contact with a seat. The tests are uniaxial and rate-controlled. Type B testing is well suited for modeling cushions as a one-dimensional spring-damper component, which is the approach taken in the demonstration. Table C.2 shows the matrix of nine CF42 (AC) material characterization tests. The test matrix spans the range from quasi-static to the highest rate measured in sled tests. The FAA has measured, Table C.1, the maximum compression rate for a few 14G sled tests. The maximum compression rate for 2.0” CF or DAX foams is ~ 20/s, and allowing that data could be used for 19G environments, a maximum compression rate of ~27/s might be expected. Compression rates for AF foams are higher. (SAE, 2021) recommends testing rates up to 30 in/s, which is lower than the rates shown in Table C.1.

Compression rates in the test matrix are anchored at the (ASTM, 2010) testing rate with equal ln(rate) spacing on either side. Two tests (M23075 and M23080) in the rate-dependent regime are replicates. Tests M23079 and M23071 were conducted at the two lowest rates but are effectively replicates because they are both well within the quasi-static regime. An important determination that a test is either quasi-static or rate-dependent is to observe the load measurement when compression is stopped at the end of the test. The test is quasi-static if the load holds constant and rate-dependent if the load relaxes asymptotically to a lower value.

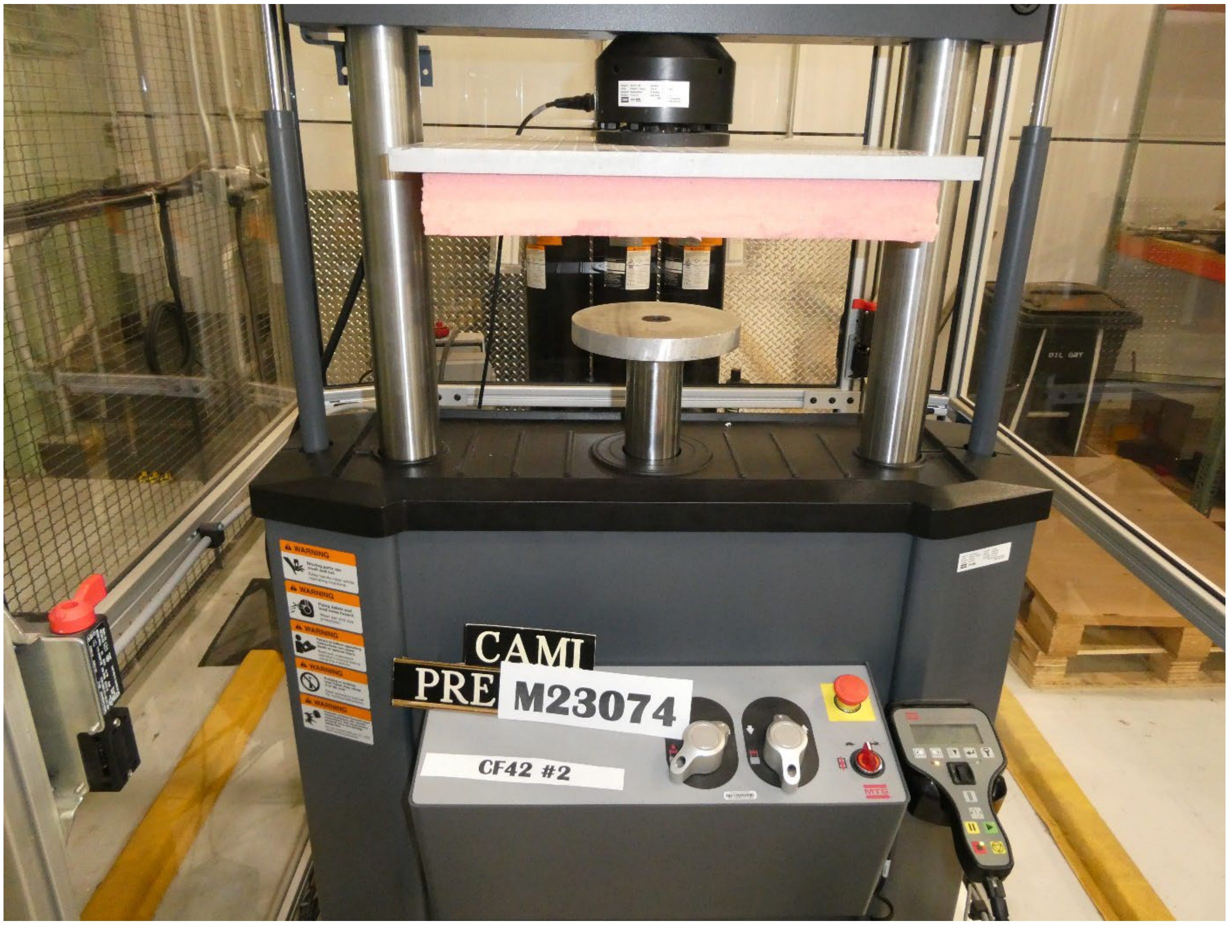


Figure C.1: ASTM D 3574 Type B test (IFD)

Table C.1: Maximum compression rates observed in 14G sled tests

Test #	Foam	Thickness (in)	Env: G's	Max Rate (in/s)	Max Rate 1/s
A04071	CF47	4.20	14	43	10.3
A04035	DAX26	2.00	14	27	13.7
A04036	DAX26	2.00	14	41	20.6
A04055	DAX26	3.25	14	46	14.0
A04058	DAX26	4.50	14	48	10.8
A04057	DAX90	2.00	14	42	21.0
A04062	DAX90	3.25	14	60	18.4
A04063	DAX90	3.25	14	68	20.8
A04066	DAX90	4.50	14	89	19.8
A04061	AF5565	2.00	14	53	26.6
A04065	AF5565	3.25	14	67	20.5
A04069	AF5565	4.50	14	93	20.6

Table C.2: Matrix of CF42 (AC) material characterization tests

	Compression Rate		
Test Number	in/s	1/s	Comment
M23079	3.3E-05	1.65E-05	Replicates
M23071	3.3E-04	1.65E-04	
M23073	3.3E-03	1.65E-03	
M23074	3.3E-02	1.65E-02	ASTM D 3574
M23075	0.33	0.165	Replicates
M23080	0.33	0.165	
M23076	3.3	1.65	
M23077	33	16.5	
M23078	60	30	

Load cells were calibrated prior to material testing. The error in a measured load of 1126 lb_f is 0.20%, equivalent to 3 lb_f if 1500 lb_f were measured. Measurement repeatability (COV) is 0.03%, equivalent to 0.45 lb_f if 1500 lb_f were measured.

Foam samples used in testing were all cut from the same stock purchased from Skandia in April of 2022. Sample dimensions 18” by 18” by 2”. Some samples were reused when tested at the lowest rates.

C.2 CF42 (AC) material characterization test results

Figure C.2 summarizes results from the nine material characterization tests, presented in terms of increasing compression rates from bottom to top. A compression transient passes through three regimes. The first, for small compressions (<10%), is sometimes called the elastic regime. The forces and associated compressions in the elastic regime are too small to influence the calculation of initial static compression (~50% for CF42 (AC) foams) when the ATD is seated for a sled test; consequently, accurate characterization of the elastic regime is not essential.

Elastic compression quickly transitions into a plastic deformation regime characterized by large increases in compression for small increases in force. This occurs in the range of 10% to 75% compression. The calculation of initial static compression (~50%) relies on an accurate representation of the quasi-static response in this regime. The predicted maximum compression is ~70% for the baseline seat design, where the observed maximum lumbar load was ~1000 lb_f. This is because CF42 (AC) is highly rate-dependent.

The third regime is lockup, which occurs at the highest compressions. It is characterized by an exceptionally large increase in force required to achieve a slight increase in compression. Lockup is a threshold beyond which additional compression is not possible, and it occurs at about 90% compression. Lockup occurs when the foam cells are fully collapsed.

The two pairs of replicate tests can be used to estimate repeatability (test variability within a lab) in FAA material characterization tests (see Table C.3). The assessment is performed at 65% compression for two reasons: (ASTM, 2010) reports values at 65%, and 65% is midrange for predicted compressions in the assessment of the baseline seat design. In each of the two sets

of replicate data, errors are referenced to the computed median, allowing all the data to be pooled. Repeatability calculated for the database is about 8.3%, which can be compared to 1.5% reported in (ASTM, 2010) for 65% compression at a compression rate of 0.033 in/s.
The FAA was the only organization performing material characterization tests for this project. No equivalent data could be found in the open literature; consequently, there is no basis for estimating reproducibility (variability across labs). (ASTM, 2010) reports reproducibility of 3.8% at 65% compression tested and at 0.033 in/s.

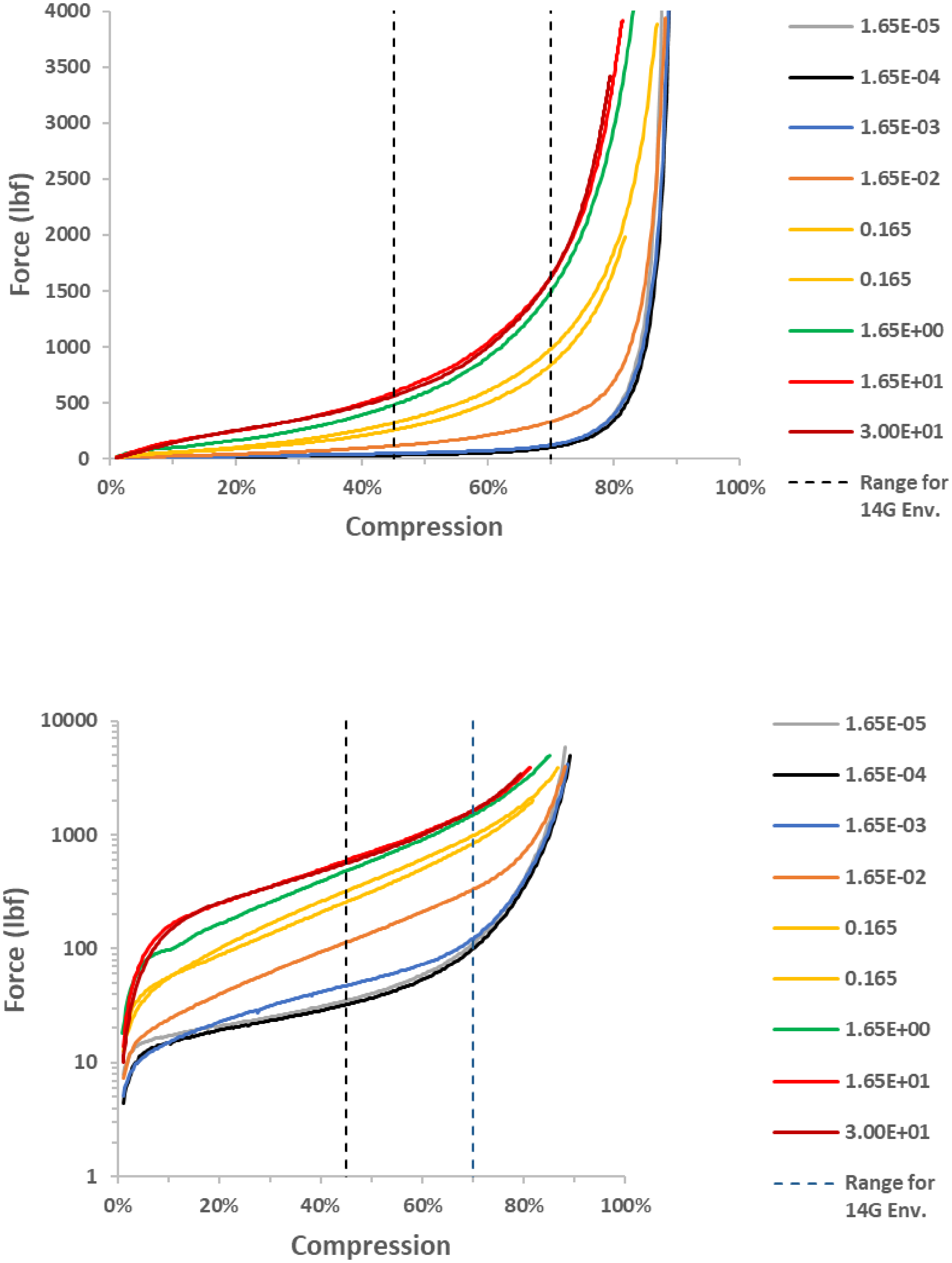


Figure C.2: Summary results from CF42 (AC) material characterization testing

Table C.3: Repeatability of FAA material characterization tests

Test Number	Regime	F(f=65%) lbf	E=ln(F/F$_{50}$)
M23079	Quasi-Static	76.23	0.04289
M23071	Quasi-Static	69.83	-0.04481
M23075	Rate Dependent	762.95	0.08711
M23080	Rate Dependent	635.65	-0.09543
		Pooled Repeatability = s	8.3%

C.3 Calibration of constitutive model for CF42 (AC)

(Johnson & Cook, 1985) developed a constitutive model for metals subjected to high strains, high strain rates, and elevated temperature. The Johnson & Cook constitutive model is widely used, and its use has expanded beyond metals. It is common to find modifications of Johnson & Cook for specific applications.

The Johnson & Cook constitutive model assumes a multiplicative segregation of three terms: a quasi-static deformation term, a rate-dependent term that hardens the response, and a temperature term that softens the repose. Here, we modify Johnson & Cook for use CF42 (AC). The modified form of Johnson & Cook that we use is given by

$$\mathrm{F}(\varphi, \dot{\varphi}) = \mathrm{F}_{\mathrm{qs}}(\varphi)\ \mathrm{F}_{\mathrm{dyn}}(\varphi, \dot{\varphi})\ \mathrm{F}_{\mathrm{temp}}. \qquad \textbf{C-1}$$

The quasi-static compression term is independent of the compression rate, $\dot{\varphi}$,

$$\mathrm{F}_{\mathrm{qs}}(\varphi) = \mathrm{F}_0 \left(\frac{1}{1 - \frac{\varphi}{\varphi_{\mathrm{c}}}} \right)^{\mathrm{a}} \ \text{for } 10\% < \varphi < \varphi_{\mathrm{c}}\ . \qquad \textbf{C-2}$$

The force required to increase compression becomes asymptotically large when compression, f, approaches the lockup value, f$_c$. The quasi-static term is simplified by restricting its application to the plastic response regime; consequently, F$_0$ is the limiting force at zero compression, and “a” is a shape parameter.

(Jeong et al., 2012) modified Johnson & Cook by including static compression in the dynamic response term for polyurethane foams,

$$\mathrm{F}_{\mathrm{dyn}}(\varphi, \dot{\varphi}) = 1 + \left(\mathrm{b} + \mathrm{c}\frac{\varphi}{\varphi_{\mathrm{c}}} \right) \max\left(0, \ln\frac{\dot{\varphi}}{\dot{\varphi}_{\mathrm{c}}} \right) \qquad \text{for } \ \dot{\varphi} > 0 \qquad \textbf{C-3}$$

$$\mathrm{F}_{\mathrm{dyn}}(\varphi, \dot{\varphi}) = 0 \qquad \text{for } \ \dot{\varphi} < 0. \qquad \textbf{C-4}$$

Jeong’s modification is adopted here because, empirically, it was found to improve the fit to CF42 (AC) data. The dynamic response term introduces three new parameters (b, c, and $\dot{\varphi}_c$). The third parameter defines a rate that marks the transition between quasi-static and dynamic response regimes. The first two parameters control sensitivity to rate effects. The decompression representation has no impact on computing maximum lumbar loads.

F_{temp} is a temperature-dependent modifier that softens the response. In the absence of fire in an aircraft cabin, ambient cabin environments are like sled test environments and environments during material characterization tests. Softening of compression curves at very high compression rates (~100/s) has been reported in (Croop & Lobo, 2009; Neilsen et al., 2007), presumably because of self-heating. The highest rate in the test matrix (Table C.2) is 30/s, and the maximum strain rate predicted for CF42 (AC) for the baseline seat design is ~5/s, which is well below where temperature effects were observed; consequently,

$$F_{temp} = 1.0\,. \qquad \textbf{C-5}$$

The constitutive model for CF42 (AC) has five parameters determined by minimizing the RMS error between data and fit values using the constitutive model. The error between a data point and the corresponding fit value is

$$E = \ln\frac{M}{Fit}\,, \qquad \textbf{C-6}$$

where M is the measured force and “Fit” is the computed force. This form of error is critical because measured forces vary by two orders of magnitude over the transient. Otherwise, a few data points near lockup would have disproportionate weight during the fitting process. In addition, this representation of the discrepancy honors the fact that the compressive force can never be negative. MS Solver determines the five parameters that minimize the RMS error.

Data are pooled across all nine material characterization tests for the fitting process. At my request, the FAA filtered test data to approximately 500 data points for each test before being submitted for fitting. This ensures equal weight for each test in the fitting process. Test data was additionally filtered before fitting. Data with compressions less than 10% were filtered out of the fitting process because the elastic regime is intentionally not represented in the constitutive model. In addition, data where the measured load exceeded 2000 lbf were also filtered out before fitting. The intent was to improve the fit where it matters most to lumbar load predictions. Table C.4 lists the optimized parameters for the CF42 (AC) constitutive model.

Table C.4: Optimized parameters for the CF42 (AC) constitutive model

F(f=0) F_0 lb_f	Compression at Lockup, f_c	Shape Param a	Rate Param-1 b	Rate Param-2 c	Critical Rate $\dot{\varphi}_c$ 1/s
15.241	0.89228	1.3206	0.87198	1.6423	6.9682e-3

C.4 Assessment of the constitutive model for CF42 (AC)

Representative comparisons of the constitutive model with replicate tests in the quasi-static regime and the dynamic regime are shown in Figure C.3 and Figure C.5 , respectively. The curve overlay comparisons are qualitatively good over two orders of magnitude in force.

Quantitative discrepancies are shown in Figure C.4 and Figure C.6 for the quasi-static replicate tests and the dynamic replicate tests, and in
Figure C.7 for all nine tests. The average error across all nine tests is negligible, $|E_{avg}| = 4.1\text{x}10^{-6}$, which is expected for curve fitting. The discrepancy plots magnify systematic model form errors, which exhibit a different character in the quasi-static and dynamic regime.

Assessment of the model for rate effects is accomplished by plotting force at 65% compression as a function of compression rate for the nine tests; see Figure C.8. The response falls into two regimes: a quasi-static regime, where forces are insensitive to compression rates, and a dynamic regime, where forces are sensitive to compression rates. The transition from quasi-static to rate-dependent behavior occurs at $6.97\text{x}10^{-3}$/s for CF42 (AC) foam. The median of the discrepancies in Figure C.8 is -1.2%.

In summary, the CF42 (AC) constitutive model successfully captures data trends for the full range of relevant compressions and compression rates. The median error between the constitutive model and data is negligible, although residual systematic model form errors exist. Alternate plausible constitutive models may better fit the data, but this constitutive model is judged acceptable for the demonstration problem and will be frozen at nominal parameters for validation and prediction purposes.

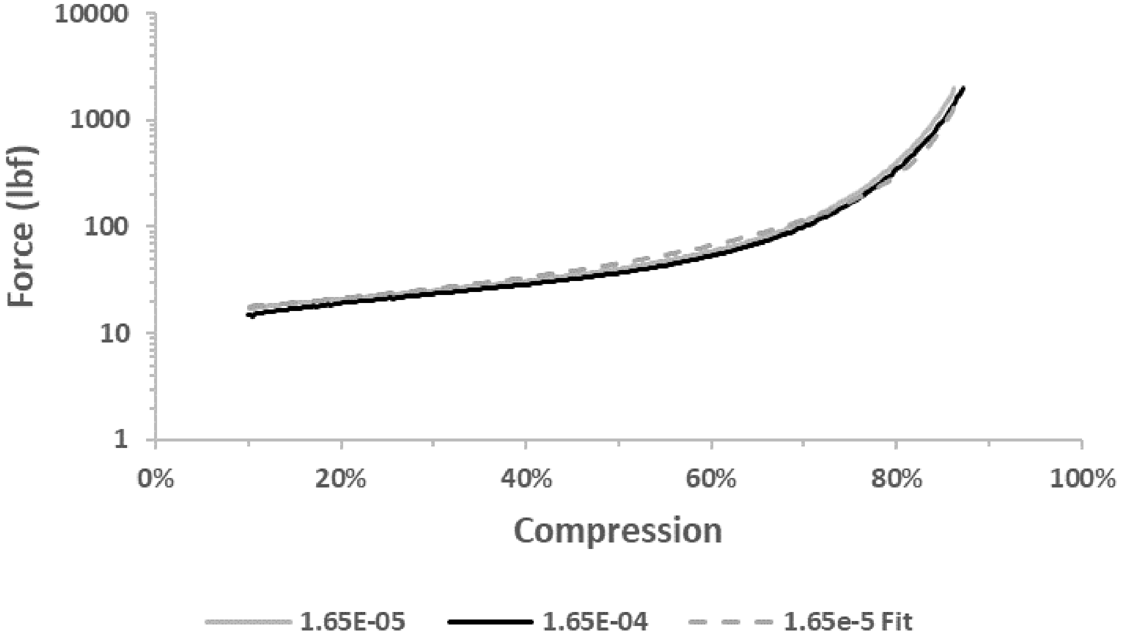


Figure C.3: Comparison of constitutive model and data for quasi-static replicate tests

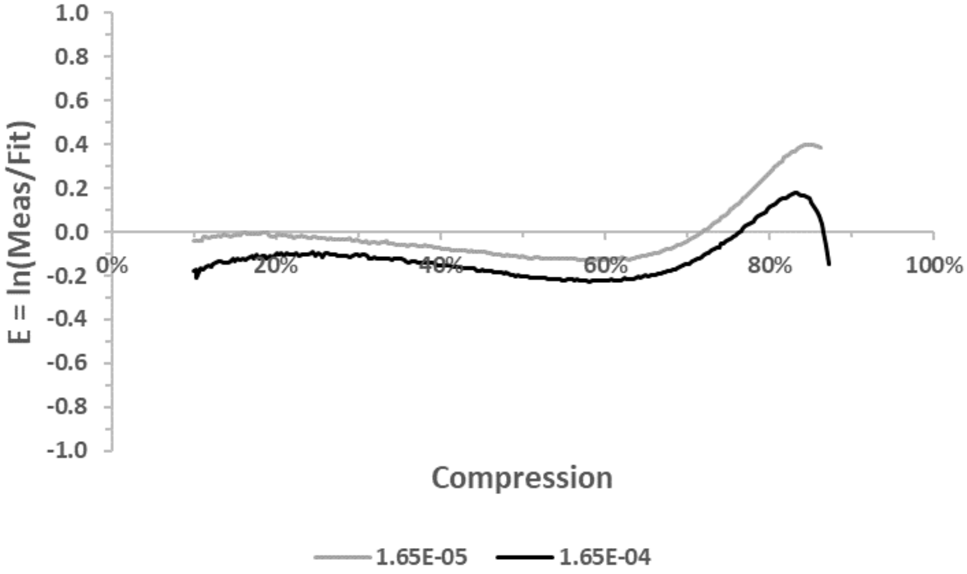


Figure C.4: Discrepancy between constitutive model and quasi-static replicate tests

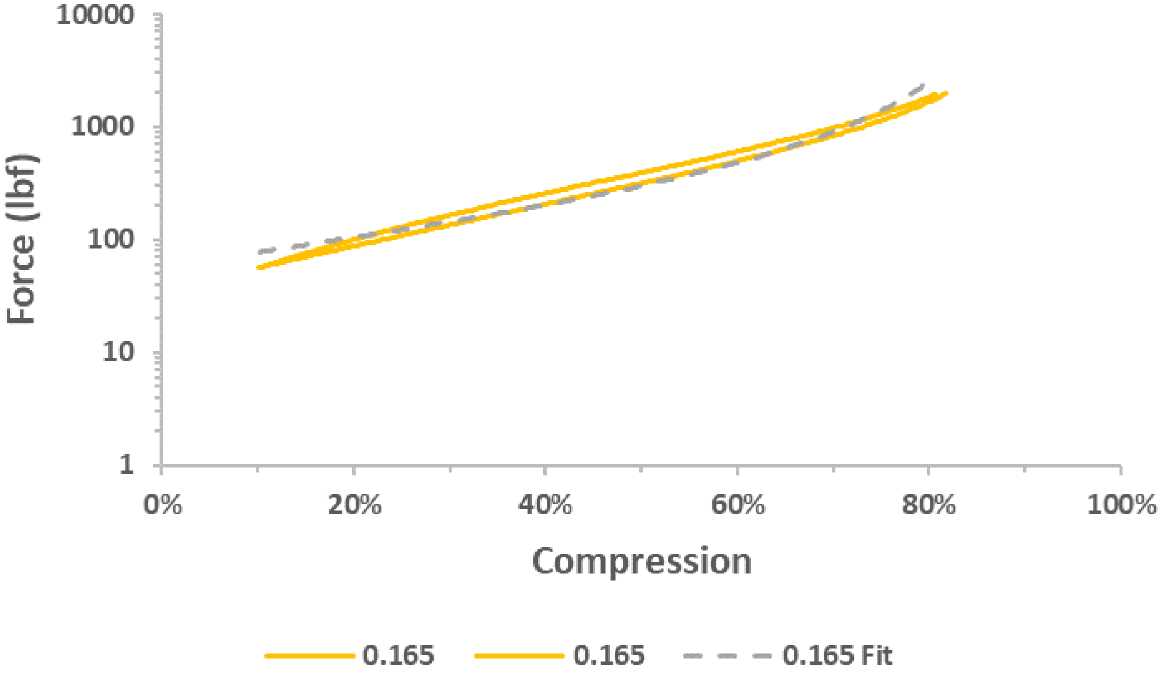


Figure C.5: Comparison of constitutive model and data for dynamic replicate tests

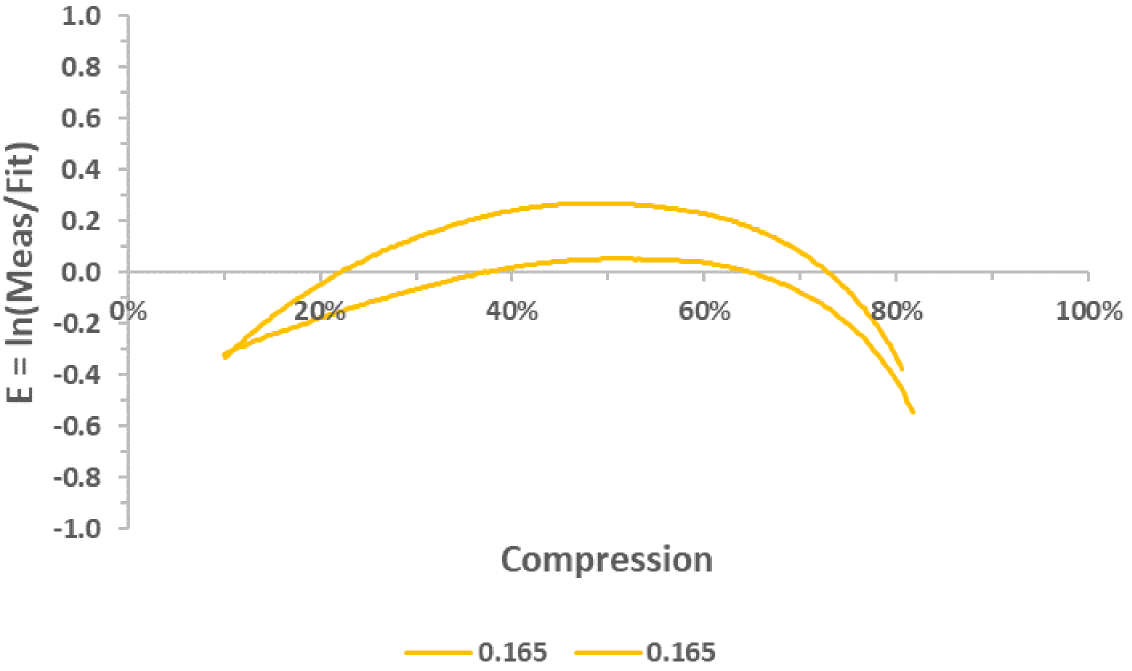


Figure C.6: Discrepancy between constitutive model and data for dynamic replicate tests

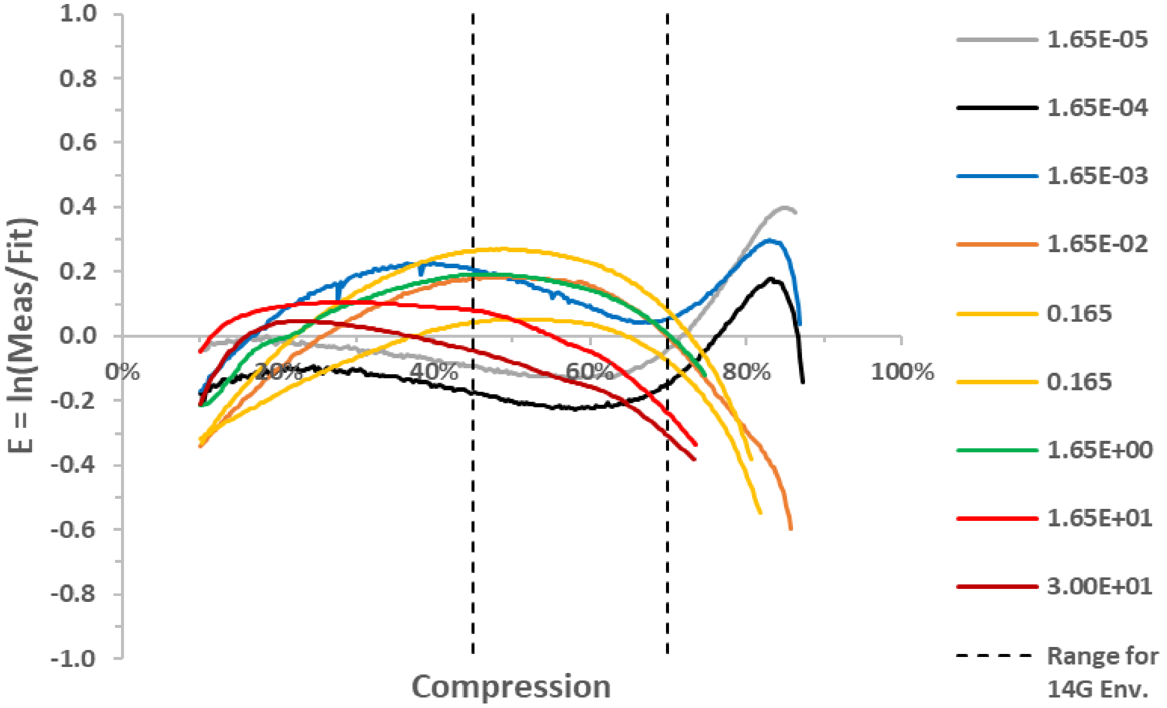


Figure C.7: Discrepancy between constitutive model and data at all rates

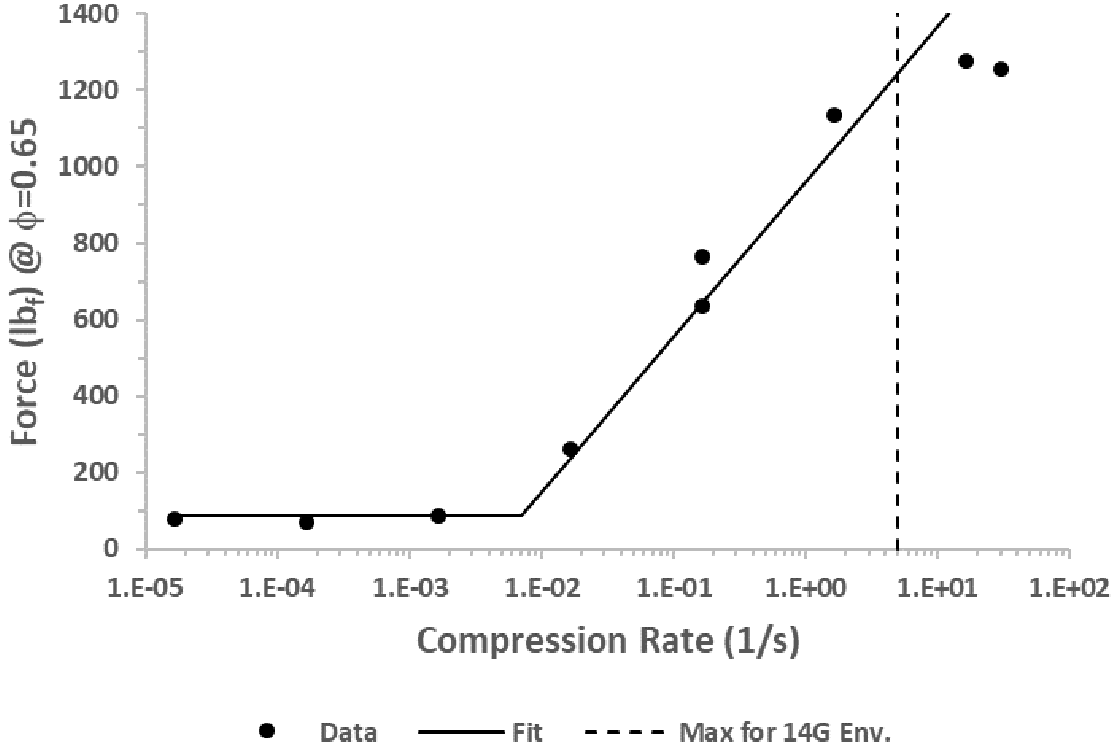


Figure C.8: Comparison of constitutive model and data for rate effects

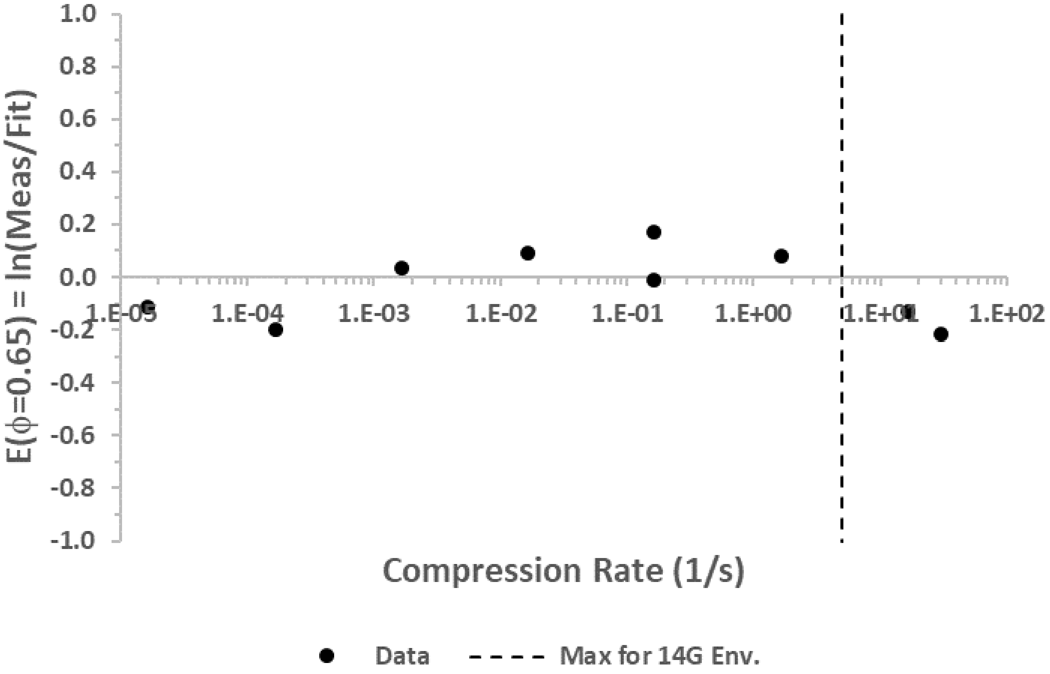


Figure C.9: Discrepancy between constitutive model and data for rate effects

Appendix D Acceptance Tests for Demonstration

The expectation is that

1. Codes are free from bugs and algorithm deficiencies,
2. Solutions will converge to the correct answer for the intended application, and
3. Simulation results are reproducible in the future.

Commercial codes are a black box to the user and regulatory communities. The software quality assurance (SQA) processes and testing strategy are considered proprietary for commercial codes. However, it should not be an act of faith that codes are free of bugs and algorithm deficiencies for specific applications. (AIAA, 2024) takes the strongest stance on this issue,

> “The users of a CFD code should also be prepared to conduct their own code verification for their specific application, or to at least audit, check, analyze, or reproduce some of the developers’ verification results or to confirm the adequacy of coverage of these results for their intended applications.”

I have advocated this perspective for regulatory applications for some time and recommend that an acceptance suite of tests be identified and referenced from code documentation or developed by the applicant. Even if relevant tests can be found in code documentation, it is preferable if the applicant reruns the tests on their own hardware and operating systems and documents the results.

Table D.1 shows the matrix of acceptance tests developed for the demonstration. Material models, initial and boundary conditions, and physics organize the required capabilities. Six regression tests were developed. The regression tests have simple analytic solutions and are re-evaluated every time the worksheet is updated. The regression tests have complete coverage of the physics and material models, one at a time, required of the computational model. The regression tests and acceptance criteria are documented in Appendix D.1.

One verification test was developed for a nearby linear system. The verification test has an analytic solution that can be used as a benchmark for the numerical solution. The observed order of accuracy is compared to the formal order of accuracy for the numerical solution scheme. The verification can be re-evaluated on demand. The verification tests have complete coverage of the physics and material models and all their interactions in a manner that closely approximates the application. The verification test and acceptance criteria are documented in Appendix D.2.

One test of sustainability was placed under configuration control. The sustainability test is the computational model for the baseline seat design (test A15008) and can be rerun on demand, with potential differences in results correlated with changes in hardware and software. A related test assesses the potential impact of roundoff errors on simulation results. The sustainability tests have complete coverage of the physics and material models and all their interactions in a manner that exactly parallels the application. The sustainability tests and acceptance criteria are documented in Appendix D.3.

Table D.1: Acceptance tests

Test Number	Test Name	Required Physics and Material Models					Coverage
		Constitutive Model	Initial Static Compression	Boundary Condition	Cushion Compression	Lumbar Load	
Regression Test Suite (RTS)							**100%**
RTS1	Constitutive model	X					
RTS2	Static compression		X				
RTS3	Seat pan $\ddot{x}$ for t*<1			X			
RTS4	Seat pan $\ddot{x}$ for t*>1			X			
RTS5	Cushion $\ddot{\varphi}$				X		
RTS6	Lumbar load no cushion					X	
Verification Test Suite (VERTS)							**100%**
VERTS1	Linear spring/mass system	X	X	X	X	X	
Sustainability Test Suite (STS)							**100%**
STS1	Baseline seat design	X	X	X	X	X	
STS2	Roundoff errors	X	X	X	X	X	

D.1 Regression test suite (RTS)

RTS1: Foam constitutive model

Date: 6/4/24

Purpose: Test implementation of the foam constitutive model (Equation 0-25) with combined quasi-static and dynamic compression. The foam constitutive model is implemented in Excel as a Function Subroutine (model).

Relevance: Dynamic events start quasi-static and transition into the rate-dependent phase for most of the transient.

Test:

The list of input parameters was chosen to produce an intuitive benchmark against which to compare the computed result.

F_0	f_c	a	b	c	$\dot{\varphi}_c$	f	$\dot{\varphi}$
0.5	1	1	1	2	1	.5	E=2.17182…

Test results and acceptance. Acceptance is machine precision for Excel.

Benchmark	Computed	$\lvert E_{rel}\rvert$	$\lvert E_{rel}\rvert$ Acceptance	Pass
3	3	0	1.0e-15	

RTS2: Initial static compression

Date: 6/4/24

Purpose: Test implementation of the initial static compression (Equation 0-6). Calculation of the initial static compression is implemented in Excel as a Function Subroutine (staticcomp).

Relevance: Initial static compression is an initial condition for the transient event.

Test:

The list of input parameters was chosen to produce an intuitive benchmark against which to compare the computed result.

F_0	f_c	a	W_0
1	1	2	4

Test results and acceptance. Acceptance is machine precision for Excel.

Benchmark	Computed	$\lvert E_{rel}\rvert$	$\lvert E_{rel}\rvert$ Acceptance	Pass
0.5	0.5	0	1.0e-15	

RTS3: Seat pan $\ddot{x}$ for t*<1

Date: 6/4/24

Purpose: Test implementation of the first phase of the FAA-prescribed boundary condition (Equation 0-10) for t*<1. Calculation of the seat pan boundary condition is implemented in Excel as a Function Subroutine (SeatPan_XDDot).

Relevance: Seat pan acceleration during phase 1 (t*<1) drives compression.

Test:

The list of input parameters was chosen to produce an intuitive benchmark against which to compare the computed result.

G	g	q	t*
10	2	60	.5

Test results and acceptance. Acceptance is machine precision for Excel.

Benchmark	Computed	$\lvert E_{rel}\rvert$	$\lvert E_{rel}\rvert$ Acceptance	Pass
5	5	1.8e-16	1.0e-15	

RTS4: Seat pan $\ddot{x}$ for t*>1

Date: 6/4/24

Purpose: Test implementation of the second phase of the FAA-prescribed boundary condition (Equation 0-11) for t*>1. Calculation of the seat pan boundary condition is implemented in Excel as a Function Subroutine (SeatPan_XDDot).

Relevance: Seat pan acceleration during phase 1 (t*<1) drives compression.

Test:

The list of input parameters was chosen to produce an intuitive benchmark against which to compare the computed result.

G	g	q	t*
10	2	60	.5

Test results and acceptance. Acceptance is machine precision for Excel.

Benchmark	Computed	$\lvert E_{rel} \rvert$	$\lvert E_{rel} \rvert$ Acceptance	Pass
5	5	1.8e-16	1.0e-15	

RTS5: Cushion $\ddot{\varphi}$

Date: 6/4/24

Purpose: Test implementation of Equation 0-32 which describes the dynamics of foam compression. The foam compression equation is implemented in Excel as Function Subroutine (PhiDDot).

Relevance: The foam compression equation is a coupled equation in the lumbar load calculation.

Test:

The list of input parameters was chosen to produce an intuitive benchmark against which to compare the computed result.

$\ddot{x}$	g	F_0	W_0	W_{UB}	H
11	1	11	1	10	2

Test results and acceptance. Acceptance is machine precision for Excel.

Benchmark	Computed	$\lvert E_{rel} \rvert$	$\lvert E_{rel} \rvert$ Acceptance	Pass
5	5	0	1.0e-15	

RTS6: Lumbar load no cushion

Date: 6/4/24

Purpose: Test implementation of Equation 0-24 which describes lumbar load for seats without cushions. The calculation of lumbar loads for seats without cushions is implemented as a Function Subroutine (LLNoCush).

Relevance: There are tests in the validation hierarchy for seats without cushions.

Test:

The list of input parameters was chosen to produce an intuitive benchmark against which to compare the computed result.

G	q	W_{UT}
10	60	100

Test results and acceptance. Acceptance is machine precision for Excel.

Benchmark	Computed	$\lvert E_{rel} \rvert$	$\lvert E_{rel} \rvert$ Acceptance	Pass
500	500	2.3e-16	1.0e-15	

D.2 Verification test suite (VERTS)

VERTS1: Linear spring/mass system

Date: 6/4/24

Purpose: Test implementation of the numerical algorithm for seats with cushions (Equations 0-10, 0-31, 0-33, 0-34, and 0-35). This test compares the numerical algorithm's convergence (order of accuracy) to an analytic benchmark for a nearby problem. The calculation of lumbar loads for seats with cushions is implemented in Excel as a Function Subroutine (ComputedMaxLoad).

Relevance: This test is "nearby" to the application and differs only in that cushion is linear without damping. All other equations, initial conditions, and boundary conditions are the same. Consequently, this test exercises the interactions of all equations in the same manner as the application.

Test:

The test involves a linear spring-mass system without damping. The foam constitutive model (Equation 0-25) is modified to accept a linear spring,

$$F(\varphi) = F_m \frac{\varphi}{\varphi_c}, \qquad \textbf{D-1}$$

where $F_m = 2000$ lb$_f$ and $f_c = 1$. The lumbar load has an analytic solution when the cushion response is linear and without damping,

$$L = L_{nc}\left[t^* - \frac{1}{\pi}\sin \pi t^*\right], \qquad \textbf{D-2}$$

where L_{nc} is the lumbar load for a seat with no cushion (Equation 0-24) and

$$\pi = \omega t_{rise} \sqrt{\frac{gF_m t_{rise}^2}{H\varphi_c W_{UB}}}\ . \qquad \textbf{D-3}$$

The analytic solution is shown in Figure D.1. Note that the presence of a cushion enhances lumbar loads by t* = 1. Figure D.2 shows the convergence of lumbar loads at t* = 1.0 to the analytic solution. Figure D.3 shows convergence of relative numerical errors when lumbar loads are referenced to the exact solution at t* = 1.0. The slope of the line is the observed order of accuracy, which can be computed from

$$p = \frac{\ln \dfrac{L_c - L_m}{L_m - L_e}}{\ln r}, \qquad \textbf{D-4}$$

where L_e is the exact solution and where L_c and L_m are the solutions on the coarse grid, N_c, and medium grid, N_m. The refinement ratio, r, is given by

$$r = \frac{N_m}{N_c}\ . \qquad \textbf{D-5}$$

The list of input parameters is given here.

N_c	N_m	L_c	L_m	L_e
4000	8000	1194.822750	1194.522571	1194.222891

Test results and acceptance.

p_{obs}	p_f	$\lvert E_p \rvert$	$\lvert E_p \rvert$ Acceptance	Pass
1.0012001	1.0	.00120	0.10	

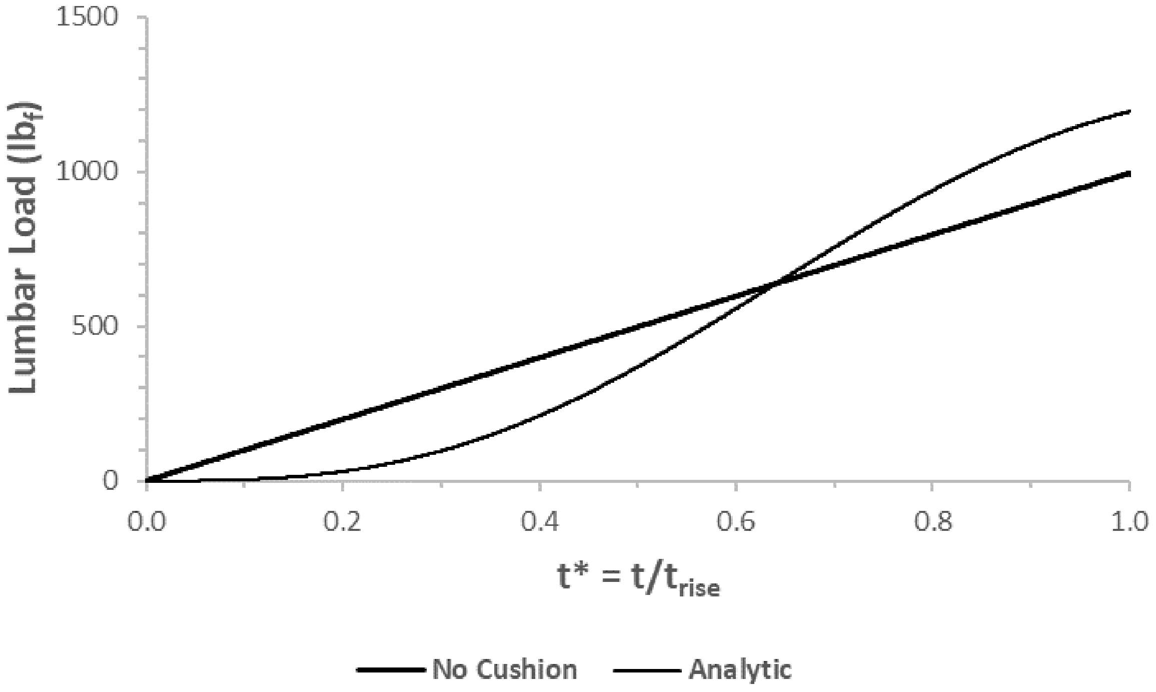


Figure D.1: Analytic solution to nearby problem

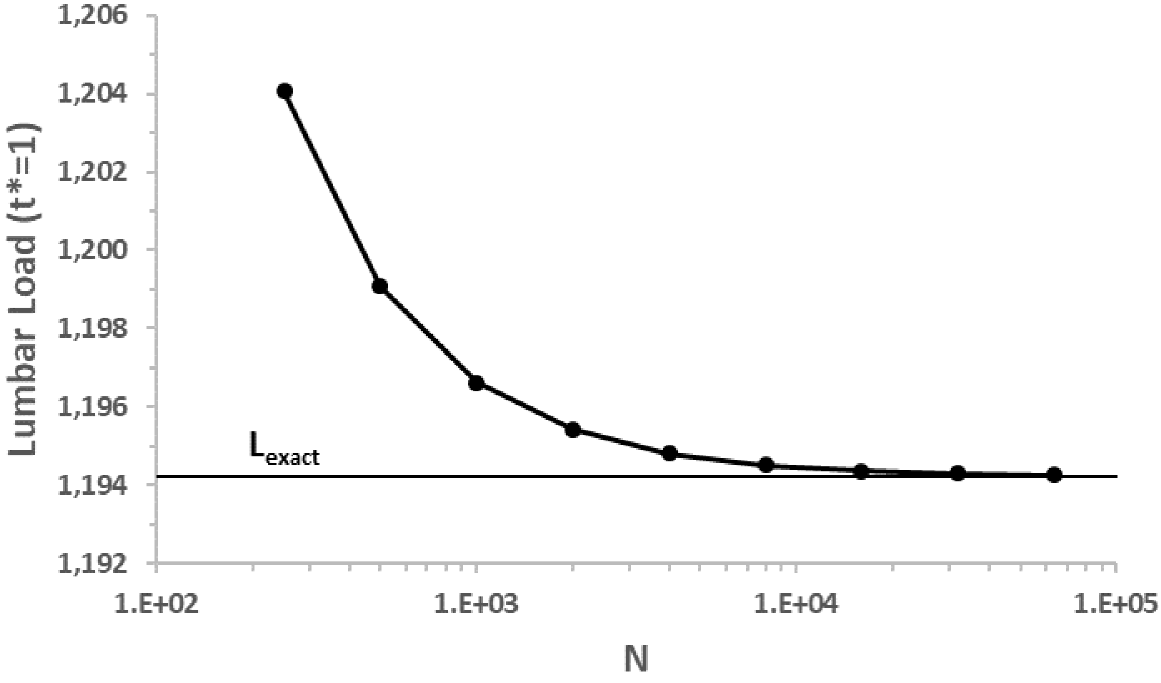


Figure D.2: Convergence lumbar loads to exact solution

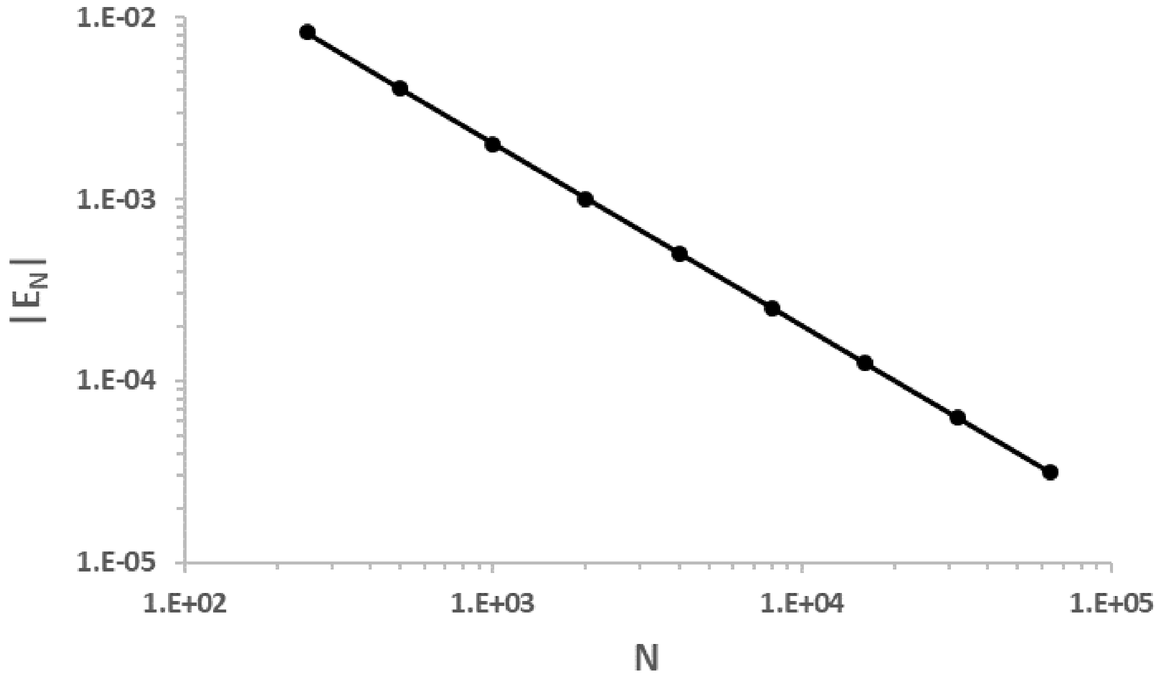


Figure D.3: Convergence of relative numerical errors

D.3 Sustainability test suite (STS)

STS1: Baseline seat design (test A15008)

Date: 6/4/24

Purpose: Test the reproducibility of results with software updates.

Relevance: Software updates occur frequently, and regulatory decisions should be insensitive to software versions.

Test:

Table D.2 formally records the version history for the platform and software used in this project. The yellow highlighting denotes changes from the previously recorded version. The platform remains unchanged over the project's history. Changes in the operating system and Excel occur routinely.

The computational model for test A15008 (baseline seat design) is placed under configuration control and rerun on demand at various times to assess possible changes in simulation results that might be correlated with changes in platform and software shown in Table D.2.

Table D.3 assesses the sustainability error (new result referenced to the previous result). Acceptance is taken as $|E_s|_{accept} < 6.67e\text{-}4$, corresponding to 1 lb_f in 1500 lb_f. Comments concerning significant changes are highlighted in yellow. A surprisingly substantial change was recorded on 5/5/24. Excel technical support says there are many reasons why roundoff errors will propagate differently for the same model rerun at a different time. Hence, the expectation is that results are not exactly reproducible. However, the magnitude of this difference suggests user error. Sustainability will be reassessed periodically until project completion. The magnitude of the change is < 1 $_{\text{lbf}}$ in 1500 lb$_f$ and can be ignored for this demonstration.

An insignificant change was recorded on 6/4/24. This change is close to machine precision and reflects a different rollup of roundoff errors.

STS2: Roundoff errors

Date: 6/5/24

Purpose: Assess sensitivity to roundoff errors.

Relevance: Excel technical support notes that there are many reasons why roundoff errors will propagate differently for the same model rerun at a different time.

Test:

The sensitivity of simulation results for the baseline seat design (test A15008) is assessed by comparing single-precision results with double-precision results. Acceptance is taken as 6.67e-4, which corresponds to 1 lb$_f$ in 1500 lb$_f$.

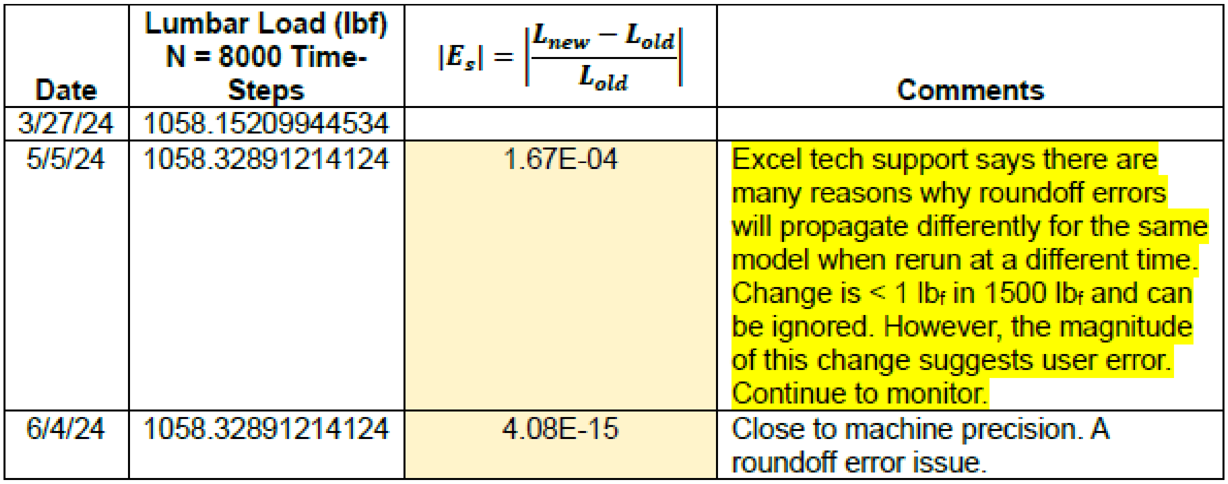
Load Double Precision	Load Single Precision	$\lvert E_{rel} \rvert$	$\lvert E_{rel} \rvert_{accept}$	Pass
1058.32891214124	1058.32885742187	5.17e-8	6.67e-4	

The sensitivity to roundoff errors is smaller than observed in STS1, calling into question the rollup of roundoff errors as an explanation. Keep in mind that the sensitivity, 5.17e-8, is assessed using the double precision solution as the benchmark and would be much less in a double precision simulation if it could be assessed.

Table D.2: Version history of hardware and software

Date	Platform	Windows OS	MS Excel
3/27/24	Dell XPS 15 9500 Intel® Core™ i9-10885H CPU @ 2.4GHz, 8 Core(s), 16 Logical Processors, 64GB RAM	Windows 11 Pro Version 23H2 (OS Build 22631.3296)	Microsoft® Excel® for Microsoft 365 MSO (Version 2403 Build 16.0.17425.20070) 64-bit
5/5/24	Dell XPS 15 9500 Intel® Core™ i9-10885H CPU @ 2.4GHz, 8 Core(s), 16 Logical Processors, 64GB RAM	Windows 11 Pro Version 23H2 (OS Build 22631.3447)	Microsoft® Excel® for Microsoft 365 MSO (Version 2403 Build 17425.20176) 64-bit
6/4/24	Dell XPS 15 9500 Intel® Core™ i9-10885H CPU @ 2.4GHz, 8 Core(s), 16 Logical Processors, 64GB RAM	Windows 11 Pro Version 23H2 (OS Build 22631.3593)	Microsoft® Excel® for Microsoft 365 MSO (Version 2405 Build 17628.20110) 64-bit

Table D.3: Assessment of model sustainability



Date	Lumbar Load (lbf) N = 8000 Time-Steps	$\lvert E_s\rvert = \left\lvert\frac{L_{new} - L_{old}}{L_{old}}\right\rvert$	Comments
3/27/24	1058.15209944534		
5/5/24	1058.32891214124	1.67E-04	Excel tech support says there are many reasons why roundoff errors will propagate differently for the same model when rerun at a different time. Change is < 1 lb_f in 1500 lb_f and can be ignored. However, the magnitude of this change suggests user error. Continue to monitor.
6/4/24	1058.32891214124	4.08E-15	Close to machine precision. A roundoff error issue.

Appendix E Database of Sled Tests

E.1 Sled tests spanning a wide range of environments and seat designs

Table E.1 presents a database of 105 sled tests where lumbar load was a primary measurement. In some cases, the initial static compression was also reported, and 105 tests were reported in eight references published over 22 years. Tests were conducted by two organizations: the Civil Aerospace Medical Institute (CAMI) and the National Institute for Aviation Research (NIAR). Test environments ranged from 9G to 19G. In recent years, the FAA-Hybrid III ATD has replaced the Hybrid II ATD. This database is restricted to research tests characterized by single seats attached to rigid frames. Tests with and without cushions are included. When present, cushions were monolithic and ranged in thickness from 1.0” to 10”. Four classes of cushion material, each with multiple options, span a wide range of stiffness and rate sensitivity.

The database serves two purposes. First, data alone can provide insight into the expected sensitivity of lumbar loads to environments and design choices. This evidence provides rationale for ePIRT assessments.

Second, the database could serve as a resource for validating the simulation conceptual model for a wide range of environments and designs. Having appropriate material data for each of the many cushion foams is a requirement; and currently, appropriate data only exist in the public domain for CF42 (AC). Consequently, only data for CF42 (AC) will be used to validate the simulation conceptual model in this report.

Table E.1: Database of sled tests

Identifying Information				Environment			ATD	Cushion		Test Results	
Entry	Reference	Test Facility	Test Number	G	T_{rise} (ms)	Angle (deg)	Type	Material	Thick (in)	Lumbar Load (lb_f)	Init Static Comp
1	(Adams et al., 2003)	NIAR	01144-001	15	60	30	Hybrid II	None	0.00	1022	
2	(Adams et al., 2003)	NIAR	01144-002	15	60	30	Hybrid II	None	0.00	1080	
3	(Adams et al., 2003)	NIAR	01144-003	15	60	30	Hybrid II	HR30	2.50	1646	
4	(Adams et al., 2003)	NIAR	01144-004	15	60	30	Hybrid II	HR30	2.50	1988	
5	(Adams et al., 2003)	NIAR	01267-001a	15	60	30	Hybrid II	None	0.00	1163	
6	(Adams et al., 2003)	NIAR	01267-001b	15	60	30	Hybrid II	None	0.00	1103	
7	(Adams et al., 2003)	NIAR	01267-002a	15	60	30	Hybrid II	None	0.00	1265	
8	(Adams et al., 2003)	NIAR	01267-002b	15	60	30	Hybrid II	None	0.00	1299	
9	(Adams et al., 2003)	NIAR	01267-007a	15	60	30	Hybrid II	DAX26	4.00	1814	
10	(Adams et al., 2003)	NIAR	01267-007b	15	60	30	Hybrid II	DAX26	4.00	1752	
11	(Adams et al., 2003)	NIAR	01267-008a	15	60	30	Hybrid II	DAX26	4.00	2068	
12	(Adams et al., 2003)	NIAR	01267-008b	15	60	30	Hybrid II	DAX26	4.00	2066	
13	(Adams et al., 2003)	NIAR	01267-005a	15	60	30	Hybrid II	DAX26	6.00	2128	
14	(Adams et al., 2003)	NIAR	01267-005b	15	60	30	Hybrid II	DAX26	6.00	2010	
15	(Adams et al., 2003)	NIAR	01267-006a	15	60	30	Hybrid II	DAX20	6.00	1960	
16	(Adams et al., 2003)	NIAR	01267-006b	15	60	30	Hybrid II	DAX26	6.00	1922	
17	(Adams et al., 2003)	NIAR	01267-009a	15	60	30	Hybrid II	DAX26	10.00	2044	
18	(Adams et al., 2003)	NIAR	01267-009b	15	60	30	Hybrid II	DAX26	10.00	2027	
19	(Adams et al., 2003)	NIAR	01267-010a	15	60	30	Hybrid II	DAX26	10.00	1856	
20	(Adams et al., 2003)	NIAR	01267-010b	15	60	30	Hybrid II	DAX26	10.00	1937	
21	(Adams et al., 2003)	NIAR	01267-011a	15	60	30	Hybrid II	DAX55	4.00	1932	
22	(Adams et al., 2003)	NIAR	01267-011b	15	60	30	Hybrid II	DAX55	4.00	1835	
23	(Adams et al., 2003)	NIAR	01267-012a	15	60	30	Hybrid II	DAX55	4.00	2212	

24	(Adams et al., 2003)	NIAR	01267-012b	15	60	30	Hybrid II	DAX55	4.00	2104	
25	(Adams et al., 2003)	NIAR	01267-003a	15	60	30	Hybrid II	DAX90	8.00	1416	
26	(Adams et al., 2003)	NIAR	01267-003b	15	60	30	Hybrid II	DAX90	8.00	1401	
27	(Adams et al., 2003)	NIAR	01267-004a	15	60	30	Hybrid II	DAX90	8.00	1469	
28	(Adams et al., 2003)	NIAR	01267-004b	15	60	30	Hybrid II	DAX90	8.00	1296	
29	(Olivares, 2013)	NIAR	07324-10	19	50	30	Hybrid II	None	0.00	1410	
30	(Olivares, 2013)	NIAR	07324-11	19	50	30	Hybrid II	None	0.00	1757	
31	(Olivares, 2013)	NIAR	07324-12	19	50	30	Hybrid II	None	0.00	1693	
32	(Olivares, 2013)	NIAR	07324-30	19	50	30	Hybrid II	None	0.00	1120	
33	(Olivares, 2013)	NIAR	07324-31	19	50	30	Hybrid II	None	0.00	1161	
34	(Olivares, 2013)	NIAR	07324-13	19	50	30	FAA-Hybrid III	None	0.00	1713	
35	(Olivares, 2013)	NIAR	07324-14	19	50	30	FAA-Hybrid III	None	0.00	1736	
36	(Olivares, 2013)	NIAR	07324-15	19	50	30	FAA-Hybrid III	None	0.00	1798	
37	(Taylor, Moorcroft, et al., 2017)	CAMI	A12013	9	100	30	Hybrid II	CF47	1.00	580	
38	(Taylor, Moorcroft, et al., 2017)	CAMI	A12031	9	100	30	Hybrid II	CF47	1.00	553	
39	(Taylor, Moorcroft, et al., 2017)	CAMI	A12028	9	100	30	FAA-Hybrid III	CF47	1.00	519	
40	(Taylor, Moorcroft, et al., 2017)	CAMI	A12011	14	80	30	Hybrid II	CF47	1.00	909	
41	(Taylor, Moorcroft, et al., 2017)	CAMI	A12032	14	80	30	Hybrid II	CF47	1.00	1040	
42	(Taylor, Moorcroft, et al., 2017)	CAMI	A12029	14	80	30	FAA-Hybrid III	CF47	1.00	874	

43	(Taylor, Moorcroft, et al., 2017)	CAMI	A12012	19	50	30	Hybrid II	CF47	1.00	1860	
44	(Taylor, Moorcroft, et al., 2017)	CAMI	A12014	19	50	30	Hybrid II	CF47	1.00	1827	
45	(Gowdy et al., 1999)	CAMI	A96041	15	80	30	Hybrid II	CF47	1.00	1277	
46	(Gowdy et al., 1999)	CAMI	A96042	15	80	30	Hybrid II	CF47	1.00	1270	
47	(Gowdy et al., 1999)	CAMI	A96043	15	80	30	Hybrid II	CF47	1.00	1238	
48	(Gowdy et al., 1999)	CAMI	A98032	15	80	30	FAA-Hybrid III	CF47	1.00	1236	
49	(Gowdy et al., 1999)	CAMI	A98033	15	80	30	FAA-Hybrid III	CF47	1.00	1258	
50	(Gowdy et al., 1999)	CAMI	A99010	15	80	30	FAA-Hybrid III	CF47	1.00	1292	
51	(DeWeese et al., 2021)	CAMI	A09001	14	80	30	Hybrid II	DAX90	4.60	1042	
52	(DeWeese et al., 2021)	CAMI	A09004	14	80	30	Hybrid II	DAX90	4.60	979	
53	(DeWeese et al., 2021)	CAMI	A09005	14	80	30	Hybrid II	DAX47	4.50	1433	
54	(DeWeese et al., 2021)	CAMI	A09006	14	80	30	Hybrid II	DAX47	4.50	1360	
55	(DeWeese et al., 2021)	CAMI	A09007	14	80	30	Hybrid II	DAX47	4.50	1349	
56	(DeWeese et al., 2021)	CAMI	A11024	14	80	30	Hybrid II	DAX26	4.00	1292	
57	(DeWeese et al., 2021)	CAMI	A11025	14	80	30	Hybrid II	DAX26	4.00	1270	
58	(DeWeese et al., 2021)	CAMI	A11026	14	80	30	Hybrid II	DAX26	4.00	1229	
59	(DeWeese et al., 2021)	CAMI	A10009	14	80	30	Hybrid II	AF4050	4.50	1796	
60	(DeWeese et al., 2021)	CAMI	A10010	14	80	30	Hybrid II	AF4050	4.50	1873	

61	(DeWeese et al., 2021)	CAMI	A10011	14	80	30	Hybrid II	AF4050	4.50	1993	
62	(DeWeese et al., 2021)	CAMI	A10002	14	80	30	Hybrid II	AF4050	3.50	1599	
63	(DeWeese et al., 2021)	CAMI	A10003	14	80	30	Hybrid II	AF4050	3.50	1865	
64	(DeWeese et al., 2021)	CAMI	A10004	14	80	30	Hybrid II	AF4050	3.50	1941	
65	(DeWeese et al., 2021)	CAMI	A10005	14	80	30	Hybrid II	AF4050	3.50	1908	
66	(DeWeese et al., 2021)	CAMI	A10006	14	80	30	Hybrid II	AF4050	2.00	1526	
67	(DeWeese et al., 2021)	CAMI	A10007	14	80	30	Hybrid II	AF4050	2.00	1621	
68	(DeWeese et al., 2021)	CAMI	A10008	14	80	30	Hybrid II	AF4050	2.00	1464	
69	(Pellettiere et al., 2019)	NIAR	06165-5	14	80	30	Hybrid II	None	0.00	817	
70	(Pellettiere et al., 2019)	NIAR	06165-6	14	80	30	Hybrid II	None	0.00	921	
71	(Pellettiere et al., 2019)	NIAR	06165-25	14	80	30	Hybrid II	None	0.00	800	
72	(Pellettiere et al., 2019)	NIAR	06165-26	14	80	30	Hybrid II	None	0.00	912	
73	(Pellettiere et al., 2019)	NIAR	06165-7	14	80	30	FAA-Hybrid III	None	0.00	971	
74	(Pellettiere et al., 2019)	NIAR	06165-8	14	80	30	FAA-Hybrid III	None	0.00	972	
75	(Pellettiere et al., 2019)	NIAR	06165-28	14	80	30	FAA-Hybrid III	None	0.00	906	
76	(Hooper & Henderson, 2005)	CAMI	A04061	14	80	30	Hybrid II	AF5565	2.00	1856	0.032
77	(Hooper & Henderson, 2005)	CAMI	A04035	14	80	30	Hybrid II	DAX26	2.00	1173	0.628
78	(Hooper & Henderson, 2005)	CAMI	A04036	14	80	30	Hybrid II	DAX26	2.00	1187	0.679

79	(Hooper & Henderson, 2005)	CAMI	A04057	14	80	30	Hybrid II	DAX90	2.00	1233	0.188
80	(Hooper & Henderson, 2005)	CAMI	A04065	14	80	30	Hybrid II	AF5565	3.25	1779	0.050
81	(Hooper & Henderson, 2005)	CAMI	A04055	14	80	30	Hybrid II	DAX26	3.25	1200	0.606
82	(Hooper & Henderson, 2005)	CAMI	A04062	14	80	30	Hybrid II	DAX90	3.25	1195	0.172
83	(Hooper & Henderson, 2005)	CAMI	A04063	14	80	30	Hybrid II	DAX90	3.25	1172	0.185
84	(Hooper & Henderson, 2005)	CAMI	A04069	14	80	30	Hybrid II	AF5565	4.50	1699	0.081
85	(Hooper & Henderson, 2005)	CAMI	NAwa2-450	14	80	30	Hybrid II	AF4050	4.50	2137	
86	(Hooper & Henderson, 2005)	CAMI	A04058	14	80	30	Hybrid II	DAX26	4.50	1254	0.610
87	(Hooper & Henderson, 2005)	CAMI	A04066	14	80	30	Hybrid II	DAX90	4.50	1156	0.178
88	(Taylor, DeWeese, et al., 2017)	CAMI	A15005	14	80	30	FAA-Hybrid III	CF45	2.00	970	0.425
89	(Taylor, DeWeese, et al., 2017)	CAMI	A15006	14	80	30	FAA-Hybrid III	CF45	2.00	960	0.330
90	(Taylor, DeWeese, et al., 2017)	CAMI	A15003	19	50	30	FAA-Hybrid III	CF45	2.00	1509	0.425
91	(Taylor, DeWeese, et al., 2017)	CAMI	A15004	19	50	30	FAA-Hybrid III	CF45	2.00	1604	0.445
92	(Taylor, DeWeese, et al., 2017)	CAMI	A15018	14	80	30	FAA-Hybrid III	CF45	4.00	956	0.378
93	(Taylor, DeWeese, et al., 2017)	CAMI	A15017	14	80	30	FAA-Hybrid III	CF45	4.00	970	0.433
94	(Taylor, DeWeese, et al., 2017)	CAMI	A15014	19	50	30	FAA-Hybrid III	CF45	4.00	1514	0.418

95	(Taylor, DeWeese, et al., 2017)	CAMI	A15013	19	50	30	FAA-Hybrid III	CF45	4.00	1470	0.435
96	(Taylor, DeWeese, et al., 2017)	CAMI	A15007	14	80	30	FAA-Hybrid III	CF42	2.00	983	0.515
97	(Taylor, DeWeese, et al., 2017)	CAMI	A15008	14	80	30	FAA-Hybrid III	CF42 (AC)	2.00	1048	0.550
98	(Taylor, DeWeese, et al., 2017)	CAMI	A15001	19	50	30	FAA-Hybrid III	CF42	2.00	1694	0.520
99	(Taylor, DeWeese, et al., 2017)	CAMI	A15002	19	50	30	FAA-Hybrid III	CF42 (AC)	2.00	1660	0.550
100	(Taylor, DeWeese, et al., 2017)	CAMI	A15019	14	80	30	FAA-Hybrid III	CF42	4.00	1100	0.433
101	(Taylor, DeWeese, et al., 2017)	CAMI	A15020	14	80	30	FAA-Hybrid III	CF42 (AC)	4.00	1153	0.448
102	(Taylor, DeWeese, et al., 2017)	CAMI	A15016	19	50	30	FAA-Hybrid III	CF42	4.00	1771	0.440
103	(Taylor, DeWeese, et al., 2017)	CAMI	A15015	19	50	30	FAA-Hybrid III	CF42 (AC)	4.00	1962	0.505
104	(Taylor, DeWeese, et al., 2017)	CAMI	A15022	19	50	30	FAA-Hybrid III	CF42 (AC)	4.00	1975	
105	(Taylor, DeWeese, et al., 2017)	CAMI	A15021	19	50	30	FAA-Hybrid III	CF42 (AC)	4.00	1951	

E.2 Precision uncertainties in sled tests

Section 0 presents and discusses uncertainties associated with testing. The appendix addresses the assessment of precision uncertainties, which include:

1. Repeatability uncertainty, i.e., replicate testing within an organization.
2. Reproducibility uncertainty: Replicate testing by different organizations.
3. Sample-to-Sample (S2S) uncertainty: Testing with different but nominally identical samples.
4. Gauge-to-Gauge G2G) uncertainty: Testing with different but nominally identical gauges.

Often, we pool data from different references found in the literature to increase the amount of data available for analysis. This type of meta-analysis violates the formalism of repeatability and reproducibility; however, pooled data can be more useful. Pooled data is more likely to encompass truth when the sources of precision uncertainty are represented more completely.

Table E.2 summarizes a database of 101 tests to assess precision uncertainty for aircraft seat testing. The database was extracted from 9 references and includes 42 series of replicate tests for 101 tests. Typically, only 2 or 3 replicate tests are represented in each series, but 4 replicates are represented twice, and 5 replicates are represented once.

All sources of precision uncertainty are well represented in the database, which has entries

- spanning 22 years,
- with at least 3 testing organizations,
- with 9G, 14G, 15G, and 19G environments represented,
- with both Hybrid II and FAA-Hybrid III ATDs represented,
- seats with and without foam cushions and with
- four classes of cushion material ranging from 1” to 10” when cushions were present.

The replication error,

$$E = \ln \frac{L}{L_{50}}, \qquad \textbf{E-1}$$

is computed for each replicate test in a given series. Here, L_{50} is the computed median for that series. The replication errors are plotted in Figure E.1 with three affinity groupings: seats with rigid frames and no cushions, seats with rigid frames and cushions, and seats more representative of real aircraft seats and cushions. Seats with and without cushions are indistinguishable, suggesting that the primary source of precision uncertainty is associated with the seating of the ATD. There is little to distinguish the three groups; consequently, the data can be pooled. The blue curve in Figure E.2 shows the empirical distribution of precision uncertainties for the 101 tests in the database, and the red curve shows that the parametric Laplace distribution well represents the data,

$$E = \ln \frac{L}{L_{50}} = Laplace(0{,}0.058197)\,, \qquad \textbf{E-2}$$

and is representative of the precision uncertainty of any tested or untested design that falls within the broad range of applicability.

Table E.2: Database for the assessment of precision uncertainty in sled tests

Identifying Information			Environ		ATD	Cushion		Lumbar Load (lb$_f$)						$E = \ln \frac{L}{L_{50}}$				
Reference	**Year**	**Facility**	**G**	**t_{rise}**	**Type**	**Mat**	**H(in)**	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 4**	**Rep 5**	**L_{50}**	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 4**	**Rep 5**
(Adams et al., 2003)	2003	NIAR	15	60	HII	None	0.00	1022	1080				1051	-0.0281	0.0274			
(Adams et al., 2003)	2003	NIAR	15	60	HII	None	0.00	1163	1103				1133	0.0261	-0.0268			
(Adams et al., 2003)	2003	NIAR	15	60	HII	None	0.00	1265	1299				1282	-0.0133	0.0132			
(Olivares, 2013)	2013	NIAR	19	50	HII	None	0.00	1410	1757	1693	1120	1161	1410	0.0000	0.2200	0.1829	-0.2303	-0.1943
(Olivares, 2013)	2013	NIAR	19	50	FAA-HIII	None	0.00	1713	1736	1798			1736	-0.0133	0.0000	0.0351		
(Pelletiere et al., 2019)	2019	NIAR	14	80	HII	None	0.00	817	921	800	912		865	-0.0565	0.0633	-0.0775	0.0535	
(Pelletiere et al., 2019)	2019	NIAR	14	80	FAA-HIII	None	0.00	971	972	906			971	0.0000	0.0010	-0.0693		
(Adams et al., 2003)	2003	NIAR	15	60	HII	HR30	2.50	1646	1988				1817	-0.0986	0.0898			
(Adams et al., 2003)	2003	NIAR	15	60	HII	DAX26	4.00	1814	1752				1783	0.0172	-0.0175			
(Adams et al., 2003)	2003	NIAR	15	60	HII	DAX26	4.00	2068	2066				2067	0.0005	-0.0005			
(Adams et al., 2003)	2003	NIAR	15	60	HII	DAX26	6.00	2128	2010				2069	0.0281	-0.0289			
(Adams et al., 2003)	2003	NIAR	15	60	HII	DAX20	6.00	1960	1922				1941	0.0097	-0.0098			
(Adams et al., 2003)	2003	NIAR	15	60	HII	DAX26	10.00	2044	2027				2036	0.0042	-0.0042			
(Adams et al., 2003)	2003	NIAR	15	60	HII	DAX26	10.00	1856	1937				1897	-0.0216	0.0211			
(Adams et al., 2003)	2003	NIAR	15	60	HII	DAX55	4.00	1932	1835				1884	0.0254	-0.0261			
(Adams et al., 2003)	2003	NIAR	15	60	HII	DAX55	4.00	2212	2104				2158	0.0247	-0.0253			
(Adams et al., 2003)	2003	NIAR	15	60	HII	DAX90	8.00	1416	1401				1409	0.0053	-0.0053			
(Adams et al., 2003)	2003	NIAR	15	60	HII	DAX90	8.00	1469	1296				1383	0.0607	-0.0646			

(Taylor, Moorcroft, et al., 2017)	2017	CAM I	19	50	H II	CF47	1.00	1860	1827				1844	0.0089	-0.0090			
(Gowdy et al., 1999)	1999	CAM I	15	100	H II	CF47	1.00	1277	1270	1238			1270	0.0055	0.0000	-0.0255		
(Gowdy et al., 1999)	1999	CAM I	15	100	FAA-H III	CF47	1.00	1236	1258	1292			1258	-0.0176	0.0000	0.0267		
(DeWeese et al., 2021)	2021	CAM I	14	80	H II	DAX90	4.60	1042	979				1011	0.0307	-0.0317			
(DeWeese et al., 2021)	2021	CAM I	14	80	H II	DAX47	4.50	1433	1360	1349			1360	0.0523	0.0000	-0.0081		
(DeWeese et al., 2021)	2021	CAM I	14	80	H II	DAX26	4.00	1292	1270	1229			1270	0.0172	0.0000	-0.0328		
(DeWeese et al., 2021)	2021	CAM I	14	80	H II	AF4050	4.50	1796	1873	1993			1873	-0.0420	0.0000	0.0621		
(DeWeese et al., 2021)	2021	CAM I	14	80	H II	AF4050	3.50	1599	1865	1941	1908		1887	-0.1653	-0.0115	0.0285	0.0113	
(DeWeese et al., 2021)	2021	CAM I	14	80	H II	AF4050	2.00	1526	1621	1464			1526	0.0000	0.0604	-0.0415		
(Hooper & Henderson, 2005)	2005	CAM I	14	80	H II	DAX26	2.00	1173	1187				1180	-0.0057	0.0057			
(Hooper & Henderson, 2005)	2005	CAM I	14	80	H II	DAX90	3.25	1195	1172				1183	0.0097	-0.0098			
(Taylor, Moorcroft, et al., 2017)	2017	CAM I	9	100	H II	CF47	1.00	580	553				567	0.0236	-0.0241			
(Taylor, Moorcroft, et al., 2017)	2017	CAM I	14	80	H II	CF47	1.00	909	1040				975	-0.0696	0.0651			
(Taylor, Moorcroft, et al., 2017)	2017	CAM I	19	50	FAA-H III	CF42	4.00	1975	1951				1963	0.0061	-0.0061			
(Soltis & Forest, 1999)	1999	CAM I	14	80	H II	Real		1337	1166				1252	0.0662	-0.0709			
(Soltis &	1999	Other	14	80	H II	Real		1596	1398				1497	0.0639	-0.0683			

Forest, 1999)																		
(Soltis & Forest, 1999)	1999	NIAR	14	80	H II	Real		1194	1139				1166	0.0231	-0.0237			
(DeWeese et al., 2021)	2021	CAMI	14	80	H II	Real	4.50	1178	1238	1194			1194	-0.0135	0.0362	0.0000		
(DeWeese et al., 2021)	2021	CAMI	14	80	H II	Real	4.00	1640	1599	1750			1640	0.0000	-0.0253	0.0649		
(DeWeese et al., 2021)	2021	CAMI	14	80	H II	Real	4.00	1362	669				1016	0.2936	-0.4174			
(Bhonge et al., 2019)	2019	CAMI	14	80	H II	Real		1617	1503				1560	0.0359	-0.0372			
(Bhonge et al., 2019)	2019	CAMI	14	80	FAA-H III	Real		1590	1622				1606	-0.0100	0.0099			
(Bhonge et al., 2019)	2019	Other	14	80	H II	Real		1202	1217				1210	-0.0062	0.0062			
(Bhonge et al., 2019)	2019	Other	14	80	FAA-H III	Real		1551	1628				1590	-0.0245	0.0239			

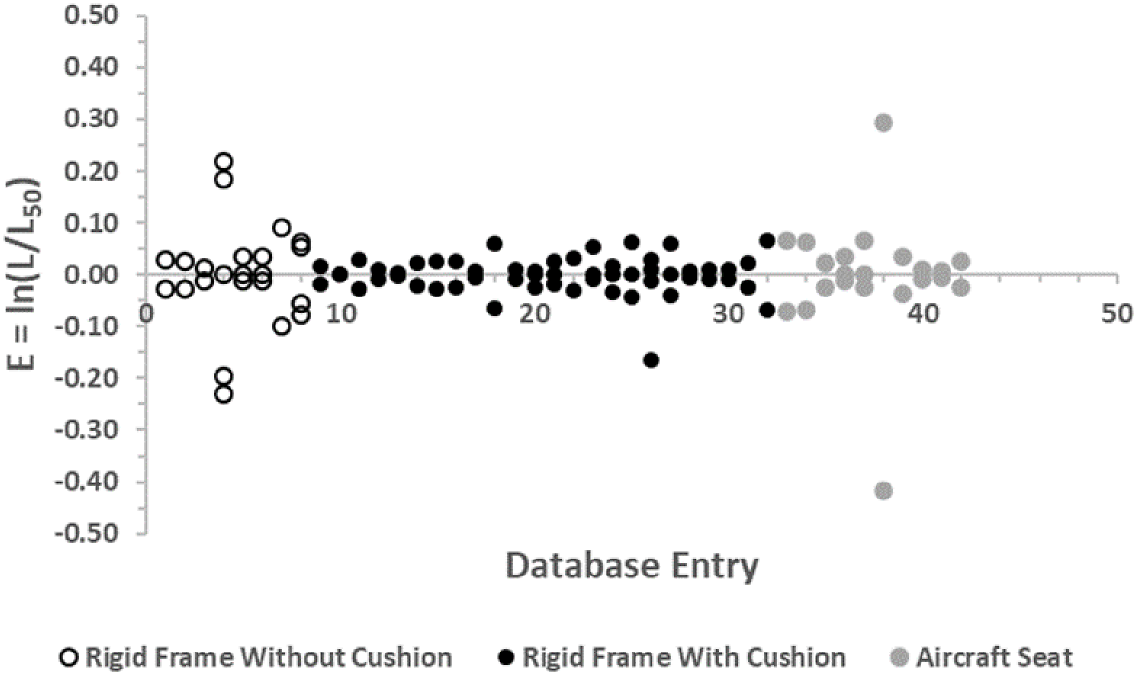


Figure E.1: Precision uncertainty can be pooled across all designs

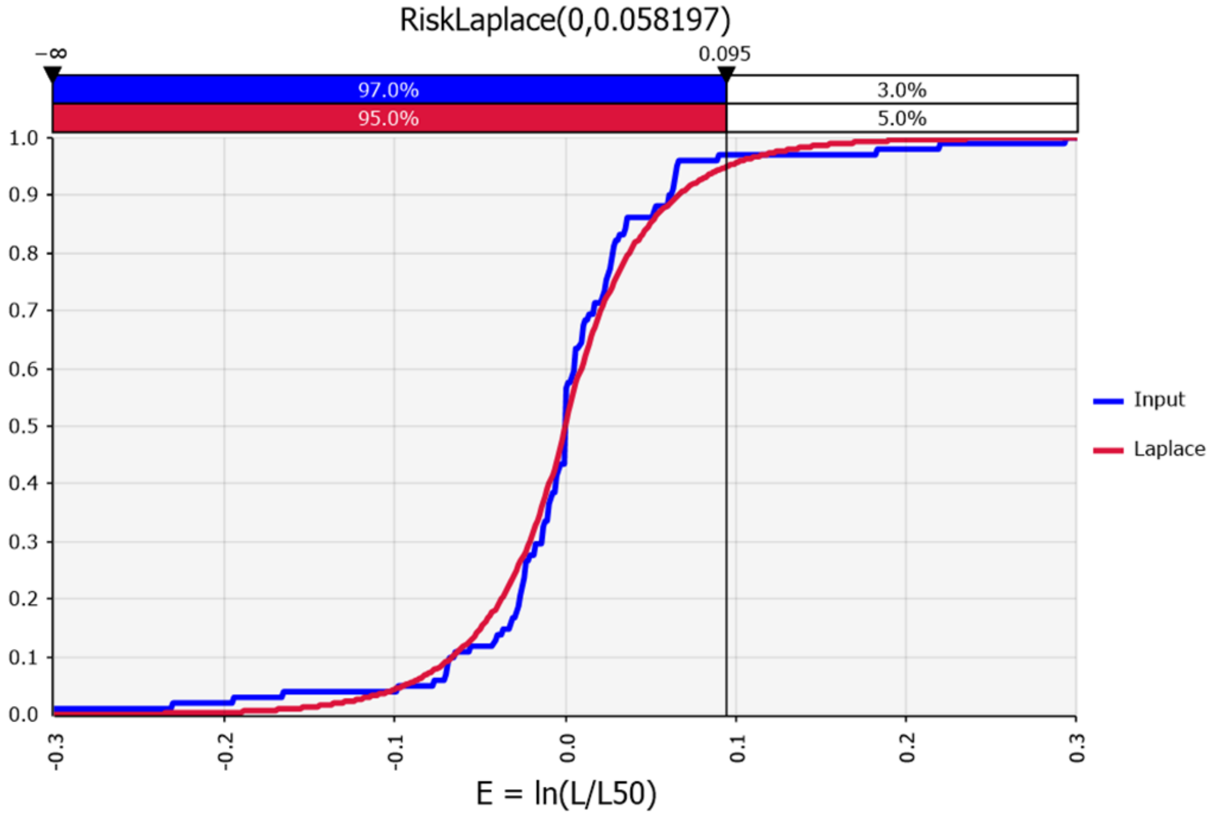


Figure E.2: Precision uncertainties are well represented by a Laplace distribution

Appendix FUncertainties in the Results of Monte Carlo Simulation

The aggregation of uncertain inputs into uncertain outputs occurs twice in the BEPU process. First, the validation metric is epistemically uncertain because of uncertainties in testing (M, Section 0) and uncertainties in simulation solution errors (P, Section 0). Second, risk integration is epistemically uncertain because of uncertainties in simulation solution errors (S(DP), Section 0) and uncertainties in model form error (E_{mf}, Section 0 In general, the propagation of input uncertainties is multivariate, meaning there can be many sources of input uncertainties.

There are many methods for propagating input uncertainties through to validation metrics and risk QOIs. Examples are Monte Carlo (MC), Latin Hypercube Sampling (LHS), method of moments, polynomial chaos, and Belief and Plausibility. Each is a numerical algorithm, and the computed results are subject to numerical errors and uncertainties analogous to discretization errors. The quantification of these errors and uncertainties is consistent with the concepts of BEPU. Are numerical errors acceptable in the computed percentiles of interest for a computational budget of N evaluations of the computational model?

F.1 Propagation of uncertainties via Monte Carlo (MC) simulation

This report uses Monte Carlo (MC) simulation as the method of choice. MC is considered the gold standard because it does not rely upon strong assumptions for inputs or outputs, such as linearity, monotonicity, or continuity. Threshold phenomena such as failure and resonant responses can be modeled.

MC randomly samples a value from each input distribution to form a set of inputs for the numerical computation of the QOI. Random means that the samples are independent and identically distributed (iid). Independent means that the input distributions are not correlated. Identically distributed means that all samples are taken from the same probability distribution. The process is repeated many times with a total sample size of N, where each set of inputs results in a new computed value for the QOI. The process produces a distribution of outputs of size N from which percentiles of interest (e.g., L_{95}) can be calculated.

The computed percentiles of interest are subject to sampling uncertainties and are epistemically uncertain because the sample size N is finite. Conceptually, the process can be replicated many times, and each replicate gives a different result for the percentiles of interest.

This is illustrated in Figure F.1. The solid black line represents truth, Normal(1370,68.5). The grey curves represent five replicates with sample size N=100, and the grey dots represent the 95th percentile for each of the five replicates shown i.e., L(95%tile). The goal is to compute the 95 percent confidence bound on the percentiles of interest, e.g., $L_{95,95}$.

The variance in the lumbar load percentile of interest, e.g., $V(L_{\%})$, converges as

$$V(L_{\%}) \sim \frac{V}{N}, \qquad \textbf{F-1}$$

for simple random sampling, where V is the variance of the full distribution, and N is the sample size. Recognize that $V(L_{\%})$ is a function of both sample size *and* the distribution variance. The latter cannot be known a priori.

Conceptually informative, the replication process is impractical because it requires an enormous number of simulations with the computational model. For a sample size N, the process would have to be replicated thousands of times to compute the confidence bounds. This weakness will be addressed in Appendix 0 by leveraging the attributes of iid; consequently, there are three advantages to MC.

1. Sample size is easily increased: New samples can be pooled with existing samples without repeating simulations of the computational model that have already been performed.
2. Guidelines for sample size: Guidelines for the a priori selection of sample size can be developed (Appendix 0) i.e., before any simulations of the computational model are performed.
3. Confidence bounds: Confidence bounds on the percentiles of interest for a computational budget of N simulations can be computed with either Bootstrap methods (Appendices 0) or application of Wilks method (Appendix 0) with no new computational model simulations.

Latin Hypercube Sampling (LHS) is a related methodology that converges as

$$V(L_{\%}) \sim \frac{V}{N^2}. \qquad \textbf{F-2}$$

LHS is a type of stratified sampling. The idea is to divide the cumulative probability into N equal probability intervals and randomly select one sample from each stratum. Samples from each input distribution are "shuffled" before forming N sets of inputs for the computational model, introducing an element of randomness.

The second-order convergence is very appealing, making LHS the common default choice. However, stratified sampling does *not* share the advantages of random sampling and iid; specifically,

1. Sample size is *not* easily increased: In general, new samples cannot be pooled with existing samples without repeating computational model simulations. A process called hierarchical LHS allows existing samples to be reused, but this capability is not always available[19]. Hierarchical LHS requires sample size doubling to add new samples to what already exists. Sample size doubling quickly becomes impractical.
2. Guidelines for sample size *do not* exist: Guidelines for the a priori sample size selection do not exist, so there is no way of knowing if the sample size is inadequate or overkill.
3. Confidence bounds *cannot* be estimated: Confidence bounds on the percentiles of interest *cannot* be computed without replicating the process thousands of times, which is impractical.

[19] Hierarchical LHS is not available in the @Risk software used here.

The $1/N^2$ convergence rate is not always fully realized in applications using LHS. The $1/N^2$ convergence rate is only realized when sampling a single distribution or multivariate sampling when only one input uncertainty dominates the total output uncertainty. Multivariate sampling is the expected norm, and there are applications where only one input uncertainty dominates. LHS convergence rates degrade towards 1/N when more than one input uncertainty contributes significantly to the total output uncertainty. This is a consequence of the random shuffling of inputs. Empirically, any advantage of LHS convergence rates is lost when four or more input uncertainties similarly contribute to the total output uncertainty. In addition, the convergence advantage of LHS may be less for nonlinear models. (Manteufel, 2000) showed that LHS convergence degraded to 1/N for the product of two uncertain inputs.

In any actual application, the convergence advantage of LHS cannot be assumed a priori and can only be assessed with a computationally expensive convergence study. However, LHS's convergence rates are never less than MC and are sometimes better, so the conventional wisdom is to accept LHS as the default. However, practical assessment of confidence bounds using LHS is not possible; consequently, MC is recommended when BEPU is the governing framework.

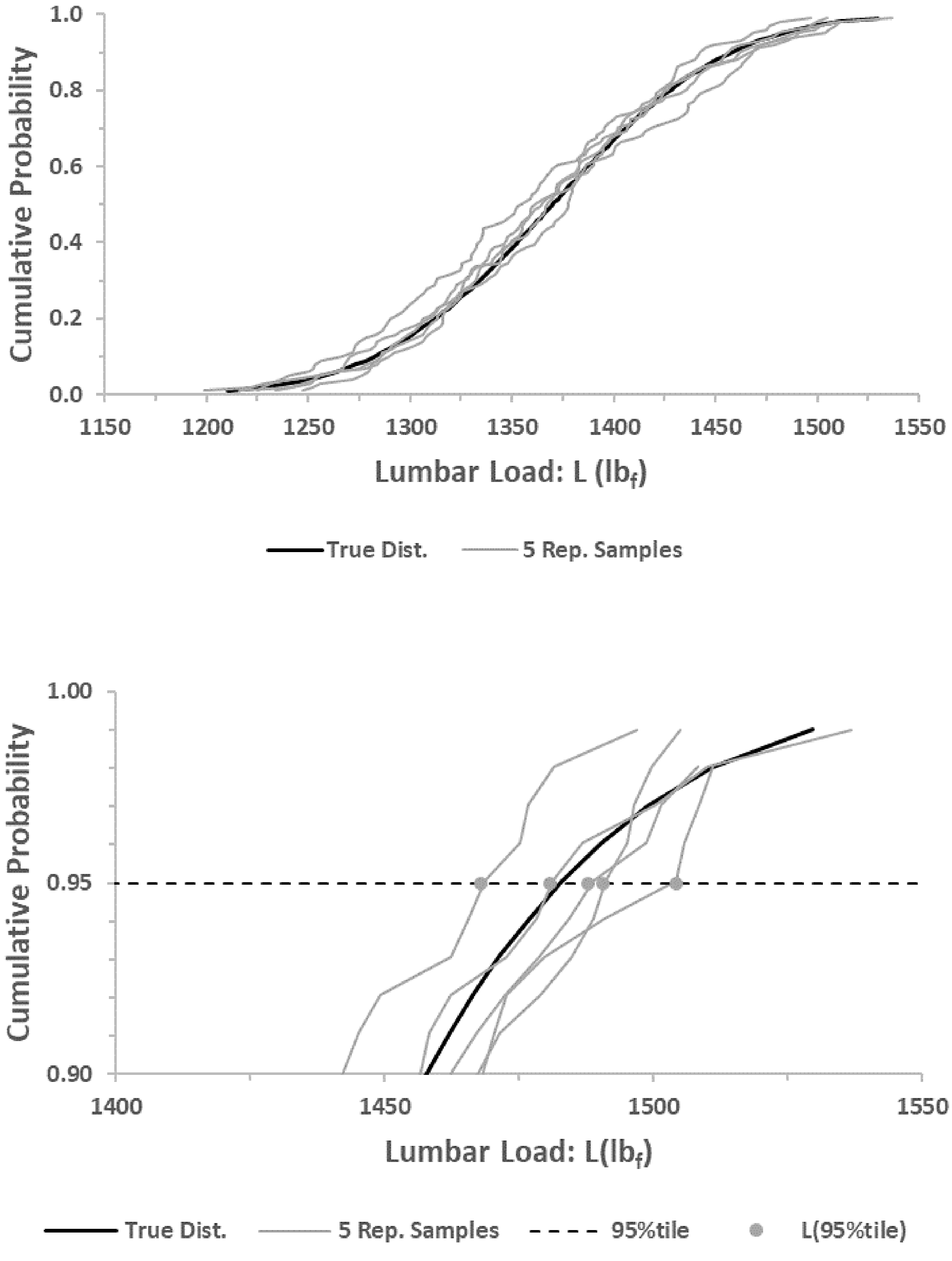


Figure F.1: Load variance for replicated sampling

F.2 Guidelines for selecting the number of Monte Carlo samples

Figure F.2 shows that replicated sampling can be viewed from the alternative perspective of probability variance for a given load. (Ang & Tang, 1975) note that for any parametric or non-parametric distribution,

$$\mathrm{V(p)} = \sigma^2(\mathrm{p}) = \frac{\mathrm{p}(1-\mathrm{p})}{\mathrm{N}}, \qquad \text{F-3}$$

with the additional constraint that

$$\sigma(\mathrm{p}) \ll 1-\mathrm{p} \qquad \text{F-4}$$

to be consistent with the large number assumptions in the derivation (theoretically, the distribution for p should not extend beyond 1). Alternatively, we can write

$$\mathrm{n}\sigma(\mathrm{p}) = 1-\mathrm{p}\,, \qquad \text{F-5}$$

where n is the number of standard deviations that fits on the interval [1-p,1], which informs the selection of sample size,

$$\mathrm{N} > \mathrm{n}^2\frac{\mathrm{p}}{1-\mathrm{p}}. \qquad \text{F-6}$$

Subjectively, $n^2 = 10$ is a minimum, and $n^2 = 100$ is required for accurate solutions.

Table F.1 summarizes recommended sample sizes for the percentiles of interest. For the 95th percentile, 190 samples are recommended as a minimum. Many studies at Sandia National Laboratories have used comparable sample sizes in conjunction with computationally expensive models; however, 1900 samples are not practical even at a National Laboratory. Even the minimum may be impractical for most organizations, in which case, application of the Wilks method (Appendix F.4) may be a more practical alternative. Note, however, that aggregation of uncertainties in the process proposed here is a spreadsheet exercise that is *not* computationally limiting; consequently, the higher guidelines are easily achieved by any organization.

Table F.1: Guidelines for sample size

	Sample Size (N)	
Probability: p	**Minimum** **$n^2 = 10$**	**Accurate** **$n^2 = 100$**
0.50	10	100
0.95	190	1900

These guidelines manage V(p) for a given load. We really want the confidence bounds on loads for percentiles of interest. The guidelines provide an informed starting point, but confidence bounds on loads are also a function of the variance (V) of the output distribution, which cannot be known a priori. Bootstrap and Wilks methods are two ways of addressing this need a posteriori.

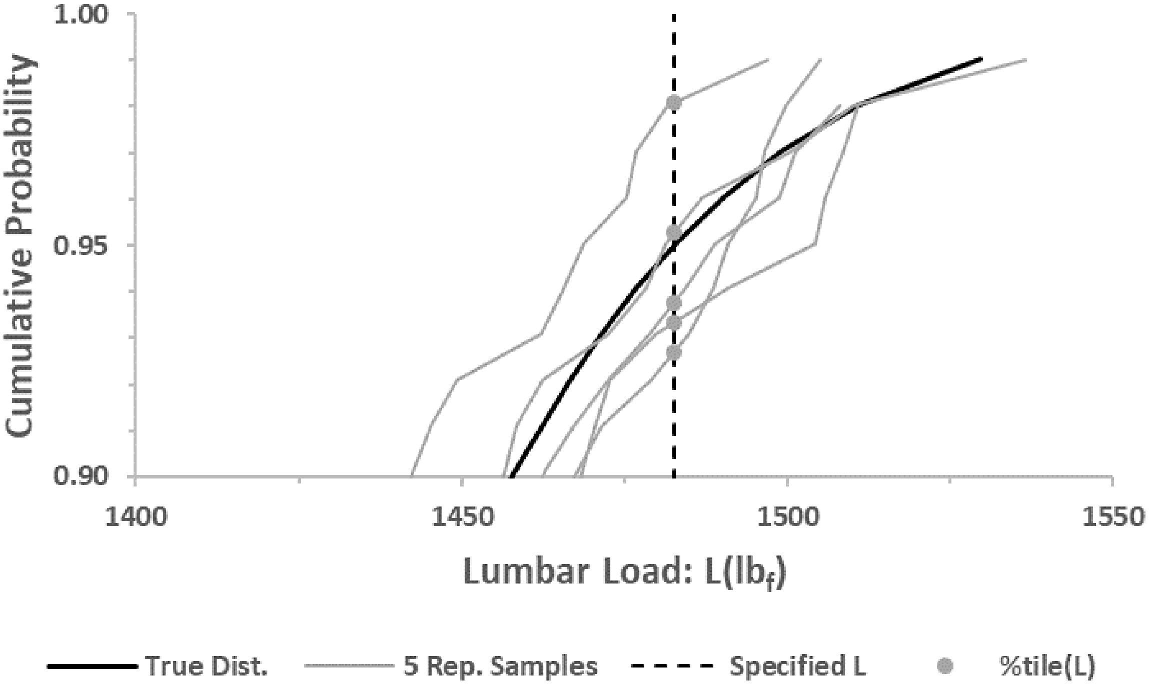


Figure F.2: Probability variance for replicated sampling

F.3 Bootstrap method for computing confidence bounds

Any statistic, like loads at a given percentile, computed from a sample of finite size, N, is subject to sampling uncertainty, i.e., another sample of the same size will produce different results. Non-parametric bootstrap is a practical means of computing confidence bounds without any assumptions about the parent distribution.

The process is shown in Figure F.3. The grey box at the top represents the universe of all values of all uncertain inputs, all values of lumbar loads associated with the uncertain inputs, and an exact value of the percentile of interest. The unknown universe, or parent, is approximated discretely by sampling. The percentile bootstrap method proceeds as follows:

1. Sample all input uncertainties N times, creating sample sets of size N. The guidelines given by Equation F-6 is a good starting point if the computational budget allows it. Each of N sets contains one random value for each of the uncertain inputs. Bootstrap requires simple random sampling (MC) i.e., stratified sampling (LHS) is not consistent with assumptions of bootstrap.

2. Using the simulation computational model, compute lumbar loads for each of the N sets of inputs, creating a set of N lumber loads. The product of this step is a distribution of lumbar loads, L(1).
3. Extract the percentile of interest, e.g., $L_{95}(1)$.
4. Resample with replacement the set of N lumbar loads computed in step 2. Note that this does not require any new computational model simulations and is computationally insignificant. The product of this step is a new lumbar load distribution L(…), referred to as a bootstrap distribution.
5. Extract the percentile of interest, e.g., $L_{95}(\ldots)$.
6. Repeat steps 4 and 5 NR-1 times, producing a new lumbar load distribution and new values for the percentile of interest each time.
    a. One product of this step (combined with steps 2 and 3) is a total of NR different lumbar load distributions reflecting sampling uncertainty. The bottom left graphic in Figure F.3 shows just 3 of NR load distributions and corresponding loads at the 95th percentile.
    b. The bottom right graphic in Figure F.3 quantifies sampling uncertainty in our estimate of the percentile of interest. The upper confidence bound, e.g., $L_{95,95}$, can be extracted from this distribution.

Bootstrap is a powerful and computationally inexpensive tool for estimating confidence bounds (or any statistic) associated with sampling errors, but there are assumptions and limitations worth mentioning.

1. Bootstrap resamples the first distribution of loads computed by MC, which is used as a surrogate for the true underlying distribution (universe). A key assumption is that this surrogate is representative of the "universe," which is more likely to be true if the distribution is not heavily skewed and the number of samples, N, is large. This is an unquantifiable judgment when the "universe" is unknown priori and should be acknowledged as an intangible uncertainty.
2. Resamples are limited to what is observed in the surrogate distribution. This limitation can be relaxed using parametric Bootstrap, but it comes at the expense of introducing intangible uncertainties in representing tail behavior.
3. The number of resamples must be sufficiently large, NR>>N, such that resampling uncertainties are insignificant compared to the original sampling uncertainties. This can be achieved by taking NR as a factor of 10 or more times the guidelines given by Equation F-6.

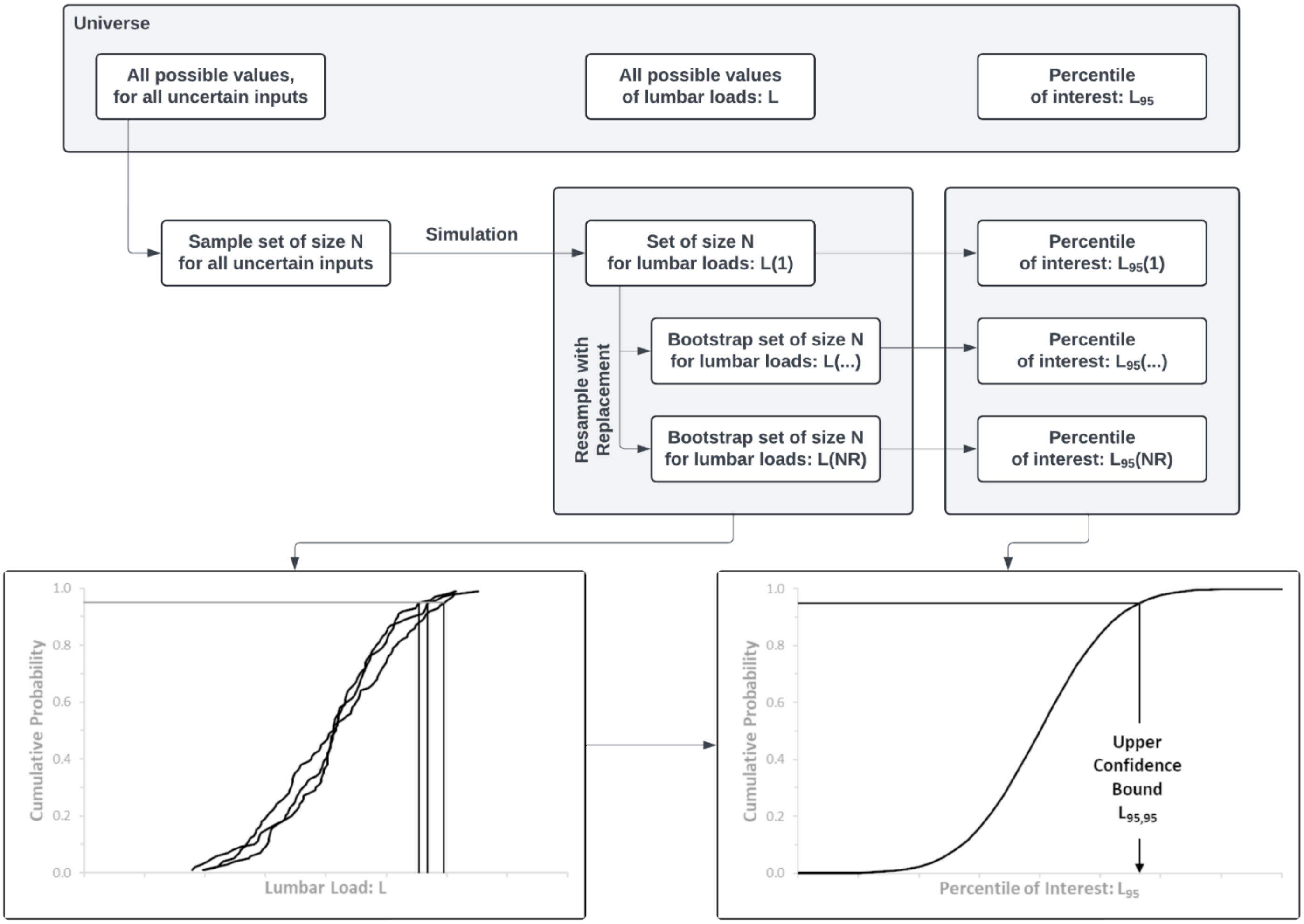

Universe
All possible values, for all uncertain inputs
All possible values of lumbar loads: L
Percentile of interest: L_{95}
Sample set of size N for all uncertain inputs
Simulation
Set of size N for lumbar loads: L(1)
Percentile of interest: $L_{95}(1)$
Resample with Replacement
Bootstrap set of size N for lumbar loads: L(...)
Percentile of interest: $L_{95}(...)$
Bootstrap set of size N for lumbar loads: L(NR)
Percentile of interest: $L_{95}(NR)$
1.0
0.8
0.6
0.4
0.2
0.0
Cumulative Probability
Lumbar Load: L
1.0
0.8
0.6
0.4
0.2
0.0
Cumulative Probability
Upper Confidence Bound $L_{95,95}$
Percentile of Interest: L_{95}

Figure F.3: Non-parametric bootstrap

F.4 Wilks method for computing tolerance limits

The Wilks method (Wilks, 1942) was developed to inform physical sampling but has been applied to computational simulation in recent years, mainly by the nuclear power industry (Lee et al., 2014; Mousseau & Williams, 2017; Porter, 2019; Toptan et al., 2022), where the application of BEPU requires the quantification of uncertainties using computationally expensive computer models.

Figure F.4 aids understanding of Wilks' method. The distribution represents the universe of all possible lumbar loads associated with all values of uncertain inputs. For regulatory purposes, we want to compare the upper tolerance bound ($L_{(t)}$) of the true load (L^t) associated with the percentile (p) of the distribution to some regulatory requirement (L_{req}).
Application of the Wilks method starts with an ordered sample set of size N, i.e., $L_{(1)} < L_{(2)}$ … $<L_{(t)}$ … $<L_{(N)}$. Here, t is an index counting from 1 to N. Theory requires that the samples be randomly selected; consequently, samples must be generated by MC. LHS violates the foundational assumption. Wilks' method identifies the upper one-side tolerance bound, $\mathrm{L}^{\mathrm{U}}_{\mathrm{p},(1-\alpha)} = \mathrm{L}_{(\mathrm{t})}$, such that $L_{(t)}>=L^t$ for the specified percentile (p) with a specified confidence (1-a). Note that L(t) is intended to bound L^t on the high side (within specified confidence); it is not intended to be an accurate estimate of L^t. It could be much larger for smaller sample size choices and the variance of the underlying true distribution.

There are assumptions and limitations to note.

1. Results are not dependent on an assumed distribution form, i.e., the application of Wilks' method is non-parametric.
2. Results depend on assumptions of a smooth distribution; consequently, threshold phenomena (e.g., failure) cannot be assessed with the Wilks method. This can be an important limitation for certain classes of applications.
3. Application of the Wilks method is limited to a single QOI. This limitation can be relaxed (Wald, 1943), but it is well outside the discussion here.

Wilks' method has historically been applied when simulation models are computationally expensive. The question asked was, "What is the minimum number of simulations required to compute the upper tolerance bound, $\mathrm{L}^{\mathrm{U}}_{\mathrm{p},(1-\alpha)}$?" For this question, the minimum sample size is given by

$$\mathrm{N} = \frac{\ln[1-(1-\alpha)]}{\ln \mathrm{p}}, \qquad \textbf{F-7}$$

where p is the percentile of interest, (1-a) is the specified confidence, and the upper tolerance bound is the largest of the sample values, i.e., $L_{(t=N).}$ Rounding up is appropriate since Equation F-7 does not produce integers.

Table F.2 compares Equation F-7 for the Wilks method with previous guidance, Equation F-6. Wilks' method produces shockingly small sample sizes for high-confidence results, especially for the 95th percentile. There is no error, so what are we missing? Previous guidance

(Appendices 0 and 0) is aimed at finding accurate lumbar loads for the percentiles of interest. Wilks' method, on the other hand, guarantees a bounding value of lumbar load for percentiles of interest. Thus, the Wilks method achieves smaller sample sizes with guaranteed conservatism at the expense of accuracy, which could be a potential problem when margins are small.

Table F.2: Wilks sample sizes compared to previous guidance

	Sample Size (N)		
Probability: p	**Minimum** **n^2 = 10**	**Accurate** **n^2 = 100**	**Wilks** **(1-a) = 0.95**
0.50	10	100	5
0.95	190	1900	59

A far less common application of Wilks would ask, "What is the upper tolerance bound for a given computational budget, N, which is greater than the minimum." The upper tolerance bound will be less conservative when sample sizes are larger than the minimum. A more general form of the Wilks method is required:

$$1 - beta.dist(1 - p, N - t + 1, t, cumulative) \geq 1 - \alpha. \qquad \textbf{F-8}$$

Excel can evaluate the cumulative Beta function, and the SOLVER utility in Excel can solve the non-linear equation for the rank position t (rounding up for non-integer values) of the ordered samples.

A hybrid application of Wilks' method would use sample size guidance given by Equation F-6 and the Wilks formula to find the upper tolerance bound, thus eliminating the need for bootstrap.

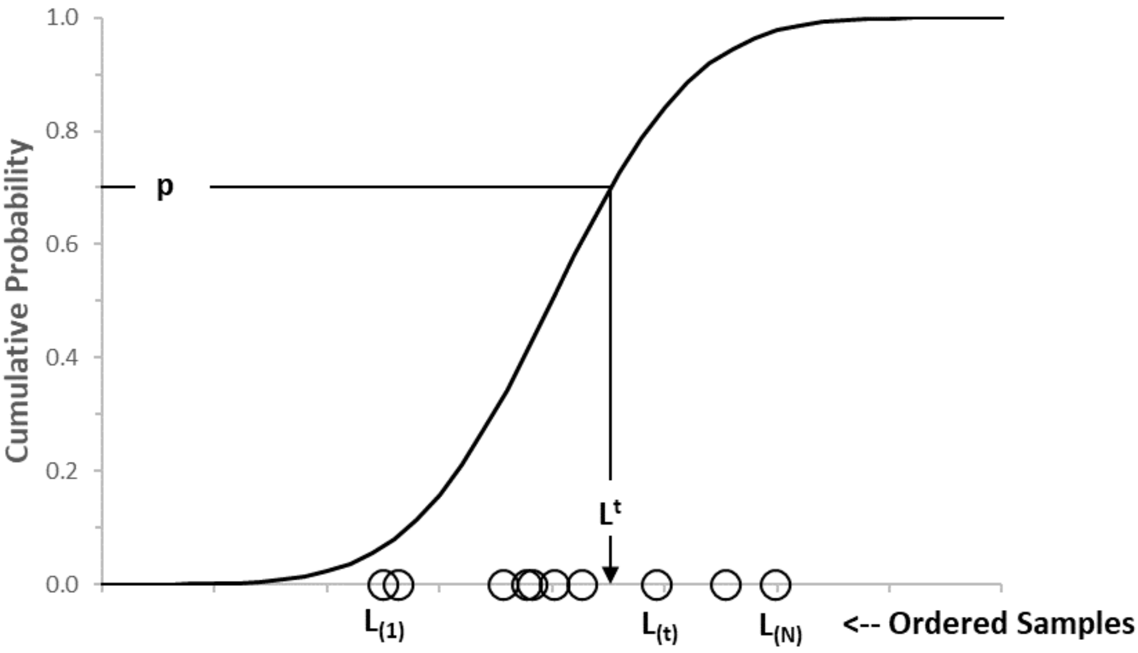


Figure F.4 Understanding Wilks Formula

F.5 Assessment of Monte Carlo sampling errors using bootstrap and Wilks methods

The concepts discussed in Appendices 0, 0, and 0 are illustrated here for the demonstration of aircraft seat certification. Table F.3 provides an assessment of Monte Carlo (MC) sampling errors for best-estimate decisions based on the median of uncertain predicted lumbar loads computed with a sample size N. A sample of N=5 is the minimum for computing the upper 1-sided tolerance bound of L_{50} using the Wilks method (AppendixF.4). A sample size of N=10 corresponds to the minimum guideline for computing L_{50}, and a sample size of N=100 corresponds to the guideline for an accurate estimate of L_{50} (Appendix 0). Lastly, a sample size of N=100000 is chosen to oversample L_{95} and certainly L_{50} intentionally. With this magnitude of oversampling, L_{50} and FoS are expected to be close to truth. The resample size, $N_{rs}=10^5$, was chosen to ensure that resampling errors are insignificant compared to the original sampling errors for the percentile of interest.

When regulating on L_{50}, Table F.3 shows that the FoS is fully converged for a sample size N=100. This is the sample size guideline for accurate estimates of L_{50}. Smaller sample sizes incur a penalty in the form of a reduced FoS associated with increased sampling errors; however, the penalty is not excessive and only a concern if the margins are small. A sample size of N=5 can produce useful results!

Similar observations can be made when the intent is to regulate with high confidence. Table F.4 provides an assessment of MC sampling errors for a high-confidence decision based on the 95th percentile of uncertain predicted lumbar loads computed with a sample size N. A sample of N=59 is the minimum for computing the upper 1-sided tolerance bound of L_{95} using Wilks' method (Appendix 0). MC and bootstrap are not evaluated for this sample size because estimates of L95 would require excessive extrapolation outside the sample base. A sample size of N=190 corresponds to the minimum guideline for computing L_{95}, and a sample size of N=1900 corresponds to the guideline for an accurate estimate of L_{95} (Appendix 0). Lastly, a sample size of N=100000 was chosen to oversample L95 intentionally. With this magnitude of oversampling, L_{95} and FoS are expected to be close to truth. The resample size, $N_{rs}=10^5$, was chosen to ensure that resampling errors are insignificant compared to the original sampling errors for the 95th percentile.

When regulating on L_{95}, Table F.4 shows that the FoS is fully converged for a sample size N=1900. This is the sample size guideline for accurate estimates of L_{95}. Smaller sample sizes incur a penalty in the form of a reduced FoS associated with increased sampling errors; however, the penalty is not excessive and only a concern if the margins are small. A sample size N=59 can produce useful results, even when MC and bootstrap should be avoided!

The upper 1-sided tolerance bound using the Wilks method is greater than the 95th percentile confidence bound computed using the bootstrap method, and the corresponding FoS's are smaller using Wilks' method. The difference is greatest at the smallest sample sizes, but it is not excessive.
Table F.5 compares the pros and cons of the bootstrap and Wilks methods, which leads to the following guidance:

1. Wilks' method is preferred if the underlying assumptions are not restrictive.
    a. Use Wilks' method when the uncertainty rollup is computationally expensive. Use the minimum sample size required to compute the upper 1-sided tolerance bound. Table F.2 shows that these sample sizes are significantly less than required for accurate MC results.
    b. Use Wilks' method when more than the minimum number of samples is required to assess small margins with confidence. Use a sample size consistent with the computational budget to reduce conservatism in the upper 1-sided tolerance bound.
    c. Use Wilks' method when the uncertainty rollup is computationally inexpensive. Use guidance for accurate MC results (Table F.1) to inform sample size and minimize conservatism in the upper 1-side tolerance bound.
2. Bootstrap is required if the underlying assumptions restrict the use of the Wilks method.
    a. Bootstrap confidence bounds are suspect for small sample sizes and skewed distributions, which introduce Intangible uncertainties that must be acknowledged. Small sample sizes are associated with a computationally expensive rollup of uncertainties.
    b. Intangible uncertainties associated with sample size and skewness of the distribution are minimized for a computationally inexpensive rollup of uncertainties when sample size can be informed by the need for accurate MC results (Table F.1).

Table F.3: Assessment of sampling errors for a best-estimate decision

Monte Carlo Sampling		Bootstrap		Wilks		Decision Metrics		
Sample Size	Percentile Of Interest	Resample Size	Upper Conf. Bound	Position	Upper 1-Sided Tol. Bound	Nominal	Bootstrap	Wilks
N	L_{50}	N_{rs}	$L_{50,95}$	t	$L_{50,95}^{U}$	$\mathrm{FoS} = \frac{L_{req}}{L_{50}}$	$\mathrm{FoS} = \frac{L_{req}}{L_{50,95}}$	$\mathrm{FoS} = \frac{L_{req}}{L_{50,95}^{U}}$
5	1090	10^5	1124	5	1166	1.38	1.33	1.29
10	1084	10^5	1108	9	1129	1.38	1.35	1.33
100	1103	10^5	1102	59	1108	1.36	1.36	1.35
100000	1102					1.36		

Table F.4: Assessment of sampling errors for a high-confidence decision

Monte Carlo Sampling		Bootstrap		Wilks		Decision Metrics		
Sample Size	Percentile Of Interest	Resample Size	Upper Conf. Bound	Position	Upper 1-Sided Tol. Bound	Nominal	Bootstrap	Wilks
N	L_{95}	N_{rs}	$L_{95,95}$	t	$L_{95,95}^{U}$	$\mathrm{FoS} = \frac{L_{req}}{L_{95}}$	$\mathrm{FoS} = \frac{L_{req}}{L_{95,95}}$	$\mathrm{FoS} = \frac{L_{req}}{L_{95,95}^{U}}$
59				59	1242			1.21
190	1201	10^5	1225	186	1243	1.25	1.22	1.21
1900	1210	10^5	1217	1821	1218	1.24	1.23	1.23
100000	1212					1.24		

Table F.5: Pros and cons of bootstrap and Wilks methods

Method	Pros	Cons
Bootstrap	• Non-parametric • Not limited to smooth distribution assumptions, simulation model can include threshold phenomena • Multiple QOI can be easily assessed in a single study • Resampling is computationally inexpensive because it does not involve the simulation computational model • Commonly used statistical tool that is easy to understand and explain	• Sample distribution must be representative of the parent distribution. This is an intangible uncertainty and suspect for small sample sizes and skewed distributions. • Results are not guaranteed conservative because of previous assumption • Requires additional effort to manage the resampling and process the results
Wilks	• Non-parametric • Small sample sizes possible for computationally expensive simulation models • Guaranteed conservative results • Simple to implement — select value at identified position from sorted sample distribution	• Results are dependent on assumptions of a smooth distribution, i.e., the simulation model cannot include threshold phenomena • Limited to a single QOI • Results may be too conservative when using the minimum sample size if the margin is small • Lack of familiarity in the analysis and regulatory communities • Concepts and underlying math are difficult to understand and explain