SANDIA REPORT
SAND2007-6219
Unlimited Release
Printed December 2007

Toward a More Rigorous Application of Margins and Uncertainties within the Nuclear Weapons Life Cycle – A Sandia Perspective

Kathleen Diegert, Scott Klenke, George Novotny, Robert Paulsen, Martin Pilch, and Timothy Trucano

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

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration, under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

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

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

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

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

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2007-6219
Unlimited Release
Printed December 2007

Toward a More Rigorous Application of Margins and Uncertainties within the Nuclear Weapons Life Cycle – A Sandia Perspective

Kathleen V. Diegert
Reliability Assessment and Human Factors

Scott E. Klenke
New Mexico Stockpile Issues and Planning

George Novotny
Weapons Engineering and Product Realization

Robert A. Paulsen, Jr.
New Mexico Stockpile Issues and Planning

Martin Pilch
QMU and Management Support

Timothy G. Trucano
Optimization and Uncertainty Estimation

Sandia National Laboratories
P. O. Box 5800
Albuquerque, New Mexico 87185

Abstract

This paper presents the conceptual framework that is being used to define quantification of margins and uncertainties (QMU) for application in the nuclear weapons (NW) work conducted at Sandia National Laboratories. The conceptual framework addresses the margins and uncertainties throughout the NW life cycle and includes the definition of terms related to QMU and to figures of merit. Potential applications of QMU consist of analyses based on physical data and on modeling and simulation. Appendix A provides general guidelines for addressing cases in which significant and relevant physical data are available for QMU analysis. Appendix B gives the specific guidance that was used to conduct QMU analyses in cycle 12 of the annual assessment process. Appendix C offers general guidelines for addressing cases in which appropriate models are available for use in QMU analysis. Appendix D contains an example that highlights the consequences of different treatments of uncertainty in model-based QMU analyses.

Acknowledgments

A large number of people who work in the Nuclear Weapons program at Sandia read and commented on early versions of this paper. The authors are grateful for the advice and support they received through this process. We thank Walter Wolfe, who carefully read and commented upon an almost final draft of the paper. We also thank Rhonda Reinert of Technically Write for her skillful and essential editing of this paper to bring it into a final form suitable for publication.

Contents

Acknowledgments 4
Acronyms and Abbreviations 8
1 Introduction 9
2 Sandia Philosophy and Vision on the Role of QMU in Engineering Designs 10
3 Elements of QMU 12
4 Figures of Merit for QMU 14
5 QMU in the NW Life Cycle 18
5.1 Requirements Definition: Basis for Margins 19
5.2 Development and Qualification: Establishing Margins and Uncertainties 20
5.3 Production: Maintaining Margins and Uncertainties 21
5.4 Stockpile Assessment: Monitoring for Changes in Margins and Uncertainties 22
5.5 Retirement/Dismantlement: Monitoring for Subset of Margins and Uncertainties 23
6 Concluding Perspective 23
References 25
Appendix A – General Guidance for Physical Data–based QMU Analysis 26
A.1 Background 26
A.2 Time-Dependent Change 26
A.3 Figure of Merit 28
A.4 Physical Data 28
A.5 Populations 28
A.6 Physical Data Samples 29
A.7 Distributions 30
A.8 Margins 32
A.9 Uncertainty 32
A.10 Trends 33
A.11 Summary 34
A.12 Reference 35
Appendix B – Guidelines for Common Data Analysis and a Presentation Using Physical Data 36
B.1 Introduction to QMU Concepts as Employed in the Analysis 36
B.2 Assumptions 37
B.3 Common Steps for Analysis and Presentation 38
B.4 Summary of Analysis and Plots 49
Appendix C – Guidelines for Application of QMU Methodologies to M&S-Centric Evaluations 51
C.1 Introduction 51
C.2 Requirements Language 51
C.3 Figures of Merit 52
C.4 Graded Approach to QMU 53
C.5 Implementation Guidance 58
C.6 References 64

Appendix D – Example Application of QMU Methodologies to M&S-Centric Evaluations 66
D.1 Sample Problem Description 66
D.2 First-Order Probability 68
D.3 Second-Order Probability 74
D.4 Mixed Probability/Interval Analysis 78
D.5 Comments on the Three Methodologies 81
D.6 References 83

Figures

1 Illustration of margin and uncertainty concepts 14
2 Elements of decision paradigm shift that support science-based engineering transformation through QMU and its connection with risk-informed decisions (Pilch, Trucano, and Helton 2006) 24
A-1 Trend line for mean performance. 27
A-2 Histogram of measured performance values exhibiting bimodality. 31
B-1 Illustration of the recommended response versus production date plot. 39
B-2 Illustration of the recommended response versus test date plot. 39
B-3 Illustration of the recommended upper margin versus age plot. 41
B-4 Illustration of the recommended lower margin versus age plot. 41
B-5 Illustration of the recommended (and updated) upper margin versus age plot. 43
B-6 Illustration of the recommended K-factor upper performance threshold versus component age plot. 44
B-7 Illustration of the recommended K-factor lower performance threshold versus component age plot. 45
B-8 Normal probability plot of residuals. 46
B-9 Residuals versus age of unit plot 47
C-1 Probabilistic representation of epistemic results underrepresents true uncertainty. 60
D-1 Input distributions for model parameters *a* and *b*. 68
D-2 Residuals between code-based and surrogate model–based assessments of *R*. 70
D-3 Results of first-order probability method applied to synthetic problem. 71
D-4 Probability that the M&S-based assessment of *R* exceeds the threshold R_T. 71
D-5 Confidence factor (*CF*) format for QMU. 72
D-6 Scatterplots to assess sensitivity of QMU results to input variability and uncertainties. 73
D-7 Results of second-order probability analysis. 76
D-8 Confidence in computed safety factors resulting from epistemic uncertainties. 76
D-9 Computation of confidence factor for second-order probability. 77
D-10 Scatterplot for sensitivity analyses in second-order probability. 77
D-11 Confidence in the computed confidence factor *CF* associated with numerical errors. 78
D-12 "Uncertainty-preserving" presentation of epistemic results with a failure probability metric. 79
D-13 "Uncertainty-preserving" presentation of epistemic results with a safety factor metric. 80

Tables

1 Elementary Terminology in QMU 17
B-1 Summary Values for Example Problem..... 49
C-1 Graded Approach to Uncertainty Quantification and Sensitivity Analysis 54
C-2 Predictive Capability Maturity Model 57
D-1 Relevant Qualification Issues Exemplified by Synthetic Problem 67
D-2 Computationally Constrained Assessment of QMU for Synthetic Problem..... 69
D-3 Sensitivity to Sources of Numerical Error in First-Order Probability Results..... 74
D-4 Results of Second-Order Probability Analysis 75
D-5 Results Comparison for First-Order and Second-Order Probability Methods..... 82

Acronyms and Abbreviations

AD	Andersen-Darling
A&E	aleatory and epistemic
BC	boundary condition
CD	compatibility definition
CDF	cumulative distribution function
ES	environmental specification
F&C	features and capabilities
I/O	input/output
ICD	interface control document
IET	integral-effects tests
LANL	Los Alamos National Laboratory
LHS	Latin hypercube sampling
LLNL	Lawrence Livermore National Laboratory
M&S	modeling and simulation
MC	military characteristics
NS	nuclear safety specification
NW	nuclear weapons
PCCM	Predictive Capability Maturity Model
PD	production date
PDF	probability density function
PLOAS	probability of loss of assured safety
R	response
QMU	quantification of margins and uncertainties
QRA	quantitative risk assessment
SA	sensitivity analysis
Sandia	Sandia National Laboratories
SET	separate-effects test
SFT	stockpile flight test
SLT	stockpile laboratory test
SN	serial number
SQE	software quality engineering
SRQ	system response quantity
STS	stockpile-to-target sequence
TD	test date
UQ	uncertainty quantification

1. Introduction

Historically, the general approach to dealing with margins and uncertainties in the life cycle of the nuclear weapons (NW) stockpile involved conservative requirements, conservative scenarios, and conservative assessments, all of which were heavily dependent on judgment and testing. Because our computational modeling capability was relatively immature when the stockpile weapons were designed, modeling and simulation (M&S) tended not to significantly reduce this conservatism. Occasionally, surprises occurred when there was no full accounting of uncertainties and when performance objectives were not met.

Decades ago, the weaponeers who designed the warheads in the current stockpile incorporated margins in their designs that often reflected the above-mentioned conservatism in the treatment of uncertainties. These weaponeers developed conservative production specifications to ensure that their design intent was accurately reflected in the as-produced and fielded product. Because of the large performance margins designed into these components, margin failures were almost never observed in stockpile testing, Instead, production and assembly errors (birth defects) were usually identified as the cause of the failures. Until several years ago, much of our stockpile evaluation efforts continued to be focused on finding birth defects through functional testing of the stockpile hardware. This focus on uncovering birth defects, among other factors, has caused us to lose some of our perspective on the original margin quantities. Consequently, much of the stockpile assessment information we have dealt with is reflective of an environment rich with physical test data that are reduced to warhead ***attributes*** (i.e., data that are interpreted in a categorical pass-fail sense) and that do not explicitly reflect quantified margins and uncertainties.

We initiated an effort in the 2001–2002 time frame to reacquire the faded perspective on the original margin quantities and to increase our understanding of margins for nonnuclear subsystems in the stockpile. We began to shift from the narrower focus on attributes to a broader evaluation-based approach, where more emphasis is placed on the collection and analysis of ***variables*** data (that is, measurements of physical quantities such as temperatures, densities, and so on) for the entire range of stockpile life-cycle activities. This change in strategy during the 2001–2002 time frame was driven in part by the lessons learned from the B61 Alt 335 experience. In addition, a strong belief by Sandia leadership contributed to the shift from a "surveillance-based" approach to an "evaluation-based" approach. Today, our weapon teams are making progress as is reflected in our annual stockpile review, annual assessment meetings, and routine interactions dealing with stockpile matters throughout the year. We have discovered that this is a journey, not a destination, and that the change we are seeking must be thought of as one of "continuous improvement," with a destination that will keep moving in line with our progress.

This paper outlines the path forward for the NW program at Sandia. This path emphasizes a clear understanding of the tools and methodology that are collectively referred to as ***quantification of margins and uncertainties*** (QMU) (JASON 2005; Pilch, Trucano, and Helton 2006; Sharp and Wood 2003); the timely acquisition or

reacquisition of associated quantitative physical and M&S data and product knowledge; and the documentation, tracking, and application of these data and knowledge to support an improved understanding of product and product-aging trends, thus informing decisions about the stockpile.

This paper presents the conceptual framework we are using at Sandia to define QMU for application in our NW work. Section 1 has briefly addressed the historical context out of which this framework has evolved. Section 2 provides the vision that guides our engineering activities and identifies applications for QMU at Sandia. In Section 3, we define the key elements of the QMU methodology and the terminology used in these elements. Section 4 continues the discussion of terminology, summarizing the key terms in a table. Section 5 focuses on the role of margins and uncertainties throughout the five phases of the NW life cycle. In Section 6, we explain how QMU is an important decision-support methodology for risk-informed decision making.

The four appendices in this paper offer guidance and examples for conducting QMU analyses. Appendix A explains how to conduct a physical (experimental and test) data–based analysis. Appendix B presents a step-by-step approach for conducting and presenting an analysis using physical data. This approach was used in cycle 12 of the annual assessment process. Appendix C focuses on analyses in which M&S plays a central role. Appendix D, a companion to Appendix C, provides an example of applying the QMU methodology in a computational model-based analysis.

2. Sandia Philosophy and Vision on the Role of QMU in Engineering Designs

The following items reflect Sandia's philosophy on the role of margins and uncertainties in our engineering activities:

- Consistent with sound engineering practices, we explicitly account for margins and uncertainties in the design of nuclear warheads and their components.
- We implement and manage margins that are consistent with cost, schedule, and performance requirements.
- We explicitly account for, monitor, and analyze margins and uncertainties throughout a warhead's life cycle using QMU.
- We use the information obtained from QMU to inform decisions about warhead and bomb requirements, design and development, production, stockpile evaluation, refurbishment, and retirement.

We presently execute all the above items to some degree but are in the process of transitioning to a more formal implementation.

Addressing margins and uncertainties provides fundamental support to vital stockpile decisions, including those related to requirements, specifications, performance, evaluation needs, aging trends, and replacement decisions.

We are working to achieve the following vision that will be the foundation of our NW engineering work.

> The application of QMU is inherent in our science-based engineering *throughout the product life cycle*. Key elements of applying a QMU methodology consist of the following: (1) specification of performance thresholds; (2) identification of associated performance margins, where a performance margin is a measure of exceeding the performance threshold; and (3) quantification of uncertainty in the performance thresholds and performance margins as well as in the larger framework of the decisions being contemplated. Experiments, testing, M&S, and expert judgment are all used to acquire this information. Statistical methods are used for characterizing the performance threshold, for assessing actual performance, and for characterizing the uncertainty in that assessment. Formal identification, documentation, and tracking of margins and their associated uncertainties are initiated during the development and qualification phase of a weapon's life cycle and continue through the rest of the life cycle. This scrutiny is applied using a graded approach to key "make-the-bomb-work" functions and to "critical performance parameters."

Applications for QMU at Sandia to support our risk-informed decision making include the following:

- Managing and assuring successful system integration
- Managing and assuring successful product realization
- Improving the technical basis of products and increasing the knowledge and resulting understanding of these products
- Supporting the analysis and understanding of observed aging trends and predictions to aid NW Complex infrastructure responsiveness
- Enabling improved decisions about the quantity and allocation of surveillance samples
- Quantifying uncertainty in experimental data and in M&S data, both of which constitute the product technical basis
- Conveying a level of confidence in reliability and safety assessments

3. Elements of QMU

QMU is primarily a technical framework for producing, combining, and communicating information about performance margins of complex systems to support risk-informed decision making for stockpile stewardship over the NW life cycle. QMU also provides an important framework for organizing the complex set of key organizational roles and responsibilities that must produce and use this information. The application of QMU may vary in how uncertainties are aggregated or how margins are calculated, but the conceptual elements should be consistent, i.e., (1) specification of performance thresholds, (2) identification of associated performance margins, and (3) quantification of uncertainty in the performance thresholds and the performance margins as well as in the larger framework of the decisions being contemplated.

The three key elements of the QMU methodology require attention throughout the stockpile life cycle. Definitions and explanations of the terminology used in these elements are given below.

- **Performance threshold:** *Performance* is the ability of a bomb, a warhead, or a component to provide the proper function (e.g., timing, output, response) when exposed to the sequence of design environments and inputs (see Section 5). This definition of performance is applicable to the following functional-requirement areas: reliability, nuclear safety, use-control, and nuclear survivability. A *performance threshold* is a specification of a necessary performance achievement, typically in quantitative form. We call this the *required performance* of a system. The required performance is most often specified in a deterministic form where the performance must be greater than (or less than) a specified performance threshold. Less frequently, but in important areas such as nuclear safety, the required performance is stated in terms of a probability that the performance is greater than (or less than) a specified threshold.

  Defining performance thresholds is a key activity in two phases of the NW life cycle: (a) requirements definition and (b) development and qualification. Both experiment and test, as well as M&S, can provide key information for the specification of performance thresholds. Past experience and current expert judgment are also contributors.

  NOTE: We also use the word “threshold” as shorthand for “performance threshold” throughout this paper, unless otherwise indicated.

- **Performance margin:** A performance margin is the difference between the required performance of a system and the demonstrated performance of a system, with a positive margin indicating that the expected performance exceeds the required performance. The expected performance can be nondeterministic and may be specified by a probability or cumulative distribution function. In traditional engineering design, a positive design margin is attained through conservative design practices that incorporate worst-case assessments, safety factors, and expert engineering judgment.

The determination of performance margins is a complex engineering activity that is based on experiment/test, M&S, experience and expert judgment. The qualification activities of the product life cycle provide the most critical information that addresses achievement of performance requirements. The demand for demonstrating the science basis for achieving desired margins in system performance is increasing, and this is an important driver for the need to develop and apply a consistent QMU approach.

NOTE: We also use the word "margin" as shorthand for "performance margin" throughout this paper, unless otherwise indicated.

- **Uncertainty:** There is uncertainty in the specification of thresholds and margins, as well as in the larger framework of the decision tasks. This uncertainty begins in the requirements that provide a foundation for the definition of performance thresholds, and it accumulates and transforms as the various science and engineering activities that lead from weapons design to qualification to evaluation are executed. There are two general types of uncertainty that must be separately accounted for, quantified, and aggregated within QMU:

  1. *Aleatory uncertainty* – also called irreducible uncertainty or stochastic variability. We typically refer to this type of uncertainty as simply *variability*. Aleatory uncertainty (or variability) is naturally characterized, quantified, and communicated in terms of probability. Common examples are variability in manufacturing processes, material composition, test conditions, and environmental factors, which lead to variability in component or system performance.

  2. *Epistemic uncertainty* – also called reducible uncertainty. This type of uncertainty is due to lack of knowledge or incomplete knowledge. Common examples of epistemic uncertainty are the so-called model form uncertainty (that is, uncertainty in how well the equations in the model capture the physical phenomena of interest), both known and unknown unknowns in scenarios, and poor-quality physical test data. In some circumstances, epistemic uncertainty may be quantified by using probabilistic and statistical concepts or by using other methods.

Quantification is a basis of sound engineering. QMU as a technical methodology quantifies the three major elements discussed above and produces numbers, random variables, or some other more-general measures of uncertainty. But this is only part of the methodology, especially since the production of numbers alone demonstrates no connection to the decision process. The methodology that produces the numbers, the formal role of the methodology, and the credible impact of the methodology within the larger decision context are also important. Thus, QMU is also intended to add transparency to the overall decision process.

Figure 1 illustrates concepts related to performance margins and uncertainty. The anticipated load characterizes the environment that the system is expected to see; this load is represented with a band to depict uncertainty, which is most likely aleatoric

uncertainty. The required strength is the documented requirement that the system is expected to meet; such a requirement may be specified in a compatibility definition (CD) document. Often, the required strength is set conservatively above the anticipated load to provide some built-in or "stealth" margin. The proven strength represents the performance demonstrated (with uncertainty, most likely aleatoric uncertainty) during qualification, product acceptance, or surveillance. The difference between the proven strength and the required strength is one measure of a performance margin. In most instances, however, the system would have additional margin because the actual strength resides above the proven strength with some uncertainty (generally epistemic uncertainty, as the actual strength or failure levels have not been characterized). Thus, a QMU analysis that determined there was a lack of a positive margin would not necessarily indicate a system failure unless the demonstrated performance is precisely the actual strength or failure level.

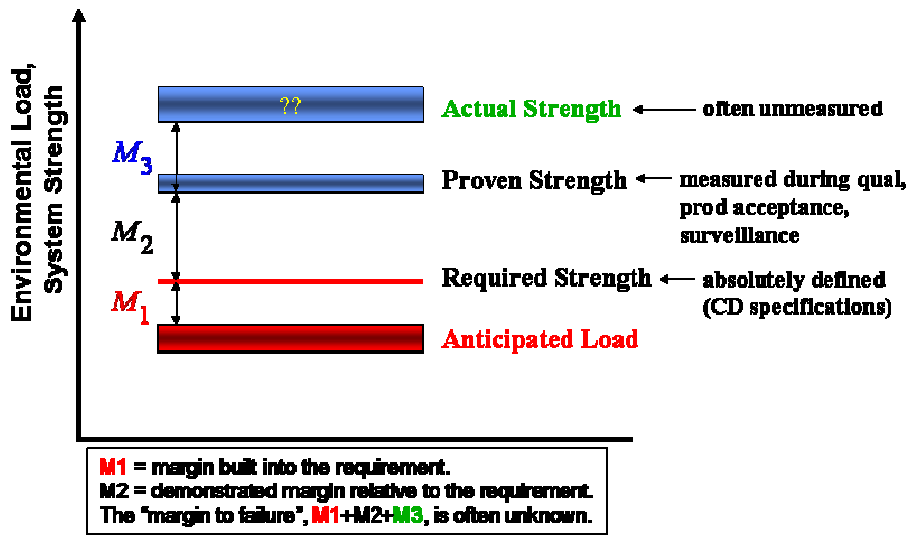


Figure 1. Illustration of margin and uncertainty concepts.

4. Figures of Merit for QMU

Table 1 summarizes elementary terminology in the quantification of margins, including potential measures of importance, i.e., figures of merit, for QMU.[1] Ideally, a single easily-understood metric, with a reference to unity, is sought to convey compliance to life-cycle requirements, regardless of the nature of variability or uncertainties encountered in the decision context. By convention, a margin is typically defined in terms of the nominal difference between the performance threshold (i.e., requirement) and the

[1]The performance metrics are defined relative to the figures in the first row of Table 1, which reflect upper-bound requirements; for lower bounds, the metric definitions would be appropriately adjusted.

assessed performance. As such, a margin alone cannot convey the impact of variability or uncertainties in the decision context.

QMU applications that are centered on assessing the performance of the nuclear explosive package are dominated by epistemic uncertainties (Sharp, Wallstrom, and Wood-Schultz 2004). Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL) have adopted the "confidence factor," *CF*, as the measure of QMU. The confidence factor has its origins in reliability theory, where it is more generally referred to as the reliability index, β; it is also referred to as the K-factor in the statistical literature. When only epistemic uncertainties are considered, a confidence factor that is greater than 1 may be interpreted to mean that reliability is essentially "ONE." (In practice, this latter statement is treated as a certification statement rather than a reliability statement. For example, the interested, and possibly confused, reader should consult Sharp, Wallstrom, and Wood-Schultz [2004].) If only aleatory uncertainties with a normal distribution are considered, a K-factor of 1 implies a reliability of .84.

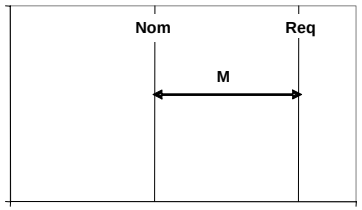
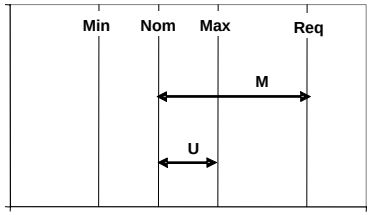
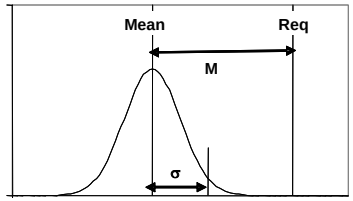
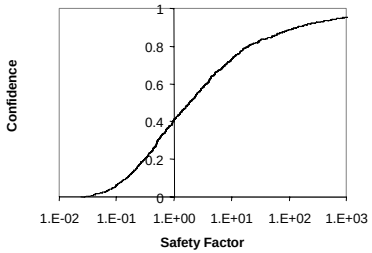
The "safety factor," which is generally defined as the ratio of requirement to assessment, must be interpreted in the context of the requirements or decision language specific to the issue. This factor applies broadly to issues where either epistemic or aleatory uncertainties dominate or where combinations of the two types of uncertainty exist. Confidence in the computed safety factors can be given a statistical specification that represents "standard errors" as well as uncertainties introduced because of lack-of-knowledge issues.

While LANL and LLNL have proposed the confidence factor as a consensus summary QMU metric, Sandia needs to accumulate experience with QMU metrics across the breadth of our stockpile work before we can follow suit. There are cases in our work where a large quantity of experimental data can be, or has been, accumulated and the dominant uncertainties are aleatoric. Application of the QMU methodology to these experimental data–based cases is considered in Appendix A and Appendix B. There are, however, other cases in our work where physical test data cannot be collected because of an inability to conduct high-fidelity tests or because the costs to collect an appropriate quantity of experimental data for a QMU analysis are prohibitive. There may also be cases in which the analysis of experimental data indicates a small margin and the experimental data do not provide enough insight into the various contributions to total uncertainty. In these cases, M&S is required to conduct a QMU analysis, and the approach must deal with both aleatoric and epistemic uncertainties. Application of the QMU methodology to M&S-based cases is described in Appendix C and Appendix D. Ultimately, our QMU methodology will integrate all relevant information sources, including physical and computational results, both of which are subject to epistemic and aleatory uncertainties to varying degrees. Future papers will address methods for this general approach, with combinations of physical and computational simulation results.

As a final point about terminology, we note that the unmodified word "probability" can create confusion when QMU results are communicated. "Probability" can carry a frequency interpretation or a belief interpretation. For example, consider this statement:

The probability of a device functioning properly is 95%. Consistent with the frequency interpretation, the expectation would be that 95 devices out of 100 (on average) would function properly. The belief interpretation would convey 95% confidence (i.e., belief) that all the devices *would* function properly, but it would reserve 5% belief that all the units *would not* function properly. Thus, the strategic implications in a decision-making context could be quite different based on the intended interpretation of probability. Because of this duality in the meaning of probability, we recommend that the words "frequency" and "belief" be directly used. Alternatively, if probability is used in communicating QMU results, we recommend that the proper interpretation be defined explicitly for the decision maker.

Table 1. Elementary Terminology in QMU

Note: These examples assume an upper-bound requirement. The terminology and figures can, of course, be appropriately adjusted to address the lower-bound case.	**No Uncertainties** 	**Epistemic Uncertainties** 	**Aleatory Uncertainties** 
Typical requirements language →	Performance threshold should not be exceeded	Performance threshold should not be exceeded	Probability of exceeding performance threshold should not exceed specified probability
Margin	Nominal Margin **M = Required – Nominal Assessed**	Nominal Margin **M = Required – Nominal Assessed**	Nominal Margin **M = Required – Mean**
Confidence factor (CF), K-factor, or reliability index (β)	Not applicable because uncertainties are not quantified	$\mathbf{CF} = \frac{\mathbf{M}}{\mathbf{U}}$	$\mathbf{K} = \beta = \frac{\mathbf{M}}{\sigma}$
Reliability (Rel) Failure probability	Rel = 1 when M > 0 Rel = 0 for M < 0	Rel = 1 when CF > 1 $p_f < 1$ for CF < 1	Rel = erf(β) for normal distribution p_f = erfc(β) for normal distribution
Safety factor $\mathbf{sf} = \frac{\mathbf{Required}}{\mathbf{Assessed}}$	$\mathbf{sf} = \frac{\mathbf{Required}}{\mathbf{Nominal\ Assessed}}$	$\mathbf{sf} = \frac{\mathbf{Required}}{\mathbf{Nominal\ Assessed} + \mathbf{U}}$	$\mathbf{sf} = \frac{\mathbf{Required\ pdf}}{\mathbf{Assessed\ pdf}}$
Confidence	"High confidence" is asserted if assessed value rigorously bounds lack-of-knowledge issues and numerical errors	"High confidence" is asserted if uncertainties are rigorously bounded for lack-of-knowledge issues and numerical errors "High confidence" is asserted if some sufficiently small percentile of the safety factor distribution exceeds the requirement	

5. QMU in the NW Life Cycle

Formal identification, documentation and tracking of margins and associated uncertainties are initiated during development and qualification and must continue through the life cycle of the weapon. Our confidence in the technical basis that underlies the performance of our weapon systems is developed, maintained, and often enhanced throughout the entire NW life cycle, until the weapon is retired or margins are lost through the aging process. There are five main phases of the NW life cycle: requirements definition, development and qualification, production, stockpile assessment, and retirement/dismantlement. These activities include well-designed tests and M&S, both of which are needed to develop, maintain, and enhance our confidence in the weapons during the stockpile life cycle. Design margins and uncertainties that are quantifiable should be appropriately *identified*, *evaluated*, *tracked*, and *applied* to ensure that our products are robust and meet the design intent.

It is important to be more precise in defining some additional terminology applicable to the following discussions. The terms "performance," "performance parameters," and "critical performance parameters" deserve attention:

> ***Performance*** is the ability of a bomb, a warhead, or a component to provide the proper function (e.g., timing, output, response) when exposed to the sequence of operational environments and inputs.
>
> ***Performance Parameters*** are **all** the parameters required to characterize a system or component design, as well as to characterize the ability of that design to meet its requirements, for which information can be generated through experiments, testing, or M&S.

Each performance parameter can be thought of as having the following:

1. requirement bounds or thresholds
2. margins related to those bounds (margins can be positive or negative)
3. variability (includes unit-to-unit differences and also measurement differences)
4. epistemic uncertainty (e.g., unknown relationships or failure modes for the performance parameter)
5. sensitivity of component or system performance related to small changes in value

Critical Performance Parameters are a subset of all performance parameters that warrant an increased level of attention from system or component engineers.

A performance parameter is considered critical if

1. the physical and/or M&S data generated to characterize that parameter shows a small margin related to the requirements, or if

2. the physical and/or M&S data generated related to the parameter indicate variability that is large or undetermined relative to the margin beyond the requirement for that parameter, or if

3. there is large total uncertainty due to lack of appropriate physical and/or M&S data to characterize the performance parameter, or if

4. the component or system performance is highly sensitive to small changes in the parameter value, or if

5. the physical and/or M&S data related to the parameter indicate a trend that over time likely leads to a small margin relative to the requirements.

The following sections discuss the role of margins and uncertainties for all performance parameters. The critical performance parameters are initially selected in the development and qualification phase of the NW life cycle. The list is amended through the phases of the life cycle as more information becomes available to characterize the current margins and uncertainties. Some performance parameters that were initially characterized as critical may be dropped from the list as the total uncertainty is reduced. Some performance parameters may be added to the list if an aging trend is confirmed, leading to reduced margins or increased epistemic uncertainty or variability. Throughout the life cycle, those performance parameters currently on the critical list should be considered carefully in decision making. Allocating additional resources may help to reduce uncertainty or monitor trends, if uncertainty or trending is indeed the reason for the criticality. Other stockpile-management actions could increase margin or reduce sensitivity.

5.1 Requirements Definition: Basis for Margins

The requirements definition phase of the NW life cycle includes the establishment of system inputs (military characteristics [MC], stockpile-to-target sequence [STS], and interface control document [ICD]) and component performance requirements (compatibility definition [CD]) based on various information sources. Traditional engineering-design approaches have historically been based on a combination of conservative performance requirements, conservative scenarios, and conservative performance assessments. These conservative design practices often meant using the available information and incorporating worst-case environments, 3-sigma enveloping, safety factors, and expert engineering judgment to develop the system requirements.

System requirements (MC, STS, and ICD) are generally propagated to the component level (CD, environmental specification [ES], and nuclear safety specification [NS]). The component response to a system input (often an envelope of a number of system-level environments) is assessed and used to generate component-level requirements. This process increases margins in the requirements that are defined in the component-level documents, i.e., the CD, the ES, and the NS. The accumulation of this conservatism adds to the margins in our systems and components, but can also impact performance and cost.

A desired goal of our engineering approach for the future is to replace this arbitrary stacked conservatism approach for developing requirements with a more quantifiable understanding of margins that forms the basis of a consistent method for managing risks. Epistemic uncertainty and variability can be reduced by applying analysis, testing, and past experience.

It is increasingly important for the design community to have a better understanding of the pedigree of their requirements. This need implies planning and understanding (or quantifying) where margins have been or could be built into the component or system requirements, as well as identification of opportunities for reducing uncertainties. *Documentation* and the ability to accurately *communicate* margins and uncertainties are important throughout the life cycle. The pedigree underlying our knowledge about the component or system allows for more design confidence and ability, when necessary, to appropriately respond to changing requirements without adversely impacting performance.

5.2 Development and Qualification: Establishing Margins and Uncertainties

In the development and qualification phase, components are designed and assessed using the tools of physics and computational simulation to arrive at a design that meets the NW requirements. Typically, there are many opportunities within this phase of the life cycle to demonstrate margins using design hardware and various experiments, testing, and M&S activities. Selection of the appropriate development and qualification testing levels supports increased confidence in performance of the design. In addition to the various experiments and tests, we increasingly use computational models, as they become more refined and predictive, for quantitative exploration of designs and underlying design requirements. The use of models can provide insight into areas of epistemic uncertainty where there is a lack of knowledge as well as a lack of test capabilities or an inability to test. The results from M&S help reduce uncertainties related to (1) scenarios that were not explored during development testing, (2) unmeasured hardware variability, and (3) interactions of complex hardware.

Two examples illustrate the utility of M&S in QMU analyses. For several years, the Enhanced Surveillance Campaign has been developing capabilities for model-based performance analysis in the area of electrical modeling and circuit simulation for nonnuclear components and subsystems. Initially, the electrical models and simulations were developed using PSPICE, a desktop application, but the current approach is based on the XYCE code, which is a supercomputer application. The electrical models and simulations have been applied to firing sets, and the models include detailed representations of all devices in that subsystem. Device-level parameters are varied for each device in the circuit through the minimum and maximum specification limits, which originate in the system requirements, to analytically determine minimum as-built margins for critical firing-set output parameters. Monte Carlo analysis is used to perform sensitivity studies to determine the key devices that could affect these critical performance parameters at the firing-set level. The resulting design-margin analysis

determines the functional limits for key device parameters and resulting performance. These functional limits are then provided to the stockpile evaluation program for monitoring. A number of aging mechanisms have also been incorporated into the models to study their effects on the output. For example, corrosion can be represented as a time-dependent leakage resistance based on a kinetic model for the specified type of corrosion. The degradation of some devices due to enhanced low-dose-rate sensitivity (ELDRS) has also been incorporated as a time-dependent change.

The second example addresses qualification of a weapon system in an abnormal thermal (fuel fire) environment. The system requirement specifies that, given the abnormal environment, the probability of producing four pounds or more of TNT-equivalent nuclear yield does not exceed 10^{-6} (one in a million). A key element of the system design to meet this requirement is the design of weaklinks that fail predictably and irreversibly before stronglink safety switches fail. As system-level testing to meet the one-in-a-million criterion is not feasible, the qualification of a weapon system in an abnormal thermal environment has traditionally relied on first-principles design and component-level testing. Computational advances have made it feasible to add modeling of the thermal race between stronglink failure and weaklink failure. Uncertainties in the fire conditions, uncertainties in heat propagation in the weapon, and uncertainties in the performance of the stronglinks and weaklinks can be combined and analyzed through a weapon system model for this fire condition. The results predict a range of probability of loss of assured safety (PLOAS). Comparing the results with the system requirement allows a statement about the quantified margin and uncertainty to be made.

We expect a robust design with known margins that are sufficient, but not excessive, for high confidence in performance and with minimized life-cycle costs. During the development and qualification phase, a baseline of margins and uncertainties for performance parameters should be assessed and documented. The very large number (hundreds to thousands) of these parameters requires a graded approach. The concept of critical performance parameters was developed to aid in this assessment. Ultimately, the subsystem and project teams must drive toward a manageable, but "critical," few parameters associated with each subsystem and the overall weapon performance for which formal identification, documentation, and tracking of margins and their associated uncertainties are initiated during the development and qualification phase and continued through the life cycle of the weapon.

5.3 Production: Maintaining Margins and Uncertainties

During the production phase of the NW life cycle, extensive actions are taken to characterize and reduce overall uncertainty in the performance of the as-built hardware. To accept the product for war-reserve application, it is important to sample a statistically significant number of new components to compare as-built performance versus as-designed performance. Product acceptance, therefore, generates a useful set of physical data for quantifying margins, variability, and epistemic uncertainty. There are further opportunities within the production phase for applying M&S to help build confidence and further deepen understanding of the margins of as-built hardware.

Product variability, both unit-to-unit and lot-to-lot, is reduced by characterizing and controlling production processes and materials. Identification of the critical performance parameters that need to be monitored during the production process is very important and can be accomplished using various analysis tools. A balance of testing and M&S is needed to characterize and apply probability distributions for the various as-built product parameters. Using M&S to address manufacturing process issues and failures can lead to improved production processes resulting in reduced uncertainties. Additionally, uncertainties in as-built product can be further reduced by characterizing the variability of physical data collected in product-acceptance tests or new-material tests conducted during the production phase. Larger margins may be established through the various testing and M&S activities defined in the production phase.

5.4 Stockpile Assessment: Monitoring for Changes in Margins and Uncertainties

In the stockpile assessment phase, performance has traditionally been measured through different stockpile flight tests (SFTs) and stockpile laboratory tests (SLTs) to strengthen the technical basis of the stated stockpile performance. Special studies and enhanced surveillance activities have also provided opportunities for developing a better understanding of the current and predicted condition of the weapon components, subsystems, or systems. All the experiments and tests, studies, and enhanced surveillance activities are data sources that should be useful and used to monitor margins and establish trends during stockpile life.

The SLT and SFT activities generate physical variables data, but the results have historically been assessed and reported primarily as attributes (pass/fail). The increased emphasis on experimental and test variables data, margins, and uncertainties, as well as on improving our stockpile technical basis, is changing this perspective. User-friendly access to SLT and SFT data is being established through the Integrated Surveillance Information System (ISIS). The precision and accuracy of new test equipment and procedures is being explicitly addressed and documented along with test results.

As-built margins and uncertainties are estimated in the production phase, and steps are taken during this phase to control these margins and uncertainties. During stockpile life the margins and uncertainties are monitored for time-dependent changes or for previously unrecognized sensitivities to storage or operational environments. Early in stockpile life, the stockpile evaluation program also monitors for quality defects that were unknowingly introduced into the stockpile. After about 10 years, defects of appreciable size have been identified (and sometimes repaired), and the evaluation increasingly focuses on time-dependent changes to performance. These changes may manifest as either a decrease in the difference between mean measured performance and the performance threshold, or an increase in variability of the measured performance, or both. Statistical analysis is necessary to perform this kind of assessment; a useful tool for this purpose is the scatterplot of measured performance as a function of age at the time of the test, for all units tested to date. Examples of this kind of statistical analysis are discussed in Appendix A.

Stockpile evaluation has historically been test based. However, where models exist, there are important simulations that should be performed to support the assessment activities. The early identification of uncertainties, along with sensitivity analyses, is helpful in defining the areas where experiments and testing could reduce uncertainty the most and should have priority. A key element for broadly applying M&S to the characterization of margins and uncertainty during stockpile evaluation is the availability of appropriate constitutive models, such as materials models and electrical-device models. A library of such constitutive models should be broadly available to the design community for use in M&S to support stockpile evaluation simulations.

5.5 Retirement/Dismantlement: Monitoring for Subset of Margins and Uncertainties

Retired weapons have no reliability requirements and are not tested or assessed for reliability during the period before the retired weapons are disassembled. However, retired weapons continue to be subject to safety requirements and use-control requirements (for systems with use control capability). A less-frequent test schedule focuses specifically on monitoring the status of critical parameters for safety and use-control components and on identifying any changes in the rest of the weapon that could negatively impact the safety and control functions of these components. Safety-margin and uncertainty assessments are needed for the critical characteristics. Because of the decreased frequency of these tests, it is important to determine whether there are trends in these characteristics and to estimate the rates of change of these characteristics based on such trends. As retired weapons usually must be transported before disassembly, margins and uncertainties in the storage, handling, and transportation environments are needed.

Once again, note that formal identification, documentation, and tracking of margins and associated uncertainties are initiated during the development and qualification phase and must be continued through the life cycle of the weapon.

6. Concluding Perspective

The intellectual framework of QMU provides a common language for NW project engineering and decision support (see Fig. 2). QMU is a decision-support methodology for risk-informed decision making. Given performance requirements, the risk in stockpile stewardship is failure to achieve required or desired performance thresholds and margins. QMU provides information that helps, indeed is necessary, to quantify and understand the various performance risks in the stockpile life cycle, and that contributes to the technical basis demanded by the decision making.

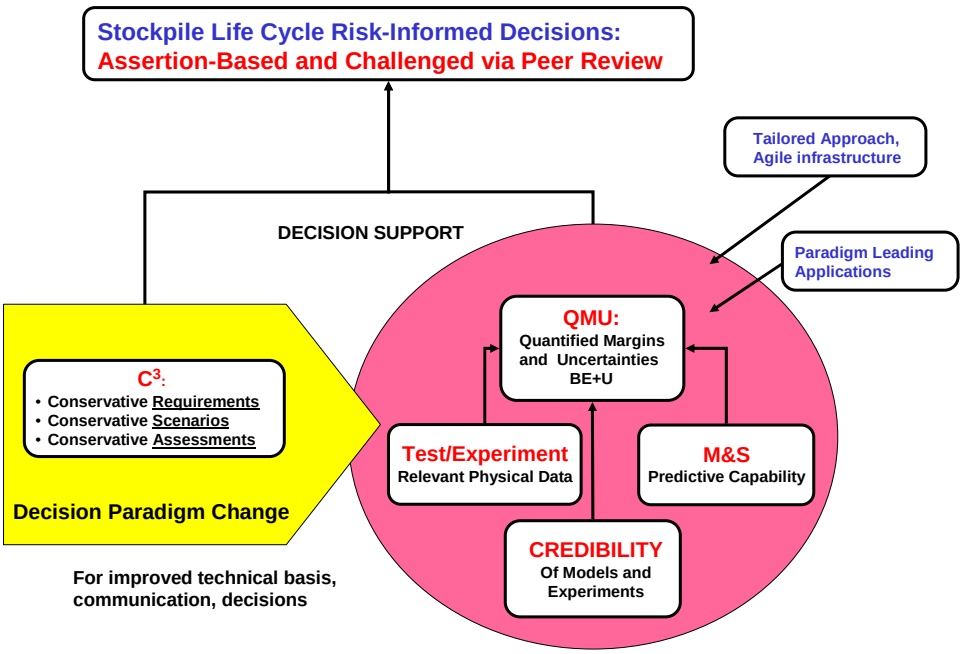


Figure 2. Elements of decision paradigm shift that support science-based engineering transformation through QMU and its connection with risk-informed decisions (Pilch, Trucano, and Helton 2006).

Risk-informed decision making does not base its actions solely on the results of QMU. QMU provides only part of the input into the decision process. Stated differently, QMU is intended for "QMU-informed decision making" rather than for "QMU-based decision making." There are several reasons for expecting a decision process that properly uses QMU would treat it as only one of several important decision dimensions. First, there will be uncertainty in the credibility of QMU results for complex problems, and subjective information will be included in these results. Second, there will always be incomplete knowledge due to both the so-called known and unknown unknowns. Third, there are also factors such as resource limitations (for example, time and money constraints), as well as the social, economic, and political factors external to the relevant scientific information, that will inevitably influence the decision process. Finally, and perhaps most importantly, there is the expectation that human scientific and engineering judgment will always be required in complex technical NW decisions.

Our future engineering environment must provide for QMU as an integral part of all our engineering activities. Our path forward must integrate margins and uncertainties as fundamental elements. Our new focus on margins and uncertainties is in process, but will require continued *support* and *encouragement* by Sandia management to assure its full and effective execution. Adopting a more formal QMU framework should result in an improved technical database for the existing stockpile, accelerated engineering and innovation, and a more credible NW stockpile for the nation!

References

1. JASON. (2005). *Quantifications of Margins and Uncertainties (QMU)*. JSR-04-330. The Mitre Corporation. (3/23/2005)

2. Pilch, M., T. G. Trucano, and J. C. Helton. (2006). *A Conception of QMU*. SAND2006-5001. Albuquerque, NM: Sandia National Laboratories.

3. Sharp, D. H., and M. M. Wood-Schultz. (2003). "QMU and Nuclear Weapons Certification." *Los Alamos Science* 28: 47–53.

4. Sharp, D. H., T. C. Wallstrom, and M. M. Wood-Schultz. (2004), *One versus 1.0*. LA-UR-04-0496. Los Alamos, NM: Los Alamos National Laboratory.

Appendix A – General Guidance for Physical Data–based QMU Analysis

Appendix A provides general guidance for conducting a physical (that is, test or experimental) data–based QMU analysis. This guidance is firmly grounded in past experience. Given the diversity of NW tasks at Sandia National Laboratories (Sandia), general guidance will sometimes be inappropriate. There may be many special cases that arise in such an analysis because of the variety of components, performance variables, test equipment, test environments, test operators, test processes, instrumentation systems, and databases. Experienced statisticians should be consulted to determine whether the methods proposed herein are appropriate for a specific analysis and to provide assistance in conducting more appropriate alternative analyses when these are needed.

A.1 Background

The conservative design practices of the past typically meant using the available information and incorporating worst-case environments, 3-sigma enveloping, safety factors, and expert engineering judgment to define the requirements. To analyze the extent to which the design intent continues to be met over time and under varying conditions, we statistically analyze the amount of stacked conservatism remaining in the component performance at various ages of the tested components.

Several assumptions underlie this discussion. First, we assume that one or more performance variables for a component have been identified. Second, we assume that the measurement of these performance variables, as well as comparison to specified limits, is sufficient to determine whether the component would have performed its intended function in actual use. Because there is a question about which requirements are needed to properly define the performance threshold, we make the assumption that CD requirements should be used to specify thresholds, as well as in evaluating trends and margins. A CD requirement typically comes closest to providing the information that is required for quantifying a threshold for acceptable component performance. As discussed in the main body of this paper, a CD requirement will likely have margin built into it. Therefore, the CD requirement is recommended as a starting point for the performance threshold in QMU analyses. If an analysis shows low margins relative to the CD requirement, further analysis of the actual system-failure level should be conducted.

As stated previously, a margin in the physical simulation experimental data–based QMU analysis is the difference between the expected component performance and the requirement. Without loss of generality, we consider a requirement to be the minimum level of component performance that is acceptable. “Negative margin” means that the expected component performance is less than the requirement, a clearly undesired state.

A.2 Time-Dependent Change

As a component sits in dormant storage in the stockpile for decades, there may (or may not) be changes in the distributions of one or more of the relevant performance variables.

These changes may be due to accumulated environmental exposure that degrades materials; chemical reactions that are generated by material incompatibilities, fatigue, or wear out (unlikely in dormancy); mechanical stress relief; low-dose radiation; or any combination of these and potentially other effects. Measured values of a performance variable may creep up or down, or the range of values may grow larger or smaller. We are most concerned if the measured values creep toward a limit or if the range of measured values expands toward the limit. Ideally, design and qualification activities for the component and the system account for anticipated changes to performance variables, as well as the influence of environmental and operational variables on performance. Design and qualification activities also seek to maintain adequate margin during the designed life of the component.

It is informative to plot certain characteristics of the statistical performance distributions, such as means or standard deviations, against age to estimate the change with age. Figure A-1 shows a notional plot of mean component performance (as the mean of a single variable) versus component age. The curve shows a downward trend with age. The notional data presented in the plot show increasing measurement error bars with increasing component age as well. A notional requirement on the performance variable threshold is shown by the red dashed line. The plot thus demonstrates an example of decreasing performance with increasing variability, with possible failure at some point in the future.

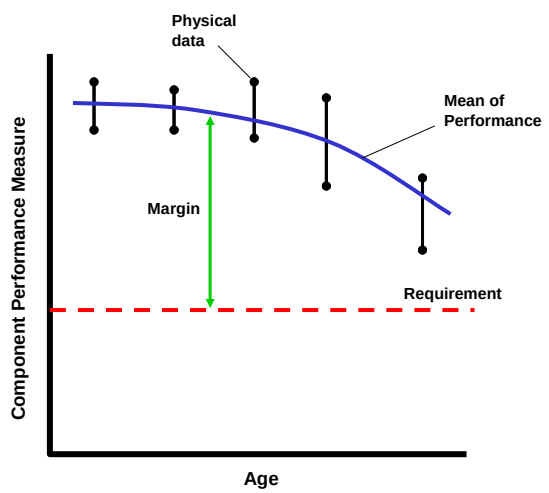


Figure A-1. Trend line for mean performance.

The ideal goal of a physical simulation data–based QMU analysis is to characterize distributions of performance, relative to a CD requirement, as a function of component age. Only in rare cases will sufficient information be available to estimate the distributions for baseline and aged states. Partial information is still useful, and conclusions can be drawn that support stockpile evaluation. In particular, we can estimate the mean and the standard deviation of performance as a function of age, and thus the ratio of margin to uncertainty (variability) as the component ages, as is illustrated previously in Figure A-1.

A.3 Figure of Merit

The figure of merit for a statistical physical data–based QMU analysis is M/U, the estimated margin of a component performance variable for a population, divided by an estimate of the variability in the performance variable over the population. For a physical simulation data–based QMU analysis, the interpretation of "margin" is mean performance minus required performance, and the interpretation of "uncertainty" (that is, variability) is the population's standard deviation. This figure of merit corresponds to the K-factor in Table 1 of the main body of this paper. The data should be examined for time dependence; if there is a time-dependent trend, the figure of merit must be estimated as a function of component age at the time of measurement of the performance variable. A confidence bound for the mean trend line can be calculated from the standard deviation of the margin estimate (sometimes called the standard error of the margin estimate). More detailed guidance is presented in Appendix B.

A.4 Physical Data

To conduct a physical data–based QMU analysis, the following information and physical data are needed:

- Identification of the component and its performance variables
- Identification of other similar components whose physical data may be statistically pooled
- Sources of physical performance data, including product acceptance processes, surveillance, and reacceptance processes. Information about test conditions is also required, especially component age.
- CD requirement

A.5 Populations

We define a population as consisting of all manufactured units of a particular design. The set of all firing sets, for example, consists of multiple populations because there are multiple firing-set designs. Each major-component number refers to a distinct population whose performance parameters under certain interface and environmental conditions are estimable from tests of sampled units of the population. Changes to suffix numbers of the serial-numbered units may represent design changes and thus distinct subpopulations. In many cases, we have known subpopulations within the general population of all manufactured units of a particular design. The subpopulations may arise from manufacturing process changes that are sufficiently large to affect one or more performance variables. Examples of types of manufacturing process changes are material changes, process changes, and manufacturing equipment changes. The impact of such changes can be judged in a plot of the performance-variable measurement against manufacturing date. The impact might be a shift in the measurements, or an increase or a

decrease in variability of the collected data after the date of the change. A statistician can perform the appropriate tests of statistical significance to determine whether an apparent shift, trend, or change in variability falls outside of the expected variation of the data prior to the change. If so, the change in material, process, or manufacturing equipment must be accounted for by subdividing the population into two distinct subpopulations based on the date of the change. The distinct subpopulations must have separate margin analyses, trend analyses, and uncertainty analyses.

A.6 Physical Data Samples

With the exception of newly introduced or infrequently tested components, most component populations in the stockpile have a long history of different types of tests with different measurements at different times on different units. Examples are qualification tests, production-acceptance (E and D) tests, new-material lab tests, new-material flight tests, stockpile lab tests, stockpile flight tests (SFTs), reacceptance tests, shelf-life tests, and tests conducted under special instruction engineering releases. The question that must be answered is, Which test data should be used for an experimental data–based QMU analysis? Test environmental conditions, test inputs, degree of accurate representation of test units for the population, and measured component output are all considerations in selecting the physical data sample for the QMU analysis. The selected data must be measured under environmental conditions that are representative of use conditions that could affect performance. The selected data must have test inputs that are representative of actual inputs during use. The units from which physical data were taken must be a random sample of the war-reserve population, or, if not, the reasons why the units were selected should be available. The collected data sample must include measurements that may be used to infer performance in actual use.

If the test units are randomly sampled from their populations, statistical inferences may be made about the applicability of conclusions drawn from the sample to the population. For example, the sample may be drawn to estimate a parameter of the population, such as mean performance. Then a range of possible values of the population parameter may be computed from the sample with a defined likelihood (a statistical confidence level) of bounding the true population parameter. This range is called a (statistical) *confidence interval*, and is usually labeled with the confidence level. Confidence intervals for high levels of confidence are wider than confidence intervals for lower levels of confidence. For consistency, we recommend calculating 90% confidence intervals.

Every measurement of component performance is subject to variation. Measurements are affected by unit variation; thus, every manufactured unit will be somewhat different from every other manufactured unit of the same design. Measurements are also subject to environmental variation; thus, changes in the environment may cause subtle or dramatic changes in the test units. Separating out the contributions to variation from environmental causes can be done through a *variance components analysis*. To conduct this type of analysis, a series of repeated measurements on different units at different environmental conditions is needed. Obtaining estimates of each of these variance components allows proper accounting for these sources in an uncertainty analysis.

If obtaining the physical data needed for a variance components analysis is not possible, other information allowing isolation of environmental effects may be available. Measurements are also subject to accuracy and precision limitations of the measuring instrument; thus, every measurement will vary due to instrumentation that is not completely stable, even on the same unit. Before a test instrument is put into service, a study of the instrument's accuracy and precision is usually completed; calibration reports provide updates to accuracy and precision calculations at specific points in time. It is good practice to obtain the results of accuracy and precision studies, and calibration reports, if they are available. (Note that these types of reports may not be available for test instruments that are no longer in service.) In the ideal case, the instrumentation bias is negligible, and the variability introduced by the measurement instrument is much less than the unit-to-unit variation. In this case, there will not be a need to specifically account for the instrumentation bias. On the other hand, some evidence suggests that similar measurements on different test instruments yield results that appear to be biased relative to each other. If this kind of bias is present, then statistically adjusting the experimentally measured values from the different test samples to account for the accuracy and precision of the test instrument used allows the pooling of these data for QMU analysis. An experienced statistician should be consulted in this situation.

A.7 Distributions

It is good practice, before calculating any summary statistics from a physical data set of performance variables, to construct a bar chart, a stem-and-leaf plot, or some other graphical display of the data to see the range of values, how those values are distributed, whether there are patterns in the data, and whether any outliers are present in the data.

Outliers are values that are far outside the typical range of variation; a stem-and-leaf plot specifically identifies outliers. The technical definition of outliers can be found in Koopmans (1987). One must determine the cause of the outliers before deciding how to statistically analyze them. Outliers may be an indication of uncontrolled environmental influences on either the test equipment or the test unit. The cause determines how to handle the outlier values. If due to equipment sensitivities, the outliers may be excluded from the QMU analysis. If test units are sensitive to various uncontrolled environments, the physical data set must be split based on environmental changes, and separate QMU physical data analyses must be conducted for each separate data set.

An example of a pattern in collected data is bimodality. Bimodality is a clumping of the data around two distinct values, as shown in Figure A-2. Bimodality can occur, for example, when a new test instrument is introduced, a change is made to the test protocol, or a material is changed in production. To obtain clues about the causes of bimodality, we recommend examination of available metadata (information about the test equipment and conditions under which the data were collected). Plotting the measured values against test dates is a good place to start. If a test date can be identified that separates the two groups, then one must determine what happened around that date that caused the change. Another factor that may signal bimodality is the production date. A shift in performance values at a specific production date may indicate an underlying cause of a change in vendors, materials, or parts.

Bimodality may be empirical evidence that points to the existence of a previously unrecognized subpopulation. A number of statistical tests may be applied to conclude the existence of distinct subpopulations. Such tests are based on the degree of separation of the two potential subpopulations and the existence of one or more factors that can be used to partition the population into subpopulations. These factors must be identifiable for each unit of the sample and the total population. Obviously, the factors must be recorded and available for each unit in the sample and the population. If the statistical test rejects the hypothesis that the data came from one distribution, then it is necessary to partition the data into separate subpopulations and perform separate margin, trend, and uncertainty analyses for the subpopulations.

Combining distinct subpopulations for one common analysis of margins, trends, and uncertainties is undesirable because it may lead to erroneous conclusions. The margin estimate from inappropriately pooled physical data sets may be too low, and the variability estimates may be too high. This situation occurs because the known difference in subpopulations is incorrectly included as an unknown in the uncertainty estimate.

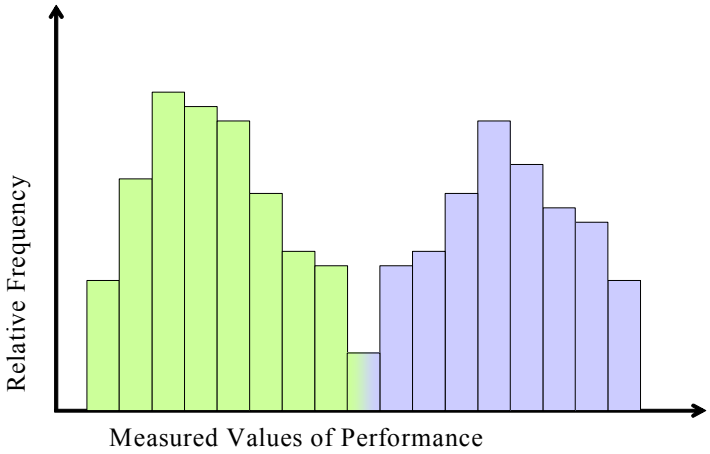


Figure A-2. Histogram of measured performance values exhibiting bimodality.

Another example of an unusual pattern in physical data is strong skewness in the data set. Skewness is defined as lack of symmetry in the distribution of measurements around the mean. By chance, measurements may appear to be slightly skewed in random samples. A statistician can conduct a statistical test to determine whether the observed skew in a random sample is too large to be due to chance. If the skewness is found to be statistically significant, the data sample cannot be modeled as a Gaussian distribution. Yet another example of an unusual pattern is kurtosis, overly heavy (or light) tails in the distribution as compared to a Gaussian distribution. Kurtosis is usually difficult to detect with small or moderate sample sizes. A statistician can also test for kurtosis, but the test is limited by the adequacy of the statistical data sample for this purpose.

In the absence of outliers, bimodality, skewness, and kurtosis, a statistician can determine whether the measurements, or some transform of the measurements, are adequately described by a Gaussian distribution. If the measurements (transformed or not) can be

described by a Gaussian distribution, then probabilities can readily be associated with various values of margin normalized by uncertainty. If the data are not Gaussian distributed, then either (1) a different distributional form must be fit to the data or (2) a larger physical data set must be collected to estimate empirical probabilities for the underlying distribution.

A.8 Margins

A *baseline mean performance* estimate is needed to calculate a baseline margin. The term "baseline" usually refers to the performance of a system or component early in its design life. If possible, obtain physical qualification data for stockpiled components to estimate the baseline mean performance. Alternatively, early production tests are a good physical data source. For components that were manufactured decades ago, it may be difficult or impossible to find qualification or early production data to establish the baseline mean performance. In some cases, a specific performance variable was not measured in early production, and measurements began at some later point during surveillance testing. In these cases, the baseline mean performance may be defined at any identified point in time, including the present time. It is necessary to record and report the baseline time period for future reference and interpretation.

The baseline margin is the baseline mean component performance minus the required performance. There should be a margin estimate for each distinct subpopulation.

A.9 Uncertainty

Variability refers to stochastic variations in measurements due to the sum effects of a number of causes, none of which is dominant. Variability can be statistically characterized with a probability density function (PDF) or an equivalent cumulative distribution function (CDF). The commonly accepted summary measure of variability is standard deviation. Some examples of variability are as follows:

- Unit-to-unit differences of nominally identical components
- Lot-to-lot differences of nominally identical materials
- Instance-to-instance differences of nominally identical environments
- Lack of measurement precision

Epistemic uncertainty is a general term that refers to imprecise or incomplete knowledge about component performance. In particular, we may believe a measurement is subject to random effects, but we may have little or no physical data to characterize those random effects statistically. Typically, in statistical analyses, epistemic uncertainty is represented as a set of discrete possible or plausible choices (e.g., model choices). Epistemic uncertainty can be reduced by targeted studies. The following are some examples of incomplete knowledge uncertainties:

- Unmeasured variability due to very small sample sizes
- Unknown and unobserved manufacturing defects
- Scenarios not represented in the test activities
- Incomplete understanding of complex interactions

In a physical data–based QMU analysis, we primarily account for population variability in the denominator of M/U, rather than epistemic uncertainty. It is appropriate to conduct sensitivity analyses to address identified epistemic uncertainties. If our knowledge about the exact proportions of two subpopulations in the stockpile is incomplete (for example, due to missing or incomplete records), we can compute the M/U factors for each subpopulation separately. Then we can hypothesize a range of values for the proportions of the subpopulations in the total population. We can recompute the figure of merit for each of the two extreme values (minimum and maximum) defined by the range of proportions for the subpopulations. The uncertainty in the figure of merit due to uncertainty in the subpopulation proportions is the range of the two computed figures of merit. This range represents possible values, not a statistical distribution of values, of the figure of merit.

A.10 Trends

Looking for trends is an important part of QMU physical simulation data–based analysis. A key question is whether the current performance margin is essentially the same as the baseline performance margin. Performance margins may not be constant for a number of reasons. Importantly, the underlying causes may be attributable to physical changes in the materials that cause changes in the performance variable. Some examples of such physical changes are material incompatibilities leading to ongoing chemical reactions like corrosion; stress relief leading to broken seals, delamination, or deformation; mechanical fatigue leading to deformed or broken parts; and diffusion through membranes leading to loss of hermeticity. Typically, all performance variables must be reviewed for trends. If a trend is identified and confirmed, an explanation must be sought for the underlying phenomena that caused the trend.

The simplest analytic tool for identifying trends is the age-based plot of a performance variable, as illustrated previously in Figure A-1. Because a number of other factors besides age can affect the performance variable, one should use plotting and statistical analysis for separating out these effects. For example, if temperature affects the performance variable, age-based plots should be made for each test temperature. A statistical analysis of covariance or an analysis of variance can also be conducted to account for both environmental effects and age-based effects. A statistician should be consulted to conduct these types of studies. An appropriate statistical analysis can determine whether a trend is too large to be attributable to chance, given the unit-to-unit variability and the measurement variability.

Certain types of physical data are preferred when age trends are being investigated. If measurements are nondestructive, and the performance variable can be measured on a unit multiple times during its life, the most efficient method of looking for age trends is to obtain data from the same set of units at different ages. The differences between the measured performance variable on the same unit at different ages are analyzed, thus eliminating the effect of the unit-to-unit variation in the analysis. The unit-to-unit variation is typically a large source of variation (based on experience with the legacy nuclear weapons stockpile) and may obscure small trends. If the measurement is destructive, it is essential to obtain test data from other units at different ages in the same lots. A less-desirable method is to obtain test data from different units of different lots at different ages. These data may be analyzed for an age trend, but the trend may be too small to be observed with the measurement, unit, and lot variation.

We can think of an age trend as a change to some aspect of the statistical distribution of the performance variable. The mean of a performance variable may decrease with age, or its standard deviation may increase with age, signaling increased variability. Any of these changes may result in an untenable margin loss if sufficiently large, so all such changes are of concern. With limited age-related physical data, it may be difficult to definitively confirm the nature of the changes to the distribution function. There are statistical techniques to test hypotheses such as a changing mean, a changing standard deviation, or a changing distributional shape. It is important to note that methods exist to test and model nonlinear change. These techniques rely on data in sufficient quantity and of sufficient quality, hence constraining the type and amount of data that should be gathered.

Stability in the performance variable is the ideal. When a time-dependent change in the performance variable is observed, the change raises concerns about whether the associated performance margin may be unacceptably reduced over the stockpile life of the component. If an age trend is observed and confirmed through statistical analysis, a statistical model of the age trend is desired for margin analysis. Some age trends are so small that they cause no serious concern about margin loss during the stockpile life. With a statistical model, the analyst can estimate the age at which margin loss equals the unreliability allocation (or some other reference value) and report that age with confidence bounds as the reliable lifetime. Alternatively, the margin loss and resulting increase in unreliability can be estimated, with confidence bounds, at a specified age.

A.11 Summary

General guidance has been presented in this appendix for conducting a physical simulation data–based QMU analysis of a component performance variable. This guidance is based on statistical techniques for understanding mean performance and the variation of performance, as a function of age, from acquired observational data. For a valid QMU analysis, it is critical to determine whether a set of test measurements can be analyzed as a single population and also to determine what distribution adequately reflects the observed variation. Guidance is given for various situations where the physical data contain multiple subpopulations of either units or test conditions. The

assistance of a statistician is recommended to improve the validity of a physical data–based QMU analysis.

Specific application of the general guidance discussed in this appendix is provided in Appendix B. The specific application covers a physical simulation data–based QMU analysis that was performed for the FY07 Annual Assessment and Stockpile Review Conference.

A.12 Reference

Koopmans, Lambert. (1987). *Introduction to Contemporary Statistical Methods*. Boston: Duxbury Press.

Appendix B – Guidelines for Common Data Analysis and a Presentation Using Physical Data

B.1 Introduction to QMU Concepts as Employed in the Analysis

The purpose of these guidelines is to establish a common approach to the analysis of physical data for QMU. The approach followed here relies on various standard plots of data to provide a basic understanding of the structure of the data and to suggest possible further investigations if needed.

Terminology consistent with the main body of this paper is used in this appendix. A margin is defined as the difference between a performance threshold specification and the measured (demonstrated) performance relative to that threshold. We recommend using the specifications in the weapon system's CDs as the appropriate performance threshold values to begin an analysis. However, it is important to recognize that, typically, margins are built into the CD specifications. Therefore, if a QMU analysis is conducted showing inadequate margins using CD specifications, a refinement of that analysis taking into account actual failure levels should be initiated. This refinement may require testing to determine actual failure modes and failure levels. In the cases where a measured response has both an upper threshold and a lower threshold, there are two margins. When present, both margins are of interest, and the analysis and plotting are pursued with each.

A direct measure of the margin is the difference between the measurement and the threshold for that measurement. Thus, for an upper performance threshold, PTU, and a measured response, R, on a given unit, the upper margin is given by $PTU - R$. Similarly, for a lower performance threshold, PTL, the lower margin is given by $R - PTL$. The units for a margin are the engineering units for the measured quantity. The only concept of the adequacy of a margin measured on a specific unit is that it should be positive in order to satisfy requirements. Further, a larger margin is better. However, a measure of goodness for a population as a whole is provided by looking at the distribution of individual margins. The degree of variation for individual margins reflects the "uncertainty" (that is, variability) with respect to the population of units. Therefore, standardizing the margin by dividing by the population standard deviation provides a measure of comparison that can be used across different responses. In essence, the standardized margin is expressing the margin in terms of multiples of the population standard deviation.

The analysis process outlined below is based primarily on the ability of scatterplots to provide visual assessments of data irregularity. A primary interest is whether the margins (equivalently standardized margins) are changing with respect to the age of the unit under test. If a time-dependent trend is evident in the data, the analysis will predict the age at which a specified fraction of the population will fall outside the chosen thresholds, with a 90% confidence bound. If no time-dependent trend is evident in the data, the analysis will provide a best estimate with confidence limits for the margin above the threshold. However, these trend estimates are based on the assumption that it is the margin that is changing with age, not the population spread. The final step in the analysis is to assess whether the data support this assumption.

B.2 Assumptions

The following assumptions underlie the description of physical data analysis presented in this appendix.

1. Physical response data are available for the component or the system.

2. Physical response data are available with the following basic structure.

Serial number (SN)	Production date (PD)	Test date (TD)	Response (R)
.	.	.	.
.	.	.	.

NOTE 1: There may be multiple pairs of test dates and responses for specific serial numbers. If these data sets originate in different test conditions (yielding different response data), then they should be considered as different statistical populations for statistical analysis purposes (see assumption 4). However, if the distinct test date–response pairs represent different test dates, such as when the same unit was included in two different stockpile cycle returns, for statistical purposes the data presentation should contain additional columns (TD_2, R_2) rather than having the SN repeated in the list of data. The data included in the TD and R columns should be the data that reflect the greatest age of the unit under test, that is, the data gathered from the most recent test.

3. Critical performance thresholds (*PTL* and *PTU*) exist for each response. The limits may have a direct relationship with the reliability of a component, but, at the very least, they represent actionable decision points. While the general case is the existence of both a lower performance threshold and an upper performance threshold, any given response may have only a single-limit lower performance threshold or upper performance threshold.

4. Additional factors related to the response, such as environmental levels (hot, cold, different stress levels, and so forth) for the testing, may exist, but it is assumed here that the physical data have been partitioned into the distinct levels and that the following analysis (in Section B.3) would be carried out for each level of these factors. Although a decision may follow the initial analysis that the data can be recombined, the starting assumption is that the response data resulting from individual factor levels are characterized separately.

5. The uncertainty in repeated measurements, represented by a standard deviation (σ_m), is known.

   NOTE 2: There are multiple sources of uncertainty that need to be considered in a full measurement capability study. Here, attention is restricted to the uncertainty that can be characterized by short-term repeated measurements. This is the primary source of variation that is needed to judge the significance of different measurements from two distinct items measured under the same conditions.

NOTE 3: The repeated-measurement standard deviation (σ_m) may not be readily available in many cases. However, we assume it is available so that it can be incorporated into a common presentation and analysis.

B.3 Common Steps for Analysis and Presentation

The steps presented in this section are illustrated with a set of data consisting of 86 responses. The plots and analysis shown here are capable of being produced within the Sandia-developed Waveform Analysis and Viewing Environment (WAVE).

1. Collect the physical response data in the format specified in Section B.2 for the analyses.

2. Plot the physical response data versus production date (*R* versus PD). Then calculate $\overline{R}$, the mean of the response data, and draw a horizontal reference line at $\overline{R}$. If an estimate of σ_m is available, error bars of length $\pm\cdot\sigma_m$ are shown at the top of the plot. Note that the variability in the data is affected by other factors besides measurement uncertainty, such as unit-to-unit variation. If the spread in the data is much larger than the error bar, this indicates that measurement uncertainty is not the major contributor to variability.

   Figure B-1 illustrates the kind of response versus production date plot that we have recommended. Note that this figure represents measured responses at varying component test dates and ages plotted against the production date of the unit. For one-shot devices, the measured responses are not available at the time of production. If there are some very significant production anomalies, this plot should reveal them. An example is a shift of responses in some interval of time when the process went out of control and was eventually brought back under control. If such anomalies are significant, the physical data set should be further segregated into subpopulations based on production date.

3. Plot the physical response data versus test date (*R* versus TD). Draw a horizontal reference line at $\overline{R}$. If an estimate of σ_m is available, error bars of length $\pm\cdot\sigma_m$ are shown at the top of this plot also.

   Figure B-2 represents measured responses at varying component ages plotted against the *test* date of the unit. If there are some very significant testing anomalies, in comparison to production anomalies or aging effects, this plot should reveal them. If possible, identify whether a change to the test hardware or test procedure coincides with the dates of the anomalies. If so, then the physical data set can be partitioned by these testing-related changes for further analysis.

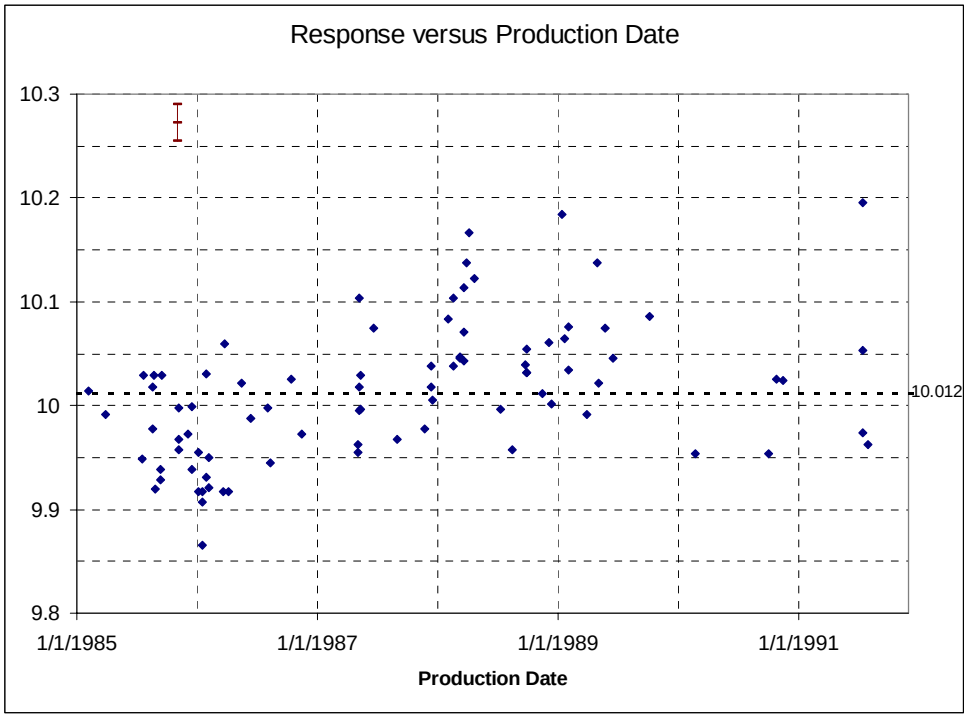


Figure B-1. Illustration of the recommended response versus production date plot.

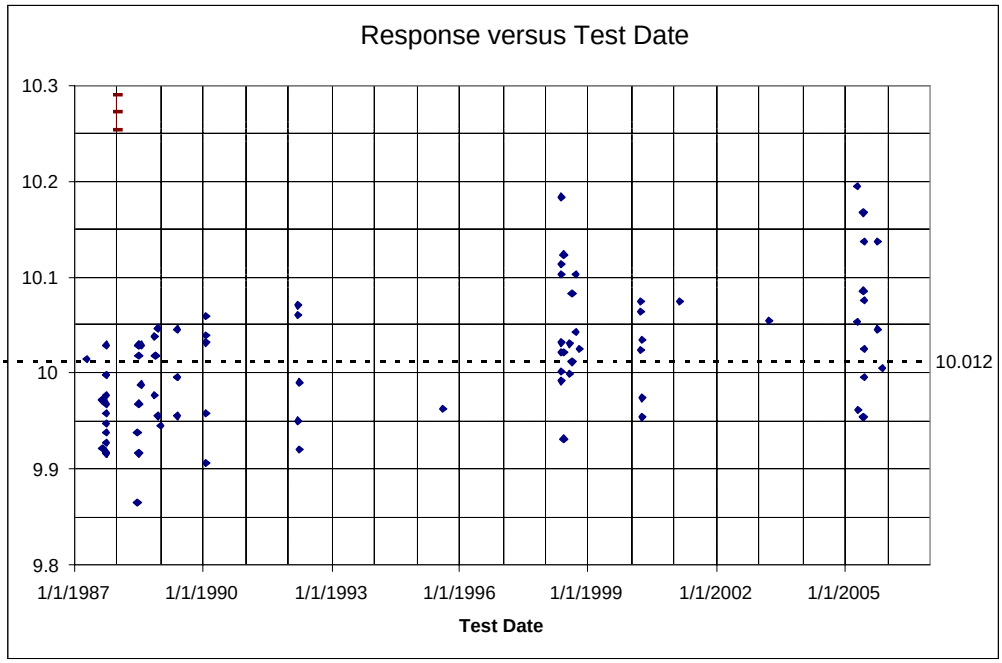


Figure B-2. Illustration of the recommended response versus test date plot.

Discussion of two previous plots. The ultimate goal of the analysis is to judge whether the response seems to change with increasing age of the units. The first two plots (Figs. B-1 and B-2) provide information on the extent to which production anomalies or testing anomalies might influence the response measurements. The analyst should determine the influence of identified production or test anomalies before looking for age effects because the anomalies could mask aging effects or lead to a false inference of aging effects.

In Figure B-1, there is an indication of a slight increase in the response with production date, which seems to stabilize after 1988. There is no indication of sudden shifts in the performance measure over a small time interval in either Figure B-1 or Figure B-2. Sudden shifts may be taken as evidence of process-control problems or adjustments. In Figure B-2, there seems to be a slight increase in the response with test date, but no apparent shift in test results, which is sometimes an indicator of a new test instrument or test protocol. There is also an indication of increased variability in Figure B-2 in the tests around 1998 and 2005. The 1998 increase is due to two data points, a high value and a low value. The 2005 data have a more general spread. The measurement error indicates that none of the individual responses are extremely "atypical" from other responses. There do not appear to be any significant production anomalies or test anomalies affecting the response, and thus there is no need to partition this physical data set.

4. Calculate the overall response mean $\overline{R}$ and the overall sample standard deviation s_R. Calculate the margin quantities $M_{PTU,i} = PTU - R_i$ and $M_{PTL,i} = R_i - PTL$, where i indexes individual unit responses. These quantities express the distance of the data from the performance thresholds. Note that both margin quantities should be positive to meet the requirement of being within the performance thresholds and that both margins are expressed in the engineering units of the original response. In addition, the scaled quantities $R_{PTU,i} = \frac{M_{PTU,i}}{s_R}$ and $R_{PTL,i} = \frac{M_{PTL,i}}{s_R}$ represent "standardized" responses, where the distance from the thresholds (0 corresponds to the threshold) is expressed in the number of population standard deviations. This standardization enables a more meaningful comparison between different responses in totally different engineering units. Then create two plots, one for the upper margin (R_{PTU}) and one for the lower margin (R_{PTL}), against age of the unit at the time of the test ($TD - PD$). For reasons of visual comparison, both plots should include 0 on the vertical scales. Note that a single plot can be used to represent both the margins and the "standardized" margins, as they only differ by a scale change. The reference line going through the mean of the data provides an estimate of the K-factor associated with the population when the trend is not considered.

   For the example under discussion, $PTU = 10.26$ and $PTL = 8.4$. The variable $TD - PD$ is in units of years. The plots are shown in Figures B-3 and B-4.

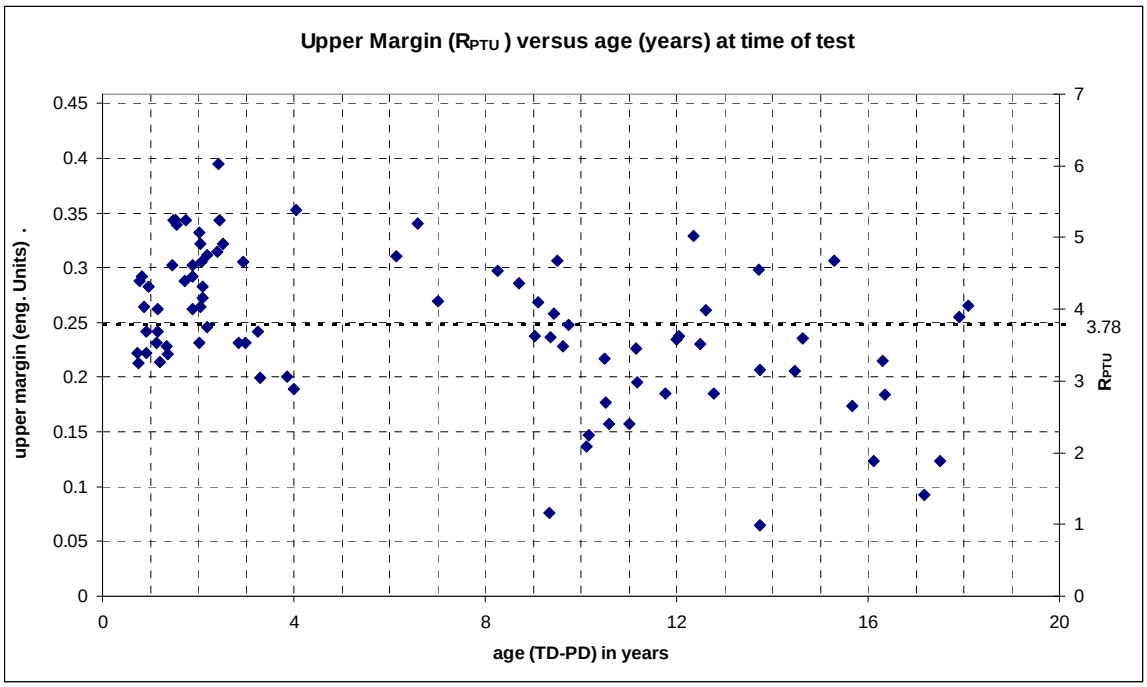


Figure B-3. Illustration of the recommended upper margin versus age plot.

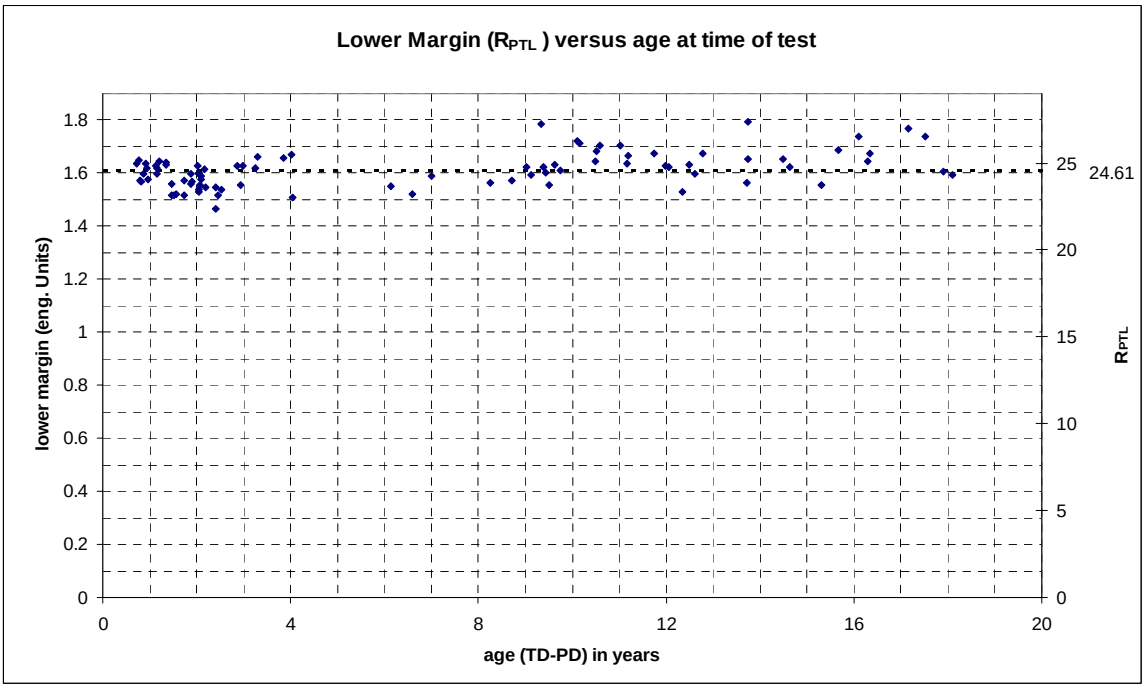


Figure B-4. Illustration of the recommended lower margin versus age plot.

5. The two previous plots in Figures B-3 and B-4 summarize the same physical response data, but mirror each other: when one has an upward trend, the other will trend downward. To assess the impact of possible trends in the data for each plot, perform a regression of $M_{PTU,i}$ and $M_{PTL,i}$ against age. Regression is a widely used statistical analysis that models the effect of one or more independent variables (x) on a dependent variable (the response). In the implementation here, there is one independent variable, age, and two dependent variables, M_{PTU} and M_{PTL}. The result will be an estimate of the mean margins as a function of age, $\bar{M}_{PTU}(x) = c_{PTU} + d_{PTU} \cdot x$ and $\bar{M}_{PTL}(x) = c_{PTL} + d_{PTL} \cdot x$, where x is the age of a unit.

   The slope or age trend estimate will be the same in magnitude for both of the computed mean margin regressions. One will be the negative of the other (that is, $d_{PTU} = -d_{PTL}$), as the data can only trend toward one of the performance thresholds. However, due to uncertainty surrounding the estimates, there may be reason to display both. This point is addressed again in step 6 when confidence statements concerning the trend are converted to a confidence about an "alarm age."

   From the regression of M_{PTU} on age using the WAVE software, the quantity s_p, which is the estimated standard deviation of the response about the mean regression line, is obtained. In almost all cases, $s_p < s_R$. An exception can arise when the estimated trend is 0 or very close to 0; the exception is a result of using a divisor of $n-2$ for estimating s_p when $n-1$ is used for s_R. An adjusted R_{PTU} is obtained by dividing by s_p; that is, the adjusted values are defined as $R'_{PTU,i} = \frac{M_{PTU,i}}{s_p} = \frac{PTU - R_i}{s_p}$. Note that this quantity differs from the earlier R_{PTU} in the use of the regression standard deviation instead of the overall population standard deviation. In almost all cases, $R'_{PTU,i} > R_{PTU,i}$. The mean line of the adjusted standardized variables, $R'_{PTU,i}$, plotted against age, is the (sub)population K-factor from which the proportion of the (sub)population expected to fall outside of the performance thresholds on the response R can be inferred as a function of age.

   The plots discussed in step 3 are updated by showing the mean regression line and updating the right-hand scale to reflect $R'_{PTU,i}$ ($R'_{PTL,i}$) instead of R_{PTU} (R_{PTL}). Added to the plot are 90% prediction intervals to help interpret irregularities in the spread of the data. A 90% prediction interval plotted versus age provides a range in which the probability is .90 that a future data point at that age would fall. The interval also provides a convenient manner for viewing the current data, as it should be the case that roughly no more than 1 in 10 of the given data points fall outside the interval.

   Figure B-5 illustrates the kind of plot generated by the actions discussed in step 5.

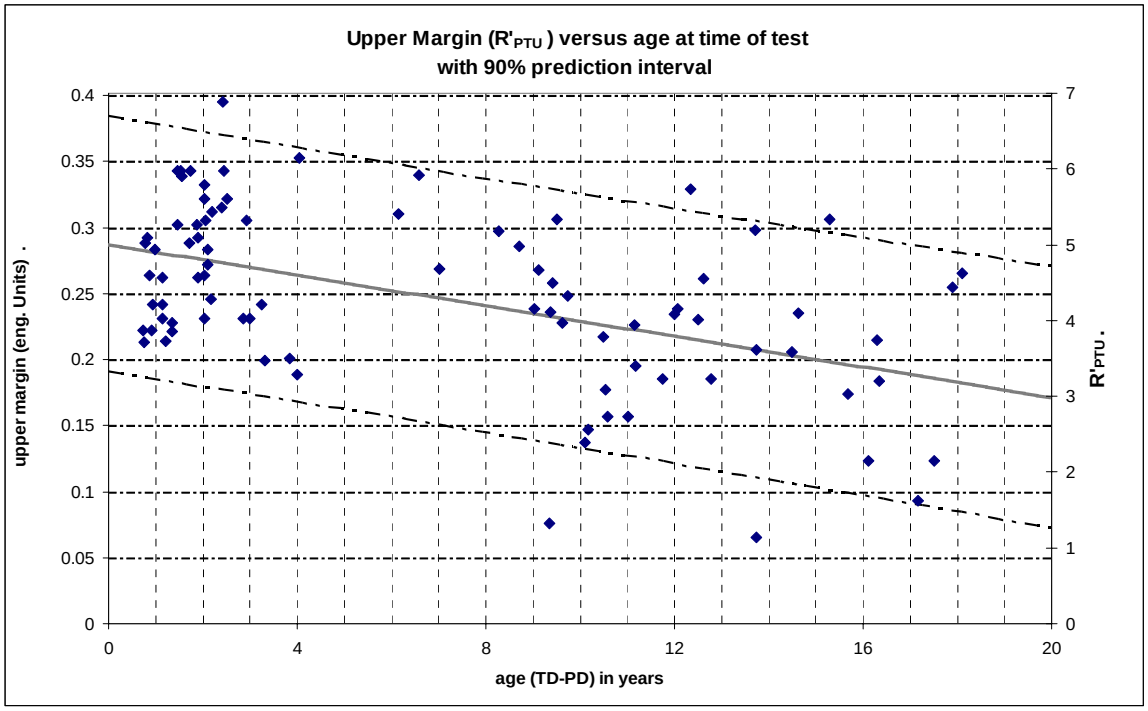


Figure B-5. Illustration of the recommended (and updated) upper margin versus age plot.

6. The next step is to translate the knowledge gained about possible trending in M_{PTU} and M_{PTL} in steps 3 and 4 to an estimate and uncertainty about the estimate of component age at which an alarm will be raised. To facilitate comparisons across components, we have defined the alarm age to be the component age at which it is predicted that a certain percentage of the population will fall outside the performance thresholds. This is accomplished by considering a confidence interval placed around the mean regression line with respect to the age-trended K–factors (R'_{PTU} and R'_{PTL}). These factors, by their very nature, are population related; that is, the standardization by the population statistics $\bar{M}_z(x)$ (where $z = PTL$ or PTU) and s_p means that these quantities are summaries of population characteristics. These standardized quantities can be referenced to probability-of-compliance levels through the assumption that the distribution of the responses is approximately Gaussian (normal) around the age-regressed mean. (Gross departures of the data from satisfying this assumption are tested in step 7.)

   To illustrate the connection to compliance levels, choose a probability level of interest, and draw a reference line on the K-factor plot along with the age-regressed mean. A default value corresponding to a threshold probability of noncompliance of .005 will be used in the absence of any other choice. The .995 percentile of a standard normal distribution is 2.576; this is the default reference line.

The regression line for R'_{PTU} is the same as that shown in Figure B-5. The confidence bounds for R'_{PTU} are extrapolated to show their intersection with the reference line. The resulting interval on component age provides a confidence interval for the component age at which an alarm is to be raised.

Figure B-6 illustrates the kind of changes generated by the actions in step 6.

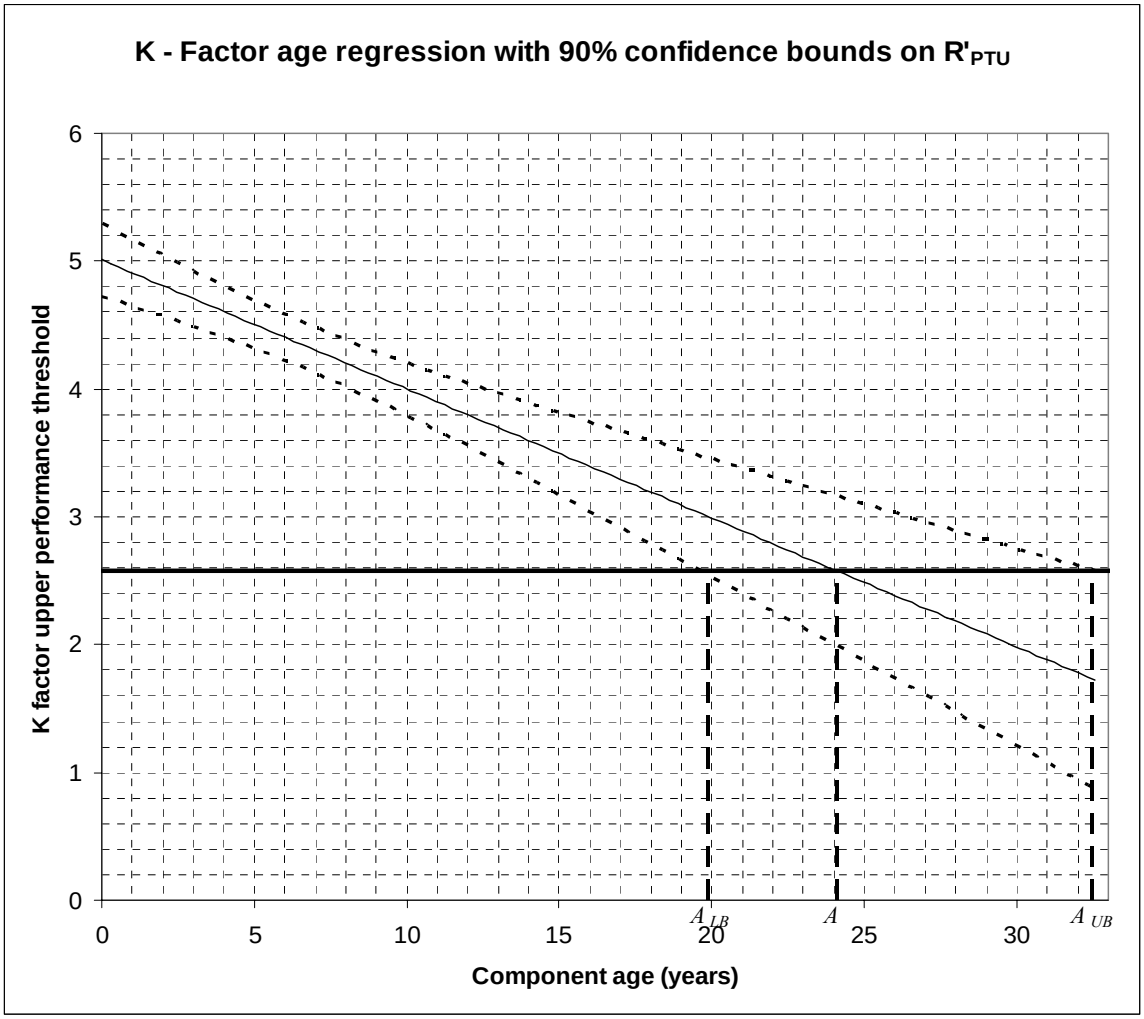


Figure B-6. Illustration of the recommended K-factor upper performance threshold versus component age plot.

The mean line in Figure B-6 can be considered as the expected change in the K-factor as a function of age. The predicted alarm age, $\hat{A}$, is calculated by determining the age at which the mean trend line crosses the reference line. The ages at which the confidence bounds cross the reference line provide confidence bounds for the predicted alarm age. These confidence bounds are denoted by the interval $\left(A_{LB}, A_{UB}\right)$. For the example, $A = 24.1$ years, and the confidence bounds are $\left(A_{LB}, A_{UB}\right)$ = (19.6 years, 32.6 years) for the trend toward the upper performance threshold.

It is possible that the upper bound, A_{UB}, does not exist, as the upper confidence curve may never reach the reference line. This would be the case only if the same level of confidence applied to the slope parameter *d* results in an upper bound that is positive. Or, in other words, the confidence interval for the slope includes the possibility that there is no trend, $d = 0$. It is also possible that none of the three curves (mean, lower bound, upper bound) intersect the reference line, as is the case for the plot produced for the data with respect to the lower limit in Figure B-7.

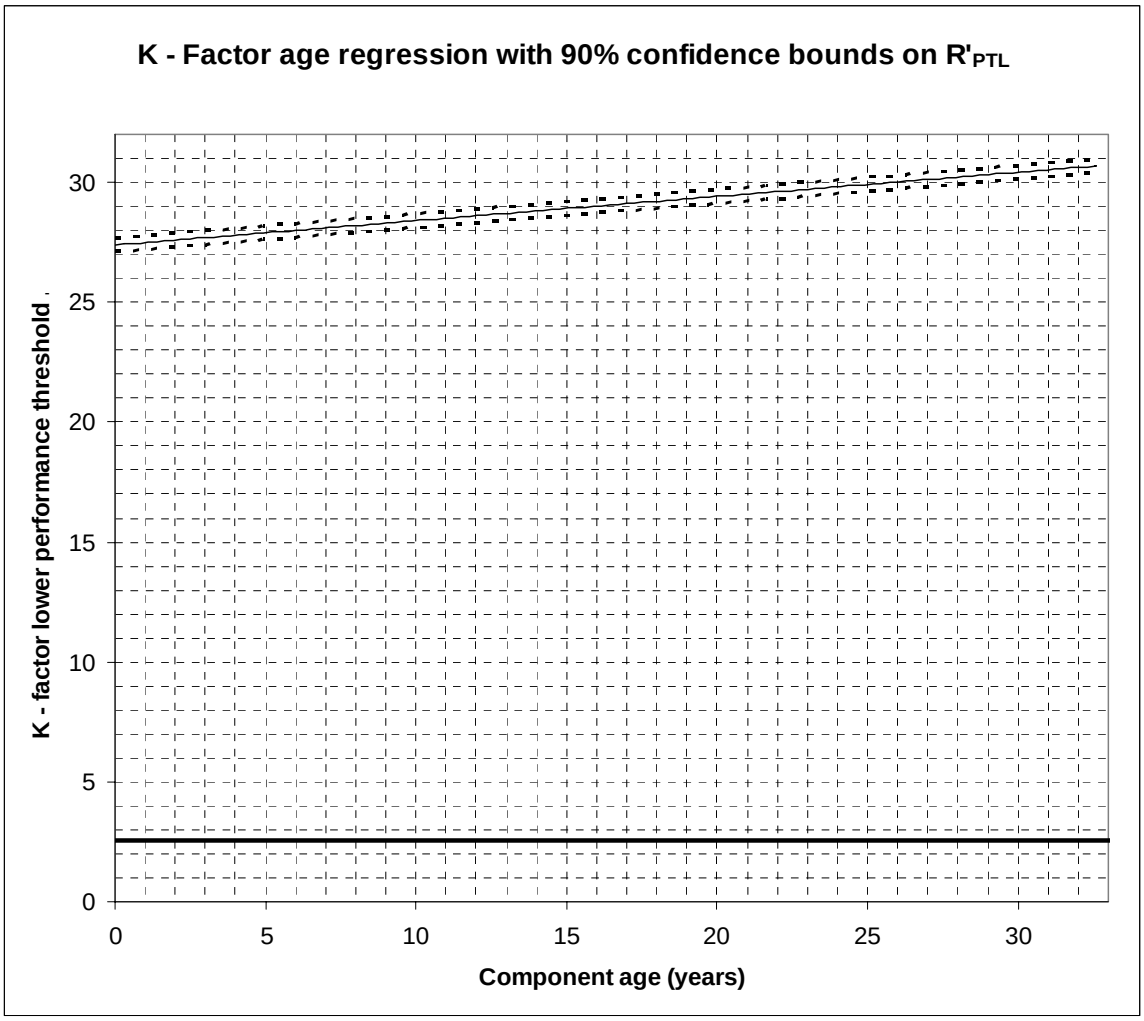


Figure B-7. Illustration of the recommended K-factor lower performance threshold versus component age plot.

Necessarily, for the lower performance threshold, the trend for the mean and both confidence bounds is positive, and therefore none of the three summary values exist.

7. Finally, the analyst should construct diagnostic plots of the residuals from the mean fit of step 4. The purpose of these diagnostic plots is to help determine whether the underlying assumptions of the age analysis are met. The residuals are the differences

of each value and the mean line for the age of the unit; that is, $res_i = M_{PTU,i} - \bar{M}(x_i)$, where x_i is the age of the unit.

Two standard plots are shown in Figures B-8 (Normal probability plot of residuals) and B-9 (Residuals versus age of unit).

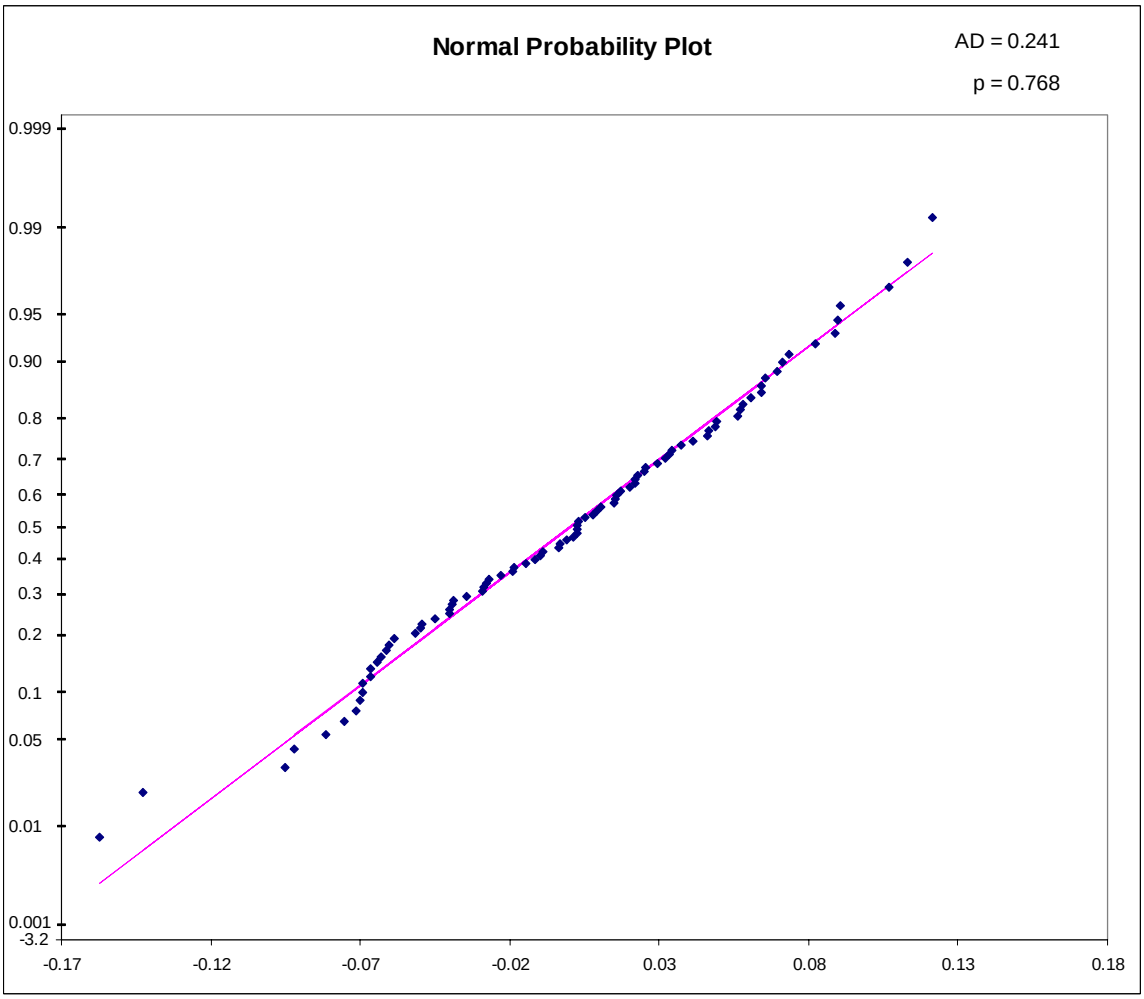


Figure B-8. Normal probability plot of residuals.

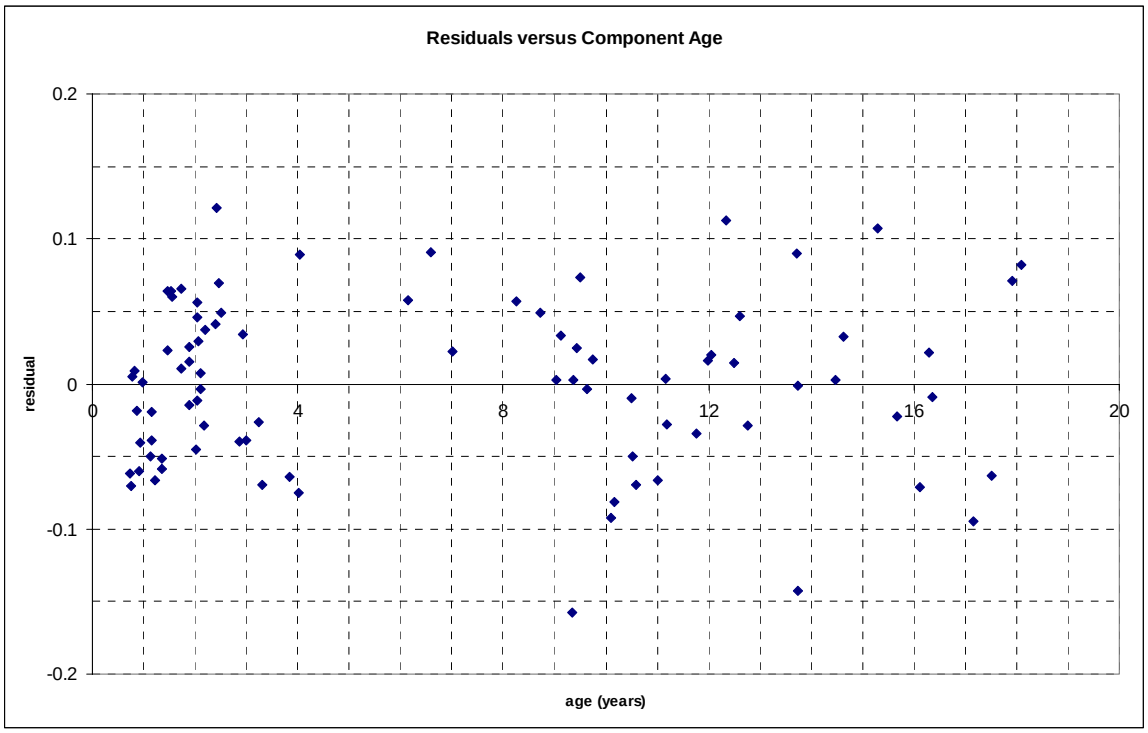


Figure B-9. Residuals versus age of unit plot.

In the normal probability plot in Figure B-8, the data are ordered from smallest to largest, the *y*-position for the *i*th data point is a transformation based on the order, and the *x*-position is the observed value. The *y*-position is appropriately scaled to yield a straight line had the data come from a normal distribution.

Accompanying the normal probability plot is a statistic known as the Anderson-Darling (*AD*) statistic, which is a weighted average of the total deviation of the empirical cumulative distribution function from the nominal straight line that a normal distribution would follow. An associated *p*-value for the statistic indicates the probability that an Anderson-Darling value would be as high or higher than the observed Anderson-Darling value had the numbers been truly drawn from a normal distribution. A low *p*-value indicates how rare the observed deviations from normal are; if the observed deviations are very rare, we conclude that the data are not likely to be normally distributed. In the above case, $AD = 0.241$ with $p = .768$, and we conclude that the data are adequately modeled by a normal distribution.

As a nominal guide for assessing the adequacy of the normal distribution assumption, it is suggested that when $p < .05$, more in-depth modeling of the response should be done or atypical values contained within the physical data should be addressed. If $p < .01$, then some re-analysis should be pursued. (NOTE: The required action may be as simple as removing (by justifiable explanation) a limited number of outliers, modeling the physical data to incorporate additional factors such as test equipment or

test protocol changes, or modeling the data as a mixture of several normal populations. The required action may also suggest that transformations of the data response are needed. Consulting a statistician is advisable if these tests for normalcy fail.)

What information is expected to be gleaned from the residual versus age plot? There are two major areas of concern for which this plot can be helpful in assessing: (1) a lack of common variation throughout the predicted range and (2) a response versus age structure that goes beyond what is being captured by the linear model. Two simple tests are recommended for finding gross departures along these lines. The first test is for equal variation in the first half of the data as compared to the second half of the data. This is accomplished by the usual F-ratio test for equal variances.

In the example, half the data are for those ages prior to four years for which the variance is calculated to be 0.00241. For the 43 points with ages exceeding four years, the variance is 0.00417. The F test statistic is the maximum over the minimum and is 1.737 = (0.00417/0.00241). This value is compared to the F distribution with 42 degrees of freedom in the numerator and 42 degrees of freedom in the denominator, yielding a p-value of .077.

The second test on the residuals is to compare the mean for the middle half of the residuals to the mean of the rest of the residuals (which roughly will be the first and last quarter of the data). The usual two-sample t test for differences in means is used to judge differences. In our example, the middle half of the 86 points is from the 22nd through the 64th points, and the mean is 0.00195. The mean of the remaining points is –0.00195. (Note: These values will nominally be the negatives of each other because the overall average across all residuals is 0. Slight differences may occur when the data contain an odd number of points.) The weighted average of the variance estimates for each group results in an overall variance measurement of 0.00329. The t statistic from these data is 0.315, which is compared to a t distribution with 84 degrees of freedom, resulting in $p = .75$. Thus, there is not evidence that nonlinearity of the relationship is of great concern.

There are many ways in which the plot of residuals can indicate problems with the assumptions of the analysis. The two tests discussed here are meant to provide a statistical test for only the most basic manifestation of problems. The residuals plotted against age are expected to behave like white noise, thus one should look for possible autocorrelation or other evidence of lack of independence of the responses. Autocorrelation is evident when there is a trend in the residuals; for example, they might follow a sinusoidal curve when plotted in observation order. The presence of autocorrelation is an indication of another effect in the responses that has not yet been explained. It may be necessary to review the metadata for the testing to explain the unknown effect. Concerns should be taken to a statistician who can use additional techniques to assess the extent of noncompliance.

Besides the plots from the above steps, the summary values in Table B-1 should be presented. The values are shown here for the example problem, for which it was

determined that only the upper limit was critical. For a given response, it may be the case that one or the other or even both of the quantities should be carried through the above steps. For completeness, the estimates for both the upper and the lower limits are shown.

Table B-1. Summary Values for Example Problem

Quantity	Value		
n (# data pts)	86		
$\overline{R}$	10.012		
σ_m	0.009		
s_R	0.0655		
s_P	0.0574		
PTU	10.26	PTL	8.4
$\hat{c}_{PTU}$	0.2875	$\hat{c}_{PTL}$	1.572
$\hat{d}_{PTU}$	–0.00579 (units/year)	$\hat{d}_{PTL}$	0.00579 (units/year)
Ref = 2.576 (0.005 margin loss)			
$\hat{A}_U$	24.1 years	$\hat{A}_L$	∞
$A_{U,LB}, A_{U,UB}$	19.6, 32.6 years	$A_{L,LB}, A_{L,UB}$	∞,∞

Only the upper alarm-age estimate or the lower alarm-age estimate can exist. The other estimate would necessarily not exist (conveniently noted as ∞), as the physical data cannot be trending simultaneously towards both limits. It is possible, however, that both limits could have a lower-bound estimate. This would occur if the confidence interval for the rate of change included both positive values and negative values. In this case, there would be a lower bound on the alarm age for each limit, but the upper bound would not exist for either of the limits. The presence of lower bounds and the absence of upper bounds for both performance thresholds indicate there are not enough data to be conclusive about the existence of a trend toward either of the performance threshold limits.

B.4 Summary of Analysis and Plots

The analysis presented in this appendix has focused on plotting physical response data and on estimating trends as a function of component age. The first two plots (Figures B-1 and B-2) presented the response data against production date and test date. Either or both of these factors could influence the data in such a way as to be confused with trends as a function of age. It is thus important to look for patterns in these data that can be explained by known production and/or test equipment and test protocol changes.

The physical data may also contain outliers, which are often considered to be points that should be removed from the data before analysis. The outliers, however, can be the important points in understanding overall trends by age, e.g., a small proportion of the

population showing aging characteristics when the bulk of the population do not. With sample sizes being low, the one or two responses pulled from the aging population may look to be outliers. This is why it is important to have an explanation to justify the removal of points from the analysis. More important than the removal of such points is the accommodation of them by identifying an appropriate statistical model for the analysis.

The segregation of a population into two parts, one of which is showing aging trends and one of which is not, would manifest itself in growing variation and/or lack of normality in the residuals. This is why it is important to look at the residuals after fitting the best linear trends.

The identification of an alarm age, the age for which units in the population will begin to fail at some predetermined rate, is based on normality assumptions concerning the distribution of the responses about a mean line. It may be the case that a transform of the data response will better fit this assumption. Statisticians can help with determining whether and how this should be done.

The analysis presented in this appendix was based on response measurements taken on units for which an age could be determined at the time of test. If, in addition, responses are known not only for the aged unit but also at the time of production, then the analysis should be done for within-unit changes. This is often referred to as "paired data," and the response that should be trended is the difference of the aged response and the production response. Considering within-unit changes thus removes some of the variation and gives a better trend estimate.

Appendix C – Guidelines for Application of QMU Methodologies to M&S-Centric Evaluations

C.1 Introduction

The purpose of this appendix is to discuss best practices and provide guidelines for the application of quantification of margins and uncertainties (QMU) methods when modeling and simulation (M&S) is centric to QMU assessments. M&S typically plays a central role when sufficient numbers of tests are not possible because of cost, environmental hazards, social or political considerations, time scales associated with the issues, or when it is physically impossible to test. QMU is applied in a decision-making context, addressing the ability to meet design, qualification, or life-cycle performance requirements. Because of the programmatic constraints of cost and schedule, there are often significant uncertainties due to lack of knowledge associated with the use of M&S for many of our high-consequence issues.

QMU, particularly for M&S-centric applications where uncertainties are dominated by lack of knowledge issues, has the technical dimensions of quantitative risk assessment (QRA). From the perspective of QRA, risk can be defined in terms of the Kaplan and Garrick (1981) risk triplet:

1. *Scenario identification* – What can happen?

2. *Likelihood of scenarios* – How likely is it to happen?

3. *Consequence of scenarios* – What are the consequences if it does happen?

A fourth component has always been an important factor in the use of QRA and will be an important factor in the application of QMU to the stockpile:

4. *Credibility* – How much confidence do you have in the answers to the first three questions?

The guidance for our QMU framework must be formulated in a manner that can easily address these four questions.

C.2 Requirements Language

Performance or safety requirements establish the metrics by which "consequence" can be measured in the context of a particular application. Consequently, performance metrics and requirements are the key elements in answering the third risk question. In the main body of this paper, specification of the performance requirements was described as deterministic:

$$R > R_T \,,$$

where R is the performance response and R_T is the performance threshold; or specification of the performance requirements was described as probabilistic:

$$\text{Prob}\left(R > R_T\right) < P_T ,$$

where R is the performance response, R_T is the performance threshold, and P_T is the specified probability threshold that must be exceeded for the requirement to be met. As an example of a deterministic requirement, fire-set voltage is required to exceed a minimum value to ensure that the detonators will fire. Nuclear safety, on the other hand, requires that the probability of inadvertent nuclear detonation in an accident be less that 10^{-6} for any initiating event, which is an example of a probabilistic requirement. In some cases, Sandia's requirements are expressed in qualitative terms. In hostile environments, for instance, successful performance requires that there be no significant degradation in the system reliability. Further, such a determination of successful performance requires a quantitative interpretation of the meaning of an associated QMU assessment is desired.

(Note that the quantitative expressions of performance threshold requirements have to be modified appropriately based on whether the requirements are stated quantitatively as upper bounds, lower bounds, or bounded intervals).

The expectation is that decisions will be made with high confidence, although the exact degree of required confidence must be tailored to the needs of a specific application. However, not all high-consequence issues have requirements that explicitly state a confidence that includes uncertainties. Some high-consequence issues of national importance (e.g., nuclear power safety [Helton and Breeding 1993]; geologic disposal of transuranic radioactive wastes, WIPP [Helton et al. 1999]; and geologic disposal of high-level radioactive wastes, Yucca Mountain [Helton and Sallaberry 2007]) involve decisions that are tied to some measure of central tendency (mean or median) and are simply "informed" by the required treatment of uncertainties.

C.3 Figures of Merit

The figure of merit recommended is the confidence factor, *CF*, which is defined as

$$CF = \frac{M}{U} ,$$

where M is defined as the margin and U is defined as the *uncertainty*. The confidence factor CF is methodologically rigorous only when U reflects epistemic uncertainties that are represented with a sharp interval. In practice, U is often used to capture both aleatory and epistemic uncertainties. To ensure consistency across applications, it is recommended that the margin M be defined in terms of the difference of median values for assessed and threshold distributions and that the uncertainty U be defined in a manner to convey "high confidence" in the context of a specific application. If the assessment of U is rigorous, then it is sufficient that $CF > 1$ to ensure that the reliability is "ONE" with high confidence. In practice, however, it is prudent to demand some robustness to unknown

unknowns or to assessments lacking rigor in the modeling processes. Consequently, it is likely that some issues will require additional attention if *CF* is too close to unity.

The QMU policy of the National Nuclear Security Administration (NNSA) also allows for a related figure of merit termed the ***K-factor***, *K*, which is defined as

$$K = \frac{M}{S},$$

where *M* is the margin and *S* is the standard deviation. The K-factor, *K*, is most appropriate when aleatory uncertainties dominate, a condition that typically occurs in physical data–rich QMU applications (e.g., see Appendices A and B), but that can sometimes occur in M&S-centric QMU applications when all the inputs have aleatory uncertainties and model errors are negligible, so that any distribution of outputs has a frequency interpretation. The K-factor is also known as the "reliability index" in the reliability literature. One component of reliability, margin sufficiency, can be directly expressed in terms of the K-factor: ***reliability*** = $\Phi(K)$, where Φ is defined as the commonly tabulated standard normal cumulative distribution function. This relation is strictly true, however, only under the assumption of Gaussian distributions (or distributions transformable to Gaussian, such as the lognormal distribution) with margins defined as the difference in means of the underlying distributions of the observed and threshold performance variable. The assessed reliability is thus 0.84 when $K = 1$, which in general does not uniquely assure compliance with the requirements. However, the K-factor does provide a measure of relative importance across applications.

The aggregation of a QMU assessment to a single figure of merit is unavoidable, but it comes at the expense of losing valuable information that could help inform decisions. For instance, the confidence factor *CF* could have an assessed value of 5, which is based on a large margin ($M = 5$) and a large uncertainty ($U = 1$). On another hand, the same confidence factor could be based on a very small margin ($M = 0.05$) and very small uncertainty ($U = 0.001$), which demands more carefully scrutiny of the evidence. Good practice, therefore, is to communicate the QMU figure of merit *and* the values from which it is derived, including all distributions to which *M* and *U* are referenced.

C.4 Graded Approach to QMU

There is a need to provide a graded approach to M&S-centric QMU activities at Sandia. This need is dictated partly because Sandia has a very large number of design and qualification requirements and, based on their potential importance or a screening of margins and uncertainties, not all of these requirements justify the highest level of rigor. We consider four levels of rigor for M&S-centric uncertainty quantification (UQ) & sensitivity analysis (SA) studies:

- High Rigor: This level of rigor is most appropriate when addressing high-consequence decisions that are made predominantly on the results of M&S and for which it is impossible or impractical to conduct relevant qualification tests.

- Medium/High Rigor: This level of rigor is most appropriate when addressing high-consequence decisions where M&S informs the decision process in some significant way by complementing, extending, or extrapolating the parameter space explored through relevant qualification testing.

- Low/Medium Rigor: This level of rigor is most appropriate when addressing decisions of lower consequence. This is the case where M&S is used in some significant way to design component or qualification tests. In these cases, it is practical to confirm decisions through relevant testing.

- Low Rigor: This level of rigor is most appropriate when addressing low-consequence decisions. Examples include scoping studies or exploratory research where important NW stockpile decisions are not being made.

Increasing rigor comes at the expense of cost and schedule, and the specialized skill sets, technologies, or infrastructure may not always be available to execute a QMU study at the highest levels of rigor. The graded approach can be used to inform a decision maker of the increased potential of being misinformed as a consequence of accepting a QMU analysis having less rigor than is desirable for the application. The graded approach can also be used as a tool to frame the negotiation of expectations (cost, schedule, and performance) at the outset of a QMU study.

Table C-1 summarizes some key attributes associated with different levels of rigor in an M&S-based QMU study.

Table C-1. Graded Approach to Uncertainty Quantification and Sensitivity Analysis

Levels of Rigor	Attributes
High Rigor (Maturity Level 3): High-consequence M&S-based decisions, no confirmatory testing available, e.g., qualification	• Aleatory and/or epistemic uncertainties represented separately and interpreted in an uncertainty-preserving manner • Rigorous quantification of the sensitivity of output uncertainties to input uncertainties • Numerical (propagation) errors rigorously quantified • No strong assumptions
Medium/High Rigor (Maturity Level 2): High-consequence M&S-informed decisions, limited confirmatory testing available, e.g., qualification support	• Aleatory and/or epistemic uncertainties represented separately • Sensitivity of output uncertainties to input uncertainties estimated quantitatively • Sensitivity to numerical (propagation) errors explored • Some strong assumptions
Low/Medium Rigor (Maturity Level 1):	• Aleatory and/or epistemic uncertainties represented and propagated without distinction • Sensitivity outputs to input uncertainties

Levels of Rigor	Attributes
Lower-consequence M&S-informed decisions, extensive confirmatory testing available, e.g., design support	explored through qualitative "what if" studies • Many strong assumptions
Low Rigor (Maturity Level 0): Low-consequence M&S-informed decisions, confirmatory testing not necessary, e.g., scoping studies	• Judgment only, uncertainties not addressed • Judgment only, or SAs not addressed • Judgment only, numerical (propagation) errors not addressed • Judgment only, strong assumptions not addressed

Note that each level of rigor in Table C-1 is identified at a given maturity value, ranging from 0 through 3. The maturity levels are used as part of the assessment approach discussed next.

Credibility of the Modeling Process

It is a fallacy to believe that analysts using generally available codes (more emphatically, research codes) with their underlying models can produce unquestionably adequate results for all intended applications. It is good practice to understand and communicate the credibility of models used in a QMU study in a concise manner, and to be prepared to defend challenges with additional evidence supporting the credibility of the QMU results. The intent then is that M&S-centric QMU assessments include the confidence factor as a summary measure of the application results, the relevant distributions or values for *M* and *U* individually, and a summary assessment of the maturity of the modeling process used to generate the QMU results. The Predictive Capability Maturity Model (PCMM) can be used as a guideline in assessing and communicating the predictive capability of models used in a specific application. The PCMM (see Table C-2) expands the risk-graded approach presented in Table C-1 for UQ/SA by addressing five elements that also contribute to the understanding of predictive capability. A brief description of the five elements follows, but a more complete discussion can be found in Oberkampf et al. (2007).

1. *Representation (geometric) fidelity*: This element addresses the question, Are important features neglected because of simplifications or stylizations that could degrade QMU results? For instance, some instabilities are inherently three dimensional in nature and might not be represented at all in one- or two-dimensional models. As another example, failures might not be adequately modeled if small critical regions where they initiate are not adequately represented in the model. The focus here is on minimizing or characterizing bias errors introduced when representation or geometric fidelity of the model is not adequate.

2. *Physics and material model fidelity*: This element addresses the questions, How fundamental are the physics and material models? and What is the level of model calibration? The focus here is on understanding the controlling physics in the application and the risk involved in using models to extrapolate outside areas where models are anchored in physical data.

3. *Code verification*: This element addresses the question, Are software errors and algorithm deficiencies corrupting the simulation results? The goal is to understand the degree to which due diligence has been applied to ensure that software features and capabilities (F&Cs) used in the specific application are free of coding errors or faulty numerical algorithms that do not perform correctly in the application parameter space.

4. *Solution verification*: This element addresses the question, Are numerical errors corrupting simulation results? Numerical errors, which should be judged in relation to the other uncertainties associated with analysis, can be associated with time, space, angle, energy, or other finite discretizations associated with the specific model. In addition, numerical errors can arise from any of the parameters (e.g., artificial viscosity, hourglass stiffness, and so forth) that are associated solely with the control of numerical algorithms.

5. *Validation*: This element addresses the question, How accurate are the simulation results at various tiers in the validation hierarchy? All models are approximations, but some models are more applicable and accurate than others. This element involves the comparison of model predictions to physical data and the characterization of variability and epistemic uncertainties in the use of the model for the specific application, which might involve interpolation or extrapolation.

Table C-2 defines the major elements of the PCCM.

Table C-2. Predictive Capability Maturity Model

MATURITY / ELEMENT	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 boundary conditions (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&Cs) 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 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 exper-imental 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 SAs conducted for parameters and models • Numerical propagation errors are demonstrated to be small • No significant UQ/SA assumptions made • Independent peer review conducted

C.5 Implementation Guidance

There are four key guidelines for implementation:

1. Account for different kinds of uncertainty.
2. Quantify sensitivities of key outputs to uncertainties in inputs.
3. Quantify numerical (propagation) errors.
4. Avoid strong assumptions.

These guidelines are discussed next from the perspective of what is desirable for the highest-consequence applications.

C.5.1 Account for Different Kinds of Uncertainty

Uncertainty can be formally classified as *aleatory uncertainty* (stochastic variability) and *epistemic uncertainty* (incomplete knowledge). Where it is practical, calculation input characterizations should separate aleatory and epistemic uncertainties. Similarly, aleatory and epistemic uncertainties should be represented separately in calculation outputs. Other M&S-based risk-informed decision processes for high-consequence issues of national importance (reactor safety [Helton and Breeding 1993], geologic disposal of radioactive wastes – WIPP [Helton et al. 1999] and Yucca Mountain [Helton and Sallaberry 2007]) have adopted this perspective.

Aleatory uncertainty characterizes the inherent randomness in the behavior of the system under study. Alternative terminologies include variability, stochastic uncertainty, irreducible uncertainty, and Type A uncertainty. Aleatory uncertainty can only be reduced by modifying the design and/or production of a component or material. Examples of aleatory uncertainty are component failures or material properties that are derived from statistical testing under conditions relevant to the application. Aleatory uncertainties are characterized by frequency distributions, and aleatory uncertainties propagated through a model with negligible error will result in distributions for key performance responses that should also carry a frequency interpretation. The second risk question in the introduction to this appendix (Section C.1) is associated with aleatory uncertainties. Consequently, the first three risk questions form the basis of a traditional reliability analysis and lead to assessment of the frequency of meeting a specified requirement.

Epistemic uncertainty characterizes the lack of knowledge about the appropriate value to use for a quantity that is assumed to have a fixed value in the context of a specific application. Epistemic uncertainty can be associated with scenarios, parameters in models, alternate plausible models (both physical and statistical), and so forth. Alternative terminologies include state-of-knowledge uncertainty, incertitude, subjective uncertainty, reducible uncertainty, and Type B uncertainty. Epistemic uncertainties are reduced, for example, by increasing one's understanding through more research, or by

increasing the total amount or overall relevance of the experimental and test data pertinent to the problem under study. Epistemic uncertainties are characterized by degrees of "belief" and should not be given a frequency interpretation. The fourth risk question in Section C.1 addresses the epistemic uncertainties in the application. Thus, the fourth question answers the question, How confident are you in your assessment of the reliability?

As discussed in the main body of this paper, "probability" should be used very carefully when one is communicating QMU results. The word "probability" can carry either a frequency interpretation or a belief interpretation. As an example of the confusion that can ensue, consider the following statement: the probability of a device functioning properly is 95%. Consistent with the frequency interpretation, the expectation would be that 95 devices out of 100 (on average within a large population) would function properly. The belief interpretation would convey 95% confidence (i.e., belief) that *all* the devices would function properly, but would reserve 5% belief that *all* the units would not function properly. The strategic implications in a decision context could be quite different based on the intended interpretation of the word "probability". Because of this duality in the meaning of the word "probability", we recommend that the words *frequency* and *belief* be used directly. Alternatively, if the word "probability" is used in communicating QMU results, we recommend that the proper interpretation be defined explicitly for the decision maker.

One should infer from epistemic results only what is justified. Epistemic performance responses can be greatly underestimated if the results are given a probabilistic interpretation. As an example, consider that, for a specific performance assessment, an important performance response is represented as the sum of 10 uncertain variables and that the performance response is required to be less than seven. All that is known about any of the 10 variables is that they are bounded between zero and one and that all values within the interval are possible with no sense of graded belief. There is no evidence to support a probabilistic distribution of the possible values for the variables. In this case, an "uncertainty-preserving" interpretation is that all values between 0 and 10 are possible for the performance response and that violation of the requirement cannot be excluded. Alternatively, one might be tempted to represent each of the uncertain input variables as uniform probability distribution functions (PDFs) that are propagated through the model with Monte Carlo. The result of such a representation is depicted in Figure C-1. When a decision maker is presented with the analysis in this figure, a conclusion could easily be reached that there is 98.6% confidence (a belief-based interpretation that is sometimes confused with a frequency-based interpretation) that the requirement is not violated. Such a representation of epistemic uncertainties is not "uncertainty-preserving" if the results are given a probabilistic interpretation. The result implies more than the inputs justify (i.e., higher confidence than supported by the knowledge). Some decision makers have indicated that they do not have confidence in risk studies because of the tendency to make "something from nothing" by implying more confidence than can be justified by the supporting evidence. The temptation to assume distributions and make confidence statements that cannot be supported should be avoided.

In this simple example, it is more logically consistent to say that any computed result should be interpreted as being possible without any sense of frequency or graded belief and to then question whether input combinations leading to maximum results have been identified. This is an "interval representation of results" that is consistent with the available input information. Advanced techniques for the representation of epistemic results in an "uncertainty-preserving" format (including treatment of cases when aleatory uncertainties are combined with epistemic uncertainties) include probability bounding (Ferson and Tucker 2006), Dempster-Schaefer representations (Helton et al. 2004), nested polynomial chaos expansions (Red Horse and Benjamin 2004), and belief scales (Pilch 2005). All these methods give the same interval-like results discussed above when epistemic uncertainties dominate and when all epistemic inputs are intervals. The methods differ slightly in their presentation of results when epistemic inputs are given characteristics of graded belief. None of these methods have yet been applied to high-consequence applications in the United States, nor have they been subjected to significant peer review at the national or international level; and as such, they are currently regarded as research topics.

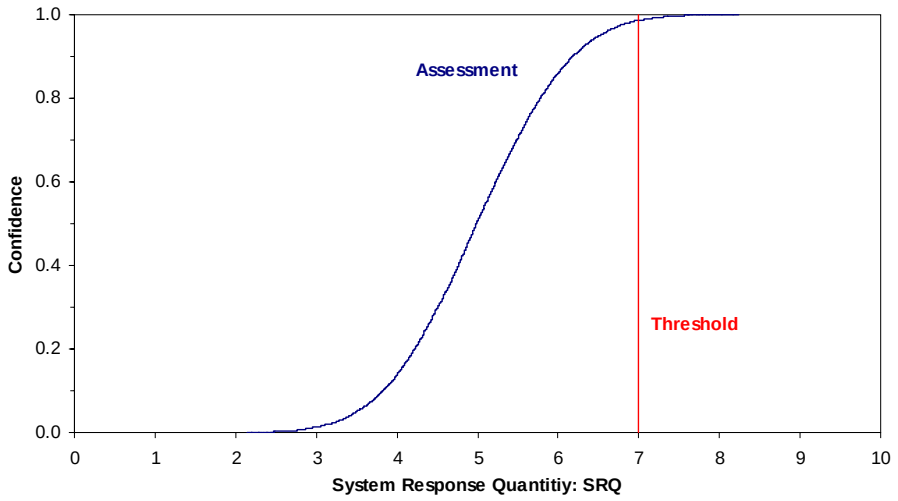


Figure C-1. Probabilistic representation of epistemic results underrepresents true uncertainty.

C.5.2 Quantify Sensitivities

QMU increases our understanding of the technical basis for decisions, in part because a quantitative sensitivity analysis (SA) can help identify the dominant input uncertainties that contribute to the uncertainties in the predicted performance response. Increased understanding allows resources and attention to be focused on those elements that have the greatest impact on the results. Most commonly available tools for SA are based on assumptions of linearity. Consequently, the analyst should justify assumptions of linearity and be cautious of highly nonlinear sensitivities, threshold behavior, or resonance behavior, all of which can lead to misleading perceptions of sensitivity under inappropriate assumptions of linearity. Spurious sensitivity results can also appear when there are too few data to reliably quantify sensitivity or when correlation amongst the inputs is present.

C.5.3 Quantify Numerical Errors in UQ/SA Results

Computed UQ/SA results (distributions and key statistical summaries such as percentiles or moments, margins, QMU figures of merit, and sensitivity metrics) are no more than estimates derived from finite numbers of model evaluations. One simple approach to assess sensitivity to numerical errors is to repeat a sampling-based study with a different ensemble of calculations derived from a different random number seed. In a similar spirit, hierarchical sampling strategies (Helton et al. 2006) can be used to assess sensitivity to numerical errors by adding additional samples until results are "converged." The rigor of such straightforward sampling-based studies is typically limited because of the prohibitive computational expense associated with many M&S-based QMU studies. For uncertainties addressed with probabilistic methodologies, the standard errors associated with statistical attributes (e.g., mean, variance, correlation coefficient) can be applied to computed confidence bounds. For some statistical attributes of interest (e.g., means or variance), analytic expressions have been developed for the standard errors. Furthermore, bootstrap methods (Davidson and Hinkley 1999) and can be used to estimate the standard errors for any general statistic of interest (often with no new function evaluations).

For other nonprobabilistic methods, the concept of numerical errors needs to be addressed using language specific to the chosen methodology. For instance, epistemic uncertainties characterized as intervals with all intermediate values treated as possible will produce an ensemble of results when propagated through a model that is evaluated a finite number of times. The ensemble coarsely defines the interval of possible outputs, but that interval is only an estimate because of the finite number of calculations. Consequently, it is unlikely that the produced output range contains the exact combinations of inputs that maximize the interval of output results.

Polynomial chaos is another class of UQ methodologies, where stochastic outputs are represented by a truncated series of stochastic functions and where the series coefficients are evaluated based on the available data (physical or code-generated). The properly posed questions are, What are the standard errors in key statistics associated with finite numbers of "data"? and "What is a truncated series expansion of the statistical outputs?

Errors in UQ/SA results can also arise if surrogate models (i.e., response surfaces) are first fit to a training set of computations and then sampling is performed on the response surface in lieu of the more detailed, but computationally intensive, physics model. Although numerical errors are a form of epistemic uncertainty, we recommend that they be addressed and represented separately when possible.

C.5.4 Avoid Making Strong Assumptions

Strong assumptions are unsubstantiated assertions or unrecognized assumptions that have the potential of influencing the results of the study in some significant way. There are many ways that strong assumptions can creep into a QMU/SA study and undermine the credibility of the results. The following practices can be used as a guide to avoid some strong assumption.

Provide justification for assumptions of linearity. QMU has its greatest value when applied to high-consequence, and sometimes controversial, issues where nonlinear coupled multiphysics models (sometimes exhibiting resonance or threshold behavior) are being applied in complex geometries. In the absence of evidence, it is dangerous to employ UQ/SA methodologies that assume a linear relationship of output uncertainties with input uncertainties. Here are two examples. The variance of a random output *y* is commonly given as $\sigma_y^2 = \sum \left(\frac{\partial y}{\partial x_i} \sigma_{x_i} \right)^2$ (where $\sigma_{x_i}^2$ is the variance of x_i), which assumes linearity of the response over each range of uncertainty. If this variance relation is used, then evidence should be provided that *y* is a linear function of each dominant x_i over the σ_{x_i} range. As another example, the square of the partial correlation coefficient $r^2\left(y, x_i\right)$, is commonly used as a measure of sensitivity of the output *y* to the uncertainty in input x_i. This relationship rests on an assumption of linearity between *y* and x_i. Scatterplots of *y* as a function of x_i could be used to justify the assumed linearity.

Provide supporting evidence for the characterization of dominant uncertainties. The potential value of QMU or risk studies is discounted in the minds of some decision makers because of the perception, in the absence of supporting evidence beyond expert judgment, that uncertainty propagation provides an illusion of making something from nothing. Although it may be useful for initial scoping studies to arbitrarily assume generous ranges on some uncertain inputs, it is best practice to provide traceable supporting evidence for the characterization of dominant uncertainties and to explicitly discuss the aleatory or epistemic nature of the uncertainty characterization. It is well documented that certain cognitive biases commonly lead to underrepresentation of epistemic uncertainties (Tversky and Kahnerman 1974). When characterizing epistemic uncertainties, the focus should be on understanding the limits of credibility, although it is acceptable to provide measures of graded belief within those limits. For epistemic issues, it is common that supporting evidence is in conflict, and any schemes to aggregate conflicting information should preserve the full range of possibilities (e.g., vertical averaging of cumulative distributions).

Provide justification for assumptions about the functional form of distributions. When possible, justify selection of the distribution form based on theory, comparison to experimental data, or sound rationale. In the case of experimental data, statistical goodness-of-fit measures should be provided. In the absence of supporting evidence or compelling argument, it is sometimes convenient to assume, for instance, that a distribution is normal or lognormal. This is bad practice. Uncertainty propagation methods rely implicitly on these distributional forms and their associated parameters; consequently, faulty assumptions can lead to errors that cannot be reliably assessed. In particular, tail behavior is sensitive to assumed distribution forms. Nonparametric and distribution-free techniques should be considered to avoid making assumptions about the distribution form. Alternately, one can treat the distribution form as epistemic by exploring sensitivity to different assumed distributions or by allowing the parameters describing a given distribution to be distributed themselves in an epistemic manner.

Characterize the dependencies among all uncertain parameters. Probabilistic models used in probabilistic risk assessments take two kinds of inputs: (1) the marginal distributions for the different variables and (2) the dependencies between these variables. The second set of inputs is arguably as important as the first, but they are commonly ignored. Variables cannot be assumed independent without theoretical or observational justification. Variables also cannot be assumed to be linearly correlated without reasonable justification. The rationale that certain combinations of parameters are not possible can be used to simplify the dependency analysis. Although it is often reasonable to assume independence among some variables in an engineered system, it is not always reasonable to do so. For instance, it is probably not tenable to assume independence between component masses and surface areas, or between age and performance. Linear correlation is not the only form of stochastic dependence (which is the reason that lack of linear correlation does not guarantee independence), and pair-wise independence does not imply mutual independence in the general multivariate case. Common-cause or common-mode failures can introduce unrecognized dependencies among the variables in the analysis. Such subtle dependencies may become more likely in abnormal operating environments.

In the past, risk analyses sometimes consciously or unconsciously assumed (or left it to the reader to assume) the uncertainty inputs in the mathematical expressions were independent even when the justification for this assumption was quite weak or nonexistent. In particular, the lack of specific empirical evidence to the contrary might be proffered as a reason to assume independence. This is bad practice because it can lead to substantial over- or underestimation of the output results and because, in either case, such output results are divorced from physical reality. For example, if the variables added or multiplied together in a probabilistic risk assessment have positive correlations that are ignored or incorrectly assumed to be zero, or if variables with negative correlation are subtracted or divided, the resulting probability distribution will likely be underestimated in the tails. The import of such an underestimate is that the analysis will yield an erroneous and potentially dangerous miscalculation of the chance of high-consequence events. Without specific analysis, there is no way to foretell whether inattention to such phenomena causes wasteful overestimation or dangerous underestimation of the actual ability to meet performance requirements.

Ensure consistency between the variability of model parameters and the underlying assumptions of the models. Avoid confusion over temporal vs. spatial, local vs. global variability by stating explicitly the underlying assumptions. For instance, to be mathematically consistent with the formulation of many conservation equations, the associated material parameters should be interpreted as spatial averages (or temporal averages, depending on the application). However, it is possible to characterize material parameters in laboratory tests using samples too small to be representative of the "true" average material (this problem is more likely for complex materials or for complex material parameters). Similarly, the sample population for material parameter tests should be relevant to that of the intended application. For example, the variability in a material parameter for a population of devices consists of the natural variability of the material, as well as the batch-to-batch material variability from a given vendor, and possibly vendor-to-vendor material variability for the application population. Characterizing the

variability from one unit selected from the application population (or worse yet, from an irrelevant population) could well underrepresent the variability exhibited by the target population.

Address alternate plausible models. Sometimes there are alternate plausible models that give substantially different results when applied in a parameter space far from where the models might be anchored in data. In such cases, it is not sufficient only to address parameter uncertainty within a selected model because the greater uncertainty may arise from the model form itself. In the absence of compelling evidence, it is bad practice to arbitrarily select a single model or to average together incompatible models or use Bayesian model averaging. It is recommended that the results from all plausible models be compared in the analysis.

Seek best estimate plus uncertainty. In the face of large epistemic uncertainty, it is sometimes tempting to assume a conservative model or parameter value in lieu of the effort and/or cost required to sufficiently understand the credible spectrum of models or parameters. This is bad practice. The inclusion of conservative assumptions with best estimates (plus uncertainty) introduces a set of conditions that undermines the ability to ascribe either a frequency or a belief interpretation to the results. This practice also undermines the value of an SA to identify contributors to uncertain results. More importantly, however, it is bad practice to assume that the extremes of the predicted performance response are always derived from the extremes of uncertain inputs. Monotonic behavior of predicted performance responses to uncertain inputs in general is not known in nonlinear complex multiphysics applications (e.g., the maximum predicted performance response may occur for the minimum of some inputs combined with the maximum of other uncertainties), and in some cases resonance behavior could maximize predicted performance responses for intermediate values of uncertain inputs. This inability to correlate output extremes with input extremes is a problem for Sandia because the scenarios specified in the STS requirements are generally judged to be conservative. The recommended practice then is to allow for no other "conservative" element in the analysis and clearly state that the QMU results are conditional on the specified scenario or environment.

C.6 References

1. Davidson, A. C. and D. V. Hinkley. (1999). *Bootstrap Methods and Their Application*. Cambridge, UK: Cambridge University Press.

2. Ferson, S., and W. T. Tucker. (2006). "Sensitivity Analysis Using Probability Bounding." *Reliability Engineering and System Safety* 91, nos. 10–11: 1435–1442.

3. Helton, J. C., and R. J. Breeding. (1993). "Calculation of Reactor Accident Safety Goals." *Reliability Engineering and System Safety* 39: 129–158.

4. Helton, J. C., D. R. Anderson, H. N. Jow, M. G. Marietta, and G. Basabilvazo. (1999). "Performance Assessment in Support of the 1996 Compliance Certification Application for the Waste Isolation Pilot Plant." *Risk Analysis* 19, no. 5: 959–986.

5. Helton, J. C., J. D. Johnson, and W. L. Oberkampf. (2004). "An Exploration of Alternative Approaches to the Representation of Uncertainty in Model Predictions." *Reliability Engineering and System Safety* 85, nos. 1–3: 39–71.

6. Helton, J. C., and C. J. Sallaberry. (2007). *Illustration 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*. SAND2007-1353. Albuquerque, NM: Sandia National Laboratories.

7. Helton, J. C., J. D. Johnson, C. J. Sallaberry, and C. B. Storlie. (2006). "Survey of Sampling-Based Methods for Uncertainty and Sensitivity Analysis." *Reliability Engineering and System Safety* 91: 1175–1209.

8. Kaplan, S., and B. J. Garrick. (1981). "On the Quantitative Definition of Risk." *Risk Analysis* 1, no. 1: 11–27.

9. Oberkampf, W. L., M. Pilch, and T. G. Trucano. (2007). *Predictive Capability Maturity Model for Computational Modeling and Simulation*. SAND2007-5948. Albuquerque, NM: Sandia National Laboratories.

10. Pilch, M. (2005). *The Method of Belief Scales as a Means for Dealing with Uncertainty in Tough Regulatory Decisions*. SAND2005-4777. Albuquerque, NM: Sandia National Laboratories.

11. Red-Horse, J. R., and A. S. Benjamin. (2004). "A Probabilistic Approach to Uncertainty Quantification with Limited Information." *Reliability Engineering and System Safety* 85, nos.1–3: 183–190.

12. Tversky, A. and D. Kahneman. (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science* 185: 1124–1131.

Appendix D – Example Application of QMU Methodologies to M&S-Centric Evaluations

The purpose of this appendix is to illustrate many of the concepts and presentation formats for M&S-centric QMU, as discussed in Appendix C, in the context of a simple "synthetic" application. We begin with a description of a sample problem. The treatment of aleatoric and epistemic uncertainties is of particular interest in this example. Following the description of the sample problem are three methods by which the problem can be worked. The appendix concludes with a commentary on the three methods and a revelation of the true values of certain parameters in the model. This revelation emphasizes the importance of proper representation of epistemic uncertainties.

D.1 Sample Problem Description

This synthetic example is characterized by a threshold distribution representing some key performance response (R_T). Here, we specify that the threshold function can be represented as a normal distribution with a precisely known mean $\mu = 8$ and standard deviation $\sigma = 2$; consequently, the threshold distribution is purely aleatory in nature.

$$R_T = N(8,2)$$

This is analogous to many Sandia applications where the threshold function is associated with component or material failure. In these cases, it is common that lot-sample testing of components or material tests can be used to characterize the threshold function. Typically, there are epistemic uncertainties associated with the characterization, partly because the database is limited and partly because of uncertainties in the diagnostics or in the applicability of the database to the application. These epistemic elements of the threshold function (distribution form itself or the parameters describing a given distribution) add complexity, but no new conceptual insight; consequently, the synthetic example is stylized to address only aleatory uncertainties for the threshold function.

In the spirit of M&S, we specify that a computer code is used as a transfer operator between the environment specifications (associated with some scenario, typically supplied in the STS requirements) and the M&S-based assessments of *R*, which will be compared with the threshold function. As is sometimes the case, the computer model can be computationally intensive; so for the synthetic example, we specify that only ~25 evaluations of the model can be anticipated with the available resources. For the example, the "model" is represented by

$$R = f(a,b) = a^b$$

where *f* denotes evaluation by a computer code, *a* is a parameter associated with the scenario-dependent environment specification, and *b* is a model parameter. There are no alternate plausible models, and model form errors are specified to be negligible relative to

other sources of uncertainty or error, which is another stylization made here to more clearly illustrate the key QMU concepts and presentation formats.

The parameters *a* and *b* are thought to have fixed but unknown values for the purpose of this application; consequently, they are purely epistemic in nature, which is typical of many of our applications. We represent the values for *a* and *b* as intervals,

$$a = [1,2] \qquad b = [0,3].$$

The epistemic characterization of *a* and *b* is thus represented by a range of values, and any value within the range is simply deemed possible without any evidence to favor one value over another.

A probabilistic requirement (typically specified or implied in the military characteristics for actual weapon issues) is provided for the synthetic problem:

$$\text{Prob}\left(R > R_T\right) < P_T < 0.01$$

For this application, there is an expectation that the probabilistic requirement must be satisfied with high confidence.

The synthetic example is not representative of all possible M&S applications, but it does embody representative elements of some of Sandia's more important safety and survivability applications, as illustrated in Table D-1.

Table D-1. Relevant Qualification Issues Exemplified by Synthetic Problem

Application	Response: R	Threshold; R_T	Requirement
Abnormal Thermal	Temperature	Stronglink and weaklink failure temperatures derived from lot sample testing or material characterization tests	Probability of inadvertent nuclear detonation < 10^{-6} per initiating event
Abnormal Mechanical	Stress or strain	Failure stress or strain of welds and metal housings derived from coupon tests	Probability of inadvertent nuclear detonation < 10^{-6} per initiating event
System-Generated Electro-Magnetic Pulse (SGEMP)	Current	Component failure currents derived from current injection tests on susceptible electronics	No significant degradation of weapon reliability in hostile environments
Thermal Mechanical Shock (TMS)	Shock induced stress	Component yield stress as a function of dynamic stress under self-heating conditions derived from material testing	No significant degradation of weapon reliability in hostile environments

There are many methods by which this sample problem can be solved. Although these guidelines emphasize the "whats," and not the "hows," to illustrate the application of

QMU at different levels of rigor it is useful to analyze this synthetic problem with three common methodologies. The three methods illustrated next are first-order probability, second-order probability, and mixed probability and intervals. The primary distinction between the methods lies in the representation and interpretation of epistemic uncertainty results.

D.2 First-Order Probability

D.2.1 Uncertainty Characterization, Propagation, and QMU Format

First-order probability methods do not distinguish between aleatory and epistemic uncertainties. Both types of uncertainties are characterized by probability distributions, propagated through the model using any probabilistic method such as Monte Carlo, and the results of the propagation are then presented in a probabilistic format. Based on the problem specification, the probability distributions for the model parameters *a* and *b* are represented by uniform PDFs (shown as linear distributions when plotted as cumulative probabilities, as shown in Figure D-1).

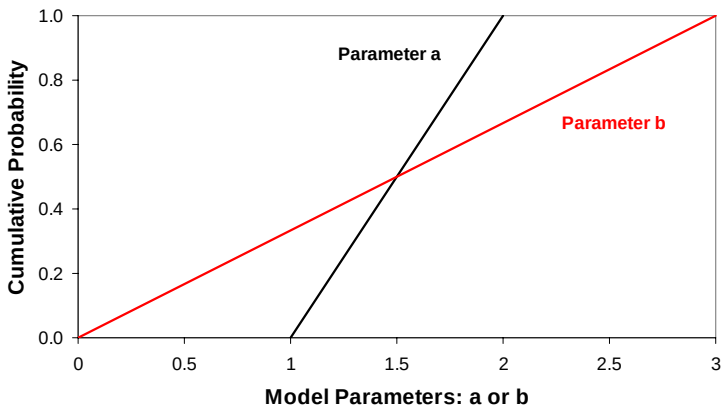


Figure D-1. Input distributions for model parameters *a* and *b*.

The analysis process proceeds as follows. Parameters *a*, *b*, and R_T are sampled 25 times (the computational budget) using Latin hypercube sampling (LHS). In doing so, the strong assumption of independence had to be invoked. A theoretical argument is likely possible for the independence of the threshold criterion from the model parameters, but there is no evidence that the model parameters are mutually independent. The 25 triplets of random inputs are listed in Table D-2. Based on the random pairing of parameters *a* and *b*, the code returns an assessed value for *R* so that the difference, $\Delta R = R_T - R$, can be computed. These quantities are listed in Table D-2 as outputs. The cumulative distributions for outputs can be formed by sorting the outputs from low to high. Positive values of ΔR mean that the threshold function is not exceeded for the conditions randomly selected for the input triplet. No negative values were computed for the 25 randomly selected input triplets; however, a sample size of 25 is not adequate to resolve probabilities on the order (0.01) defined in the requirements.

Table D-2. Computationally Constrained Assessment of QMU for Synthetic Problem

	Inputs			Outputs				Sorted Outputs	
Sample	**a**	**b**	**R_T**	**R**	**ΔR**	**CumP**	**R_T**	**R**	**ΔR**
1	1.891	2.473	9.828	4.834	4.994	0.02	3.588	1.009	2.395
2	1.082	0.456	10.781	1.037	9.745	0.06	4.653	1.010	3.227
3	1.301	2.609	8.525	1.985	6.540	0.10	5.492	1.037	3.326
4	1.846	1.202	8.999	2.090	6.909	0.14	5.729	1.087	3.623
5	1.488	2.398	6.216	2.593	3.623	0.18	6.216	1.110	3.785
6	1.145	2.094	4.653	1.327	3.326	0.22	6.570	1.193	3.979
7	1.439	1.864	11.059	1.970	9.089	0.26	6.720	1.226	4.162
8	1.195	0.991	3.588	1.193	2.395	0.30	6.860	1.309	4.482
9	1.956	0.851	6.860	1.770	5.090	0.34	7.077	1.327	4.490
10	1.337	2.227	9.417	1.911	7.506	0.38	7.477	1.338	4.994
11	1.360	1.565	8.134	1.618	6.516	0.42	7.618	1.567	5.090
12	1.461	2.838	6.720	2.934	3.785	0.46	7.788	1.579	5.827
13	1.254	2.014	9.296	1.579	7.718	0.50	7.979	1.618	6.139
14	1.995	0.014	5.492	1.009	4.482	0.54	8.134	1.770	6.508
15	1.594	2.891	7.077	3.850	3.227	0.58	8.492	1.911	6.516
16	1.212	1.399	10.171	1.309	8.863	0.62	8.525	1.961	6.540
17	1.070	1.532	7.618	1.110	6.508	0.66	8.917	1.970	6.909
18	1.668	0.569	7.477	1.338	6.139	0.70	8.999	1.985	7.482
19	1.836	1.108	7.788	1.961	5.827	0.74	9.296	2.079	7.506
20	1.638	0.168	11.788	1.087	10.702	0.78	9.417	2.090	7.691
21	1.743	0.808	5.729	1.567	4.162	0.82	9.828	2.593	7.718
22	1.689	2.645	7.979	4.000	3.979	0.86	10.171	2.934	8.863
23	1.017	0.613	8.492	1.010	7.482	0.90	10.781	3.850	9.089
24	1.774	0.355	8.917	1.226	7.691	0.94	11.059	4.000	9.745
25	1.537	1.704	6.570	2.079	4.490	0.98	11.788	4.834	10.702

There are two common approaches to dealing with the limited sample size. In the first approach, a theoretical distribution (e.g., a normal or lognormal distribution) for ΔR is fit to the limited information (based on 25 samples in the present example). The theoretical distribution can then be used to calculate $\text{Prob}\left(R > R_T\right)$ and the result can be compared to the requirement. This involves additional "strong assumptions" about the distribution form; consequently, this approach was not pursued for this example.

The second common approach, and the one illustrated here, for dealing with limited sample sizes is to develop a surrogate model for the computationally expensive computer model. A second-order polynomial, derived from a regression fit to data in Table D-2, provides an excellent surrogate model for the computer model:

$$R' = 3.77623 - 3.0562a - 2.7060b + 0.8003a^2 + 0.2032b^2 + 1.9603ab \ .$$

The statistic $r^2 = 0.986$ is a measure of how well the surrogate model fits the actual computational data, with $r^2 = 1$ denoting a perfect fit. Figure D-2 shows the residuals between the actual (code-based) values and the values provided by the surrogate model. The minimum and maximum values of the residuals are –0.20 and 0.26, respectively; there is no trend in the residuals with either parameter.

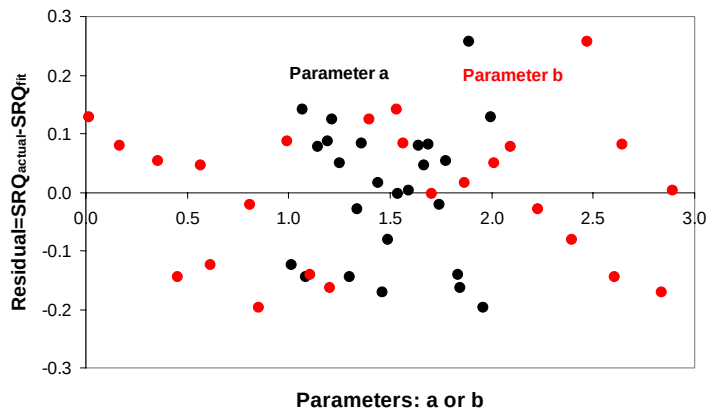


Figure D-2. Residuals between code-based and surrogate model–based assessments of *R*.

The analysis listed in Table D-2 can now be repeated using the surrogate model in place of the computationally expensive computer model. Because the surrogate model is algebraic, we can easily afford 10,000 LHS samples. The results are depicted graphically in Figure D-3. The model-based assessment of *R* is not Gaussian. The curve for ΔR is particularly meaningful and a blowup around the region of $\Delta R = 0$ is shown in Figure D-4. Negative values for ΔR occur when an M&S-based assessment of *R* exceeds the threshold value. The value of $\text{Prob}\left(R > R_T\right)$ is 0.00715 and corresponds to the point on the cumulative curve where $\Delta R = 0$.

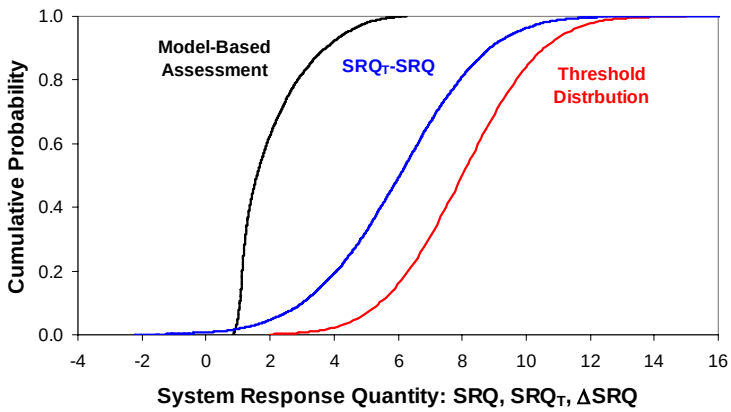


Figure D-3. Results of first-order probability method applied to synthetic problem.

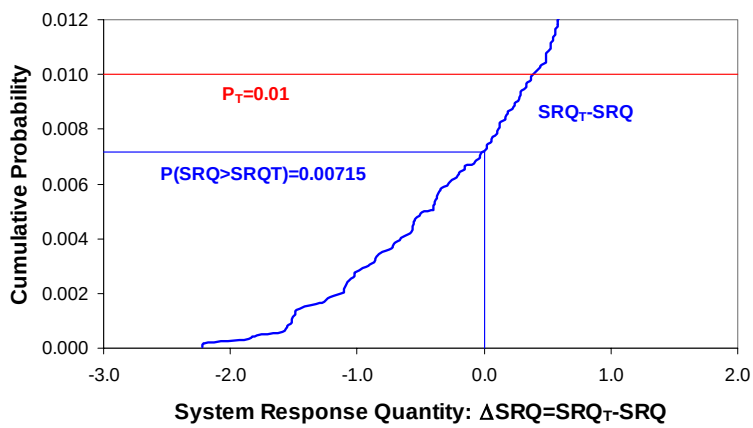


Figure D-4. Probability that the M&S-based assessment of R exceeds the threshold R_T.

Keeping in mind that the requirement is specified in probabilistic language, we are now in a position to evaluate the three QMU figures of merit. Figure D-4 shows that

$$\text{Prob}\left(R > R_T\right) = 0.00715\,,$$

which could be expressed as a reliability, $Rel = 1 - 0.00715 = 0.99285$. The safety factor ($SF$), as defined in Table 1 of the main body of this paper, is

$$SF = \frac{\text{Required Probability}}{\text{Assessed Probability}} = 0.01 \big/ 0.00715 = 1.40$$

The margin M is defined as the difference between the required probability and the assessed probability:

$$M = 0.01 - 0.00715 = 0.00285.$$

However, we are at a loss to compute the confidence factor (*CF*) because there is no "uncertainty" associated with the computed value of $\text{Prob}\left(R > R_T\right)$. This is a direct consequence of the fact that aleatory and epistemic uncertainties were not addressed separately in the method. We must conclude that the confidence factor is an ill-posed figure of merit when requirements are expressed in probabilistic language and aleatory and epistemic uncertainties are addressed without distinction.

It is possible to "force" a confidence factor representation of QMU for the synthetic problem by focusing on the M&S-assessed distribution and the threshold distribution for *R*. The necessary terms are illustrated in Figure D-5. The margin is defined as the difference in median values of the two distributions:

$$M = 8 - 1.5854 = 6.4146.$$

High-confidence uncertainties (referenced to the respective medians) are 4.170 and 2.812 for the threshold and model-based distributions, respectively, allowing an aggregate uncertainty to be defined as

$$U = \sqrt{U_T^2 + U^2} = 5.486$$

The confidence factor can now be computed as

$$CF = M/U = 6.415/5.486 = 1.17.$$

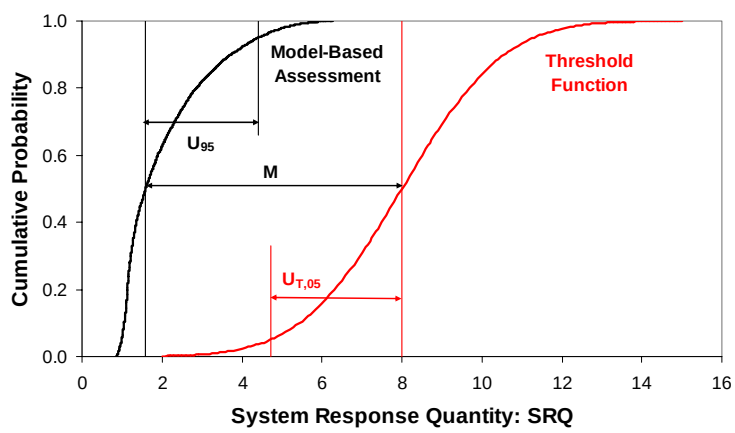


Figure D-5. Confidence factor (*CF*) format for QMU.

The confidence factor, so computed, does not directly measure compliance to the requirement of $\text{Prob}\left(R > R_T\right) < 0.01$; and, in this case, the confidence factor provides a less precise and potentially nonconservative representation of QMU results compared to the reliability or safety-factor figures of merit.

D.2.2 Sensitivity Analysis

Sensitivity analyses (SAs) are an important component of QMU. Here, we assess sensitivity by examining scatterplots of ΔR as a function of the uncertain inputs: a, b, and threshold values of R. ΔR is the appropriate dependent variable for the SA because its distribution completely determines $\text{Prob}\left(R > R_T\right)$. Figure D-6 visually illustrates that ΔR has a strong linear correlation with threshold values and no apparent correlation with the model parameters a and b. This suggests that the threshold distribution dominates the results. Because of the *observed* linear dependence, the squared partial correlation coefficients are a quantitative measure of sensitivity:

$$r^2(\Delta R;\, a) = 0.0493$$

$$r^2(\Delta R;\, b) = 0.1539$$

$$r^2(\Delta R;\, R_T) = 0.7945.$$

Quantitative SA suggests that the irreducible uncertainty associated with the threshold distribution dominates all uncertainties and that the uncertainty in parameter b exceeds the uncertainty in parameter a by a factor of ~3 amongst the reducible uncertainties.

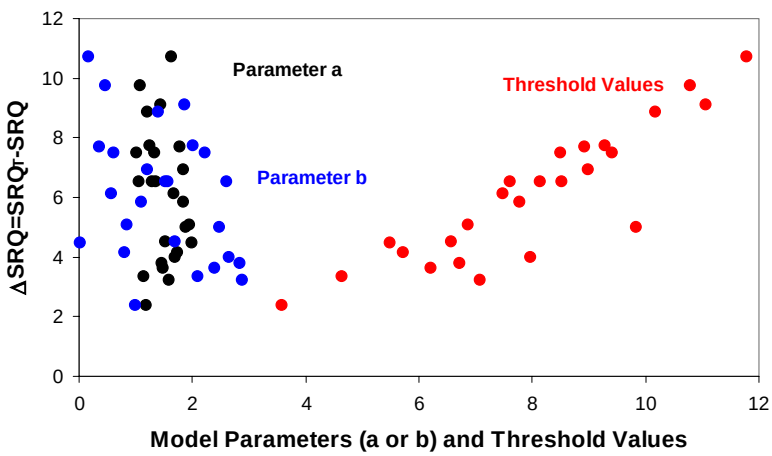


Figure D-6. Scatterplots to assess sensitivity of QMU results to input variability and uncertainties.

D.2.3 Assessment of Numerical Errors

There are two sources of numerical errors in the current assessments: standard errors resulting from limited sampling (25 function evaluations) and representational errors associated with the approximate surrogate-model representation of the complete physics model. Table D-3 summarizes the sensitivity of numerical errors in QMU results to these two sources of errors. The first column summarizes the baseline results already discussed. The next two columns explore sensitivity to representational errors associated with the use of the surrogate model. Results were obtained by shifting surrogate model results by ± 0.26. This range envelops all results of the physics-based model. Lastly, sensitivity to

finite sampling is explored in the last column. This is accomplished by randomly sampling 25 triplets (with replacement) from the original sample set of 25. This process is called "bootstrapping." The analysis process (fitting a new response surface and sampling the new response surface 10,000 times) was repeated. In no case did an exploration of sensitivity to numerical errors lead to a result that violates requirements.

Table D-3. Sensitivity to Sources of Numerical Error in First-Order Probability Results

	nom	**low -.26**	**high +.26**	**bootstrap**
Margin	6.415	6.675	6.155	6.381
U_{95}	2.812	2.812	2.812	2.835
$U_{T,05}$	4.710	4.710	4.710	4.710
U	5.486	5.486	5.486	5.498
CF	1.169	1.217	1.122	1.161
$\text{Prob}(R > R_T)$	0.00715	0.00615	0.00915	0.00685
P margin	0.00285	0.00385	0.00085	0.00315
SF	1.399	1.626	1.093	1.460

D.3 Second-Order Probability

D.3.1 Uncertainty Characterization, Propagation, and QMU Format

Second-order probability methods separate the representation and interpretation of aleatory and epistemic uncertainties. Aleatory and epistemic uncertainties are still characterized by probability distributions, but the analysis process proceeds in two steps. As preparation, we sample the model parameters, a and b, 25 times and compute the respective R values with the model. These values are shown in Table D-4 and are identical to those used in the first-order probability model.

For the first step, $\text{Prob}\left(R > R_T\right)$ is first computed conditional on one set of uncertain values for a and b. Because the threshold distribution is normal, $\text{Prob}\left(R > R_T\right)$ here can be computed exactly and analytically to be 0.05669 (interpreted as frequency). This is an example of probabilistic propagation conditional on the specific value of the epistemic R. In the second step, this process is repeated for all 25 rows with the results sorted from small to large. The corresponding safety factor (SF) is computed as the ratio of the required probability and the assessed probability.

Table D-4. Results of Second-Order Probability Analysis

Sample	a	b	R	Prob($R > R_T$)	P sort	SF sort	CumP
1	1.891	2.473	4.834	0.05669	0.00024	0.18	0.02
2	1.082	0.456	1.037	0.00025	0.00024	40.14	0.06
3	1.301	2.609	1.985	0.00132	0.00025	7.59	0.10
4	1.846	1.202	2.090	0.00156	0.00027	6.40	0.14
5	1.488	2.398	2.593	0.00343	0.00029	2.91	0.18
6	1.145	2.094	1.327	0.00042	0.00033	23.56	0.22
7	1.439	1.864	1.970	0.00128	0.00035	7.79	0.26
8	1.195	0.991	1.193	0.00033	0.00041	30.07	0.30
9	1.956	0.851	1.770	0.00092	0.00042	10.88	0.34
10	1.337	2.227	1.911	0.00116	0.00043	8.59	0.38
11	1.360	1.565	1.618	0.00071	0.00065	14.10	0.42
12	1.461	2.838	2.934	0.00566	0.00066	1.77	0.46
13	1.254	2.014	1.579	0.00066	0.00071	15.10	0.50
14	1.995	0.014	1.009	0.00024	0.00092	42.23	0.54
15	1.594	2.891	3.850	0.01899	0.00116	0.53	0.58
16	1.212	1.399	1.309	0.00041	0.00127	24.37	0.62
17	1.070	1.532	1.110	0.00029	0.00128	35.03	0.66
18	1.668	0.569	1.338	0.00043	0.00132	23.12	0.70
19	1.836	1.108	1.961	0.00127	0.00154	7.90	0.74
20	1.638	0.168	1.087	0.00027	0.00156	36.58	0.78
21	1.743	0.808	1.567	0.00065	0.00343	15.42	0.82
22	1.689	2.645	4.000	0.02274	0.00566	0.44	0.86
23	1.017	0.613	1.010	0.00024	0.01899	42.15	0.90
24	1.774	0.355	1.226	0.00035	0.02274	28.32	0.94
25	1.537	1.704	2.079	0.00154	0.05669	6.51	0.98

The sorted values can be used to construct distributions for $\text{Prob}\left(R > R_T\right)$ and *SF*. Any one realization of epistemic uncertainties results in a frequency of violating the threshold function. Doing this many times for different epistemic inputs creates a probability (confidence) distribution of frequency values. This probability of frequency format (first proposed by Kaplan and Garrick 1981) is essentially a second-order probability scheme. These results are shown in Figures D-7 and D-8. We conclude that there is ~88% confidence that the required threshold probability of 0.01 is not violated; and if we demand 95% confidence, then we reject the conclusion that requirements are met. Note that in second-order probability the epistemic results are presented in a probabilistic format.

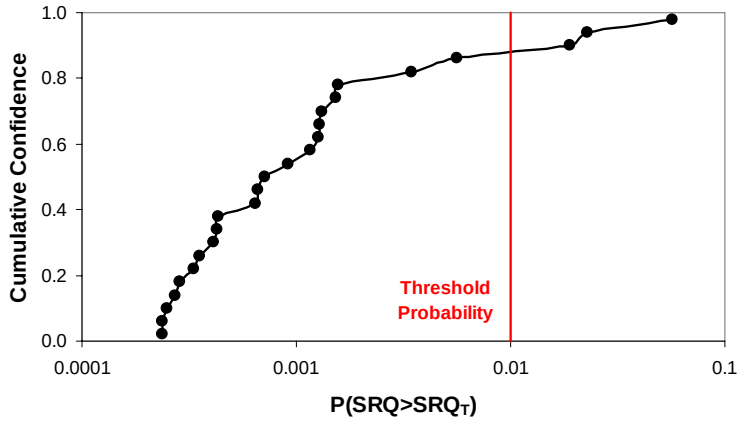


Figure D-7. Results of second-order probability analysis.

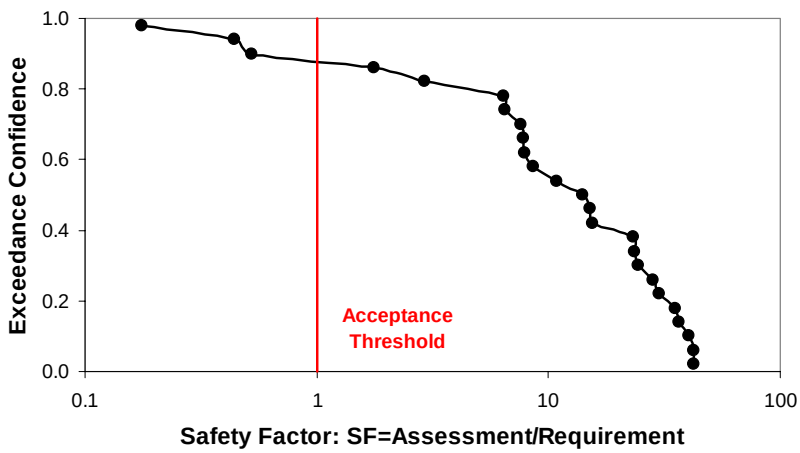


Figure D-8. Confidence in computed safety factors resulting from epistemic uncertainties.

In the context of second-order probability, the confidence factor can be rigorously computed with the aid of Figure D-9. The margin is defined as the difference in the requirement and the median of the assessed distribution:

$$M = 0.01 - 0.0007091 = 0.009291.$$

Uncertainty (at high confidence) is defined as the difference in the 95 percentile of the distribution and the median:

$$U = 0.027737 - 0.0007091 = 0.022028.$$

The confidence factor CF is then easily computed as

$$CF = M/U = 0.009291/0.022028 = 0.422.$$

The confidence factor suggests that requirements are not met at the 95% confidence level.

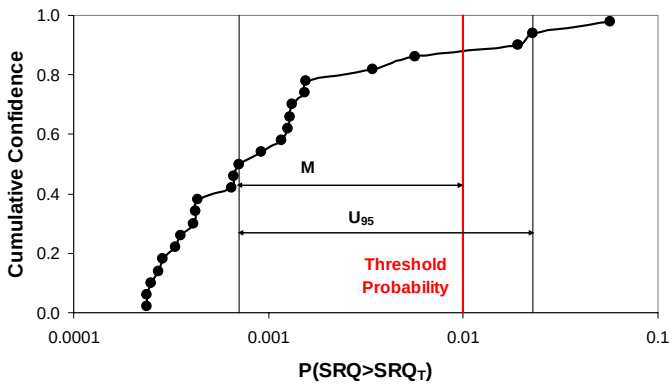


Figure D-9. Computation of confidence factor for second-order probability.

D.3.2 Sensitivity Analysis

The sensitivity of $\text{Prob}\left(R > R_T\right)$ to uncertainties in parameters *a* and *b* can be explored through a scatterplot, as depicted in Figure D-10.

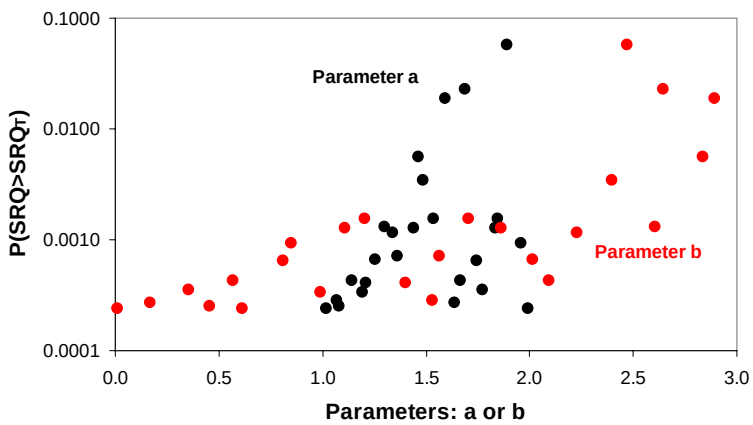


Figure D-10. Scatterplot for sensitivity analyses in second-order probability.

We see that $\ln\left[\text{Prob}\left(R > R_T\right)\right]$ is approximately linear to variations in both parameters. It is thus reasonable to quantify the sensitivity with the square of partial correlation coefficients as

$$r^2\left\{\ln\left[\text{Prob}\left(R > R_T\right)\right], a\right\} = 0.134$$
$$r^2\left\{\ln\left[\text{Prob}\left(R > R_T\right)\right], b\right\} = 0.577$$

In the case of second-order probability, note that it only makes sense to measure sensitivity against the epistemic parameters. The sensitivity due to parameter *b* exceeds the sensitivity due to parameter *a* by a factor of ~4, which is similar to results observed

with first-order probability. In this context, the threshold distribution is considered precisely known and normal.

D.3.3 Assessment of Numerical Errors

Here, we quantify the impact of numerical errors on the computation of the confidence factor, although similar results could easily be generated for the confidence that $\text{Prob}\left(R > R_T\right) < 0.01$ and the confidence that $SF > 1$. Numerical errors are associated with the finite number of samples (25) in the baseline study. To quantify and estimate of the numerical errors, we use a bootstrapping process (Davison and Hinkley, 1999) common to such applications. The process has its basis in the fact that the 25 values of $\text{Prob}\left(R > R_T\right)$ are representative of the population they approximate. In bootstrapping, we randomly select 25 values (with replacement) from the already computed set of 25 $\text{Prob}\left(R > R_T\right)$ values. This new set of 25 $\text{Prob}\left(R > R_T\right)$ values is then sorted small to large; and *M*, *U*, and *CF* are then calculated, providing for a second estimate of *CF*. This process of "resampling" can be repeated multiple times (25 in the current example), sorted, and arranged into a cumulative distribution, as shown in Figure D-11. The figure shows ~74% confidence that $CF < 0.51$; however, there is 26% belief that *CF* could exceed 1 just because of an unrepresentative sample of 25 chosen in the baseline study.

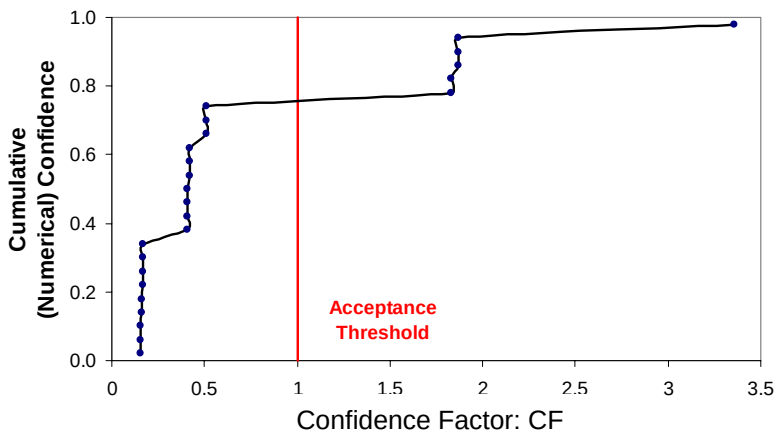


Figure D-11. Confidence in the computed confidence factor *CF* associated with numerical errors.

D.4 Mixed Probability/Interval Analysis

D.4.1 Uncertainty Characterization, Propagation, and QMU Format

In second-order probability, $\text{Prob}\left(R > R_T\right)$ is computed conditional on a set of epistemic inputs. The process is then repeated for all sets of epistemic inputs, the results sorted, and the results represented in a probabilistic format (cumulative probability distributions). This allows statements of the type "there is *x*% confidence that the requirement is not exceeded."

Mixed probability/interval analysis differs from second-order probability only in the way that epistemic results are presented and interpreted. Since all epistemic inputs were from intervals (with no evidence to believe any value more than another), the "uncertainty-preserving" presentation of results would represent all epistemic outputs as intervals as well, with no evidence to believe any value more than another. Consequently, all values of $\text{Prob}\left(R > R_T\right)$ listed in Table D-4 are interpreted as just "possible." The fact that the ensemble of 25 $\text{Prob}\left(R > R_T\right)$ values was generated from a Monte Carlo sampling of parameters *a* and *b* has no probabilistic meaning. The Monte Carlo sampling was just one method to generate the ensemble of $\text{Prob}\left(R > R_T\right)$ values for a range of *a* and *b* values that span the range of possible values and their interactions. Furthermore, there is no assumption of independence in *a* and *b*. The ensemble of $\text{Prob}\left(R > R_T\right)$ values is depicted in Figure D-12, where the 25 solid symbols are the values of $\text{Prob}\left(R > R_T\right)$ listed in Table D-4. The estimated minimum and maximum values are addressed as part of the discussion of numerical errors. Note that some possible values of $\text{Prob}\left(R > R_T\right)$ exceed the requirement; consequently, the possibility of $\text{Prob}\left(R > R_T\right)$ exceeding the requirement cannot be excluded. In a very similar way and with similar conclusions, the safety factor can be presented in an "uncertainty-preserving" format as depicted in Figure D-13.

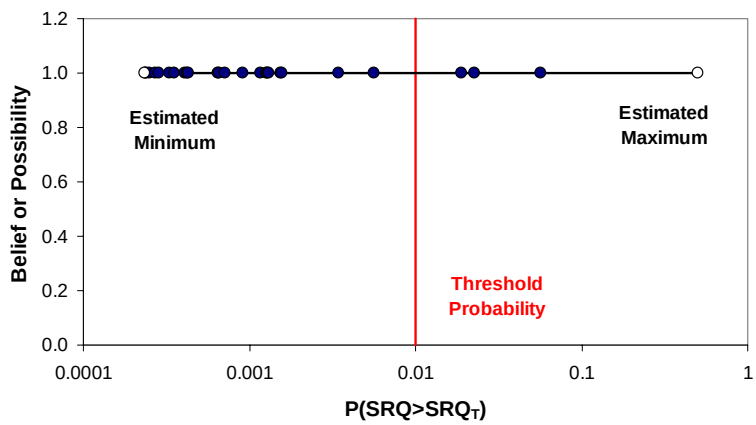


Figure D-12. "Uncertainty-preserving" presentation of epistemic results with a failure probability metric.

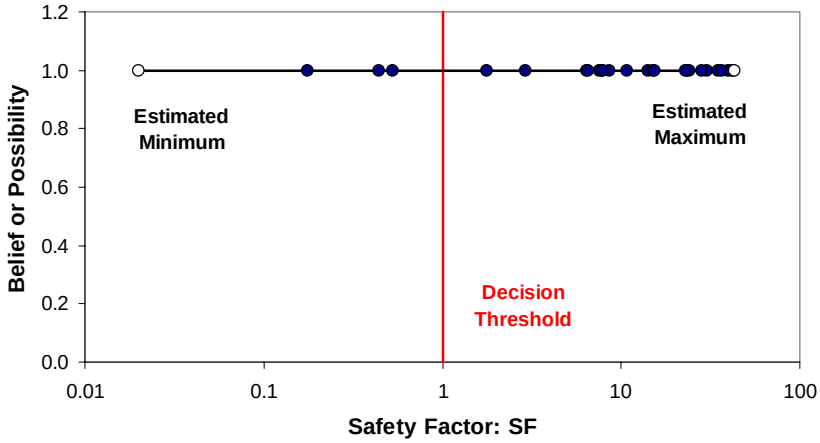


Figure D-13. "Uncertainty-preserving" presentation of epistemic results with a safety factor metric.

The confidence factor can be computed with the aid of Figure D-12. The median of the range of possible outputs is taken as the midrange of the interval:

$$\text{Median} = (0.5 + 0.000233)/2 = 0.2501.$$

The margin is then given by

$$M = \text{Threshold Probability} - \text{Median} = 0.01 - 0.2501 = -0.24,$$

which already indicates that requirements are not met because of its negative value. The uncertainty is given by

$$U = \text{Maximum} - \text{Median} = 0.5 - 0.2501 = 0.25.$$

Consequently, the confidence factor can be computed as

$$CF = M/U = -0.24/0.25 = -0.96;$$

and the fact that CF is negative indicates that the performance threshold is potentially violated because the margin itself is negative.

D.4.2 Sensitivity Analysis

The SAs performed for the second-order probability analysis are equally applicable here and need not be repeated.

D.4.3 Assessment of Numerical Errors

The 25 "possible" values of $\text{Prob}\left(R > R_T\right)$ listed in Table D-4 only coarsely define the range of possible outputs. Here, the assessment of numerical errors becomes an exploration for the maximum and minimum values defining the interval. The SA suggests

that R is positively monotonic in both parameter a and parameter b; consequently, the maximum and minimum values of $\text{Prob}\left(R > R_T\right)$ are conditional on the maximum values of a and b and the minimum values of a and b, respectively:

$$\min\left[\text{Prob}\left(R > R_T\right)\right] = 0.000233$$
$$\max\left[\text{Prob}\left(R > R_T\right)\right] = 0.5$$

D.5 Comments on the Three Methodologies

Table D-5 summarizes a comparison of the three methodologies as they were applied here. Notable is the fact that the three methodologies support different conclusions with regard to satisfying the requirements. Consistent with evidence provided in the problem specification, we can now reveal (e.g., through research in a real application) that the exact values of a and b are 1.75 and 2.5, respectively, resulting in a model-based evaluation of $R = 4.05$ and $\text{Prob}\left(R > R_T\right) = 0.024$, which exceeds the requirement. The first-order probabilistic methods underrepresent the impact of epistemic uncertainties and lead to an incorrect conclusion regarding compliance with the requirements. Second-order probability, on the other hand, produces a range of results that envelops reality; however, a decision maker might be surprised by reality because of the high confidence that values as large as 0.024 would not be achieved. This is a direct consequence of representing the epistemic results in a probabilistic format. Only the mixed probability/interval presentation format correctly enveloped reality without underrepresenting the uncertainties. Because of its greater consistency in representing epistemic uncertainties, mixed second-order probability is more appropriate when judging compliance to absolute requirements (as in qualification), while first-order probability can be useful when making relative judgments as might occur in design support (this design is better than that design). In the case of first-order probability, the figure of merit for the confidence factor could not be calculated in a manner consistent with decision requirements common to many Sandia applications. Consequently, Sandia reserves the right, on a case-by-case basis, to select figures of merit appropriate to its applications. The first two methodologies illustrated here, although in widespread use, are only two of many methodologies that could have been used. Because we used first- and second-order probability to anchor QMU concepts in an illustrative example, it should not be inferred that other methodologies cannot be applied successfully to Sandia's applications.

Table D-5. Results Comparison for First-Order and Second-Order Probability Methods

	First-Order Probability	**Second-Order Probability**	**Mixed Probability/Interval**
Uncertainties	Aleatory and epistemic uncertainties treated without distinction • All candidate figures of merit lead to conclusion that requirements *are* met • Confidence factor (*CF*) could not be calculated in a manner consistent with the decision requirements	Aleatory and epistemic uncertainties represented separately • All candidate figures of merit lead to conclusion that requirements *are not* met	Aleatory and epistemic uncertainties represented separately • All candidate figures of merit lead to conclusion that decision requirements *are not* satisfied
Sensitivity Analysis	Sensitivities quantified • Uncertainties in threshold distribution (irreducible) dominate all uncertainties • Uncertainties in model parameter *b* are ~3 times more important than uncertainties in parameter *a* amongst the reducible uncertainties	Sensitivities quantified • Uncertainties in model parameter *b* are ~4 times more important than uncertainties in parameter *a*	Sensitivities quantified • Uncertainties in model parameter *b* are ~4 times more important than uncertainties in parameter *a*
Numerical errors	Sensitivity to numerical errors explored • No case considered violates requirements	• Numerical errors quantified for *CF* • Confident that requirements *are not* met but some possibility that they might be	Numerical errors quantified • Limits of output interval quantified • Highly confident that requirements *are not* met
Assessed Level of Rigor	Low/medium rigor adequate for design support	Medium/high rigor appropriate for qualification support	High rigor appropriate for qualification

D.6 References

1. Davison, A. C. and D. V. Hinkley. (1999). *Bootstrap Methods and Their Application*. Cambridge, UK: Cambridge University Press.

2. Kaplan, S., and B. J. Garrick. (1981). "On the Quantitative Definition of Risk." *Risk Analysis* 1, no. 1: 11–27.

Distribution

External Distribution

Department of Energy (5)
Forrestal Building
1000 Independence Ave., SW
Washington, DC 20585
Attn: Dave Cameron, NA-115
Njema Frazier, NA-114
Kevin Greenaugh, NA-115
D. Kusnezov, NA-114
Karen Pao, NA-114

Los Alamos National Laboratory (9)
Mail Station 5000
P.O. Box 1663
Los Alamos, NM 87545
Attn: Mark C. Anderson, MS T080
Jerry S. Brock, MS F663
Scott Doebling, MS T080
Francois Hemez, MS F699
David Higdon, MS F600
Charles Nakhleh, MS T085
James Kamm, MS D413
David Sharp, MS B213
Robert Webster, MS F644

Lawrence Livermore National Laboratory (4)
7000 East Ave.
P.O. Box 808
Livermore, CA 94551
Attn: Frank Graziani, MS L-095
Richard Klein, MS L-023
Joe Sefcik, MS L-160
Richard Ward, MS L-160

Sandia Internal Distribution

1	MS0351	01000	R. H. Stulen
1	MS0321	01400	J. S. Peery
1	MS1110	01410	D. E. Womble
1	MS0370	01411	J. S. Stewart
10	MS0370	01411	T. G. Trucano
1	MS0316	01420	S. S. Dosanjh

1	MS0321	01430	J. E. Nelson
10	MS0828	01221	M. Pilch
1	MS0384	01500	A. C. Ratzel
1	MS0828	01544	A. A. Giunta
1	MS0457	02000	J. S. Rottler
1	MS0457	02010	J. M. Freedman
1	MS0457	02020	G. Novotny
1	MS0429	02100	B. C. Walker
10	MS0427	02118	R. A. Paulsen, Jr.
10	MS0427	02118	S. E. Klenke
1	MS9001	08000	P. J., Hommert
1	MS9153	08200	C. L. Knapp
10	MS0830	12335	K. V. Diegert

1	MS 9018	8945-1	Central Technical Files
2	MS 0899	9616	Technical Library