UCRL-ID-150874

Verification & Validation (V&V) Methodology and Quantitative Reliability at Confidence (QRC): Basis for an Investment Strategy

R. W. Logan, C. K. Nitta

July 17, 2002

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

Approved for public release; further dissemination unlimited

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of 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 or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

This report has been reproduced directly from the best available copy.

Available electronically at http://www.doc.gov/bridge

Available for a processing fee to U.S. Department of Energy
And its contractors in paper from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-mail: reports@adonis.osti.gov

Available for the sale to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/ordering.htm

OR

Lawrence Livermore National Laboratory
Technical Information Department's Digital Library
http://www.llnl.gov/tid/Library.html

Verification & Validation (V&V) Methodology and Quantitative Reliability at Confidence (QRC): Basis for an Investment Strategy

Roger W. Logan
Cynthia K. Nitta

University of California
Lawrence Livermore National Laboratory

This work represents the views of the authors and not necessarily the views of Lawrence Livermore National Laboratory or University of California

Table of Contents

0 Executive Summary: 3
1 PART I: V&V definitions, methods, metrics, guidelines, and checklists 6
1.1 Introduction: V&V, Stockpile Stewardship, and the Decision Process 6
1.2 Definitions: Verification and Validation 10
1.3 General Statement of Methodology: 12
1.3.1 Methodology and Independence Metrics 14
1.3.2 Metrics: VERIFICATION and VALIDATION 16
1.4 Verification and Validation Meters and Checklists 17
1.4.1 The Qualitative VER and VAL Meters 19
1.4.2 10 (or 30) steps to a V&V process: Summary Level Considerations 24
1.4.3 19 Verification Points to help set the VER Meter: 31
1.4.4 35 Validation Points to help in setting the VAL Meter 31
1.4.5 Relating the 84 point criteria to VER and VAL Meters and other criteria 33
1.4.6 V&V Meters and Checklist Steps: How many is enough? 35
1.5 Relation of V&V to Risk=Likelihood*Consequence 37
2 PART II: Relation of V&V to Adequacy, Acceptance, Qualification, and Certification 38
2.1 Relation to Certification Methodology, Metrics, and Prioritization 38
2.2 Moving from V&V and UQ to QRC: Quantified Reliability at Confidence 43
2.2.1 Measures of M and U : As used in FOM/QMU, CF, and QRC 44
2.2.2 Contrasting CF , FOM/QMU and QRC 47
2.2.3 Features of QRC: The Good, the Bad, and the Ugly 50
2.3 Example comparison of V&V methods with and without closure 53
2.4 Example of Process to a Quantitative Validation Statement 57
2.4.1 Model Fit and Uncertainty 57
2.4.2 Confidence Bounds on Predictive Extrapolations 62
2.5 The QRC Matrix: A first step toward prioritization 64
2.6 V&V to QRC: Relation to Risk=Likelihood*Consequence 67
3 PART III: Deployment, Resources, Value Engineering, and Investment Strategy 68
3.1 Part IIIa: Priority and Timing: Deployment and Resources 69
3.1.1 Schedule and Timing: The Ladder Chart 72
3.1.2 Examples of Ladder Success 73
3.1.3 Two Ladders — Branches — and a Trellis 74
3.1.4 How V&V Level determines need for machine capability, capacity 76
3.1.5 Speed, the Multiple Code Shootout , Timing, and Risk Mitigation 79
3.1.6 Measures of Analysis Horsepower 81
3.1.7 Earned Value expressed as Enhanced QRC in The Matrix 82
3.2 Part IIIb: Quantitative Systems Value (QSV) and Investment Strategy 84
3.2.1 Introduction to QSV: Quantitative Stockpile Value 86
3.2.2 A Simple Everyday Illustration of QSV and BCR: Stopping at a Traffic Light: 88
3.2.3 The process to get QSV and BCR: 90
4 Summary and Conclusions 95
4.1 Part I: V&V and UQ 95
4.2 Conclusions: Part II (QRC) into Part III (QSV) 96
4.3 Conclusions and the Future: Improvements and Refinements 97
5 Acknowledgements: 97
6 Appendices 98
6.1 Appendix A: Verification & Validation Definitions given by others 98
7 List of Symbols, Abbreviations, and Acronyms: 100
8 References: 101
9 Distribution 105

0 EXECUTIVE SUMMARY:

This paper represents an attempt to summarize our thoughts regarding various methods and potential guidelines for Verification and Validation (V&V) and Uncertainty Quantification (UQ) that we have observed within the broader V&V community or generated ourselves. Our goals are to evaluate these various methods, to apply them to computational simulation analyses, and integrate them into methods for Quantitative Certification techniques for the nuclear stockpile. We describe the critical nature of high quality analyses with quantified V&V, and the essential role of V&V and UQ at specified Confidence levels in evaluating system certification status. Only after V&V has contributed to UQ at confidence can rational tradeoffs of various scenarios be made. UQ of performance and safety margins for various scenarios and issues are applied in assessments of Quantified Reliability at Confidence (QRC) and we summarize with a brief description of how these V&V generated QRC quantities fold into a Value-Engineering methodology for evaluating investment strategies. V&V contributes directly to the decision process for investment, through quantification of uncertainties at confidence for margin and reliability assessments. These contributions play an even greater role in a Comprehensive Test Ban Treaty (CTBT) environment than ever before, when reliance on simulation in the absence of the ability to perform nuclear testing is critical.

Our summary is broken down into three major sections (Part I-III).

In Part I, V&V methods are described and methods to quantify Verification and Validation are presented. Methods to quantify UQ as closeness of fit at Confidence are described, including frequentist or inference type methods. A compilation of frequently used and relevant concepts for metrics are suggested for consideration when quantifying the V&V status of a given physical model for a particular Stockpile to Target Sequence (STS) environment and a given code. We also describe an idea for quick recognition of the V&V level for particular simulation capabilities in a Meter numbering system.

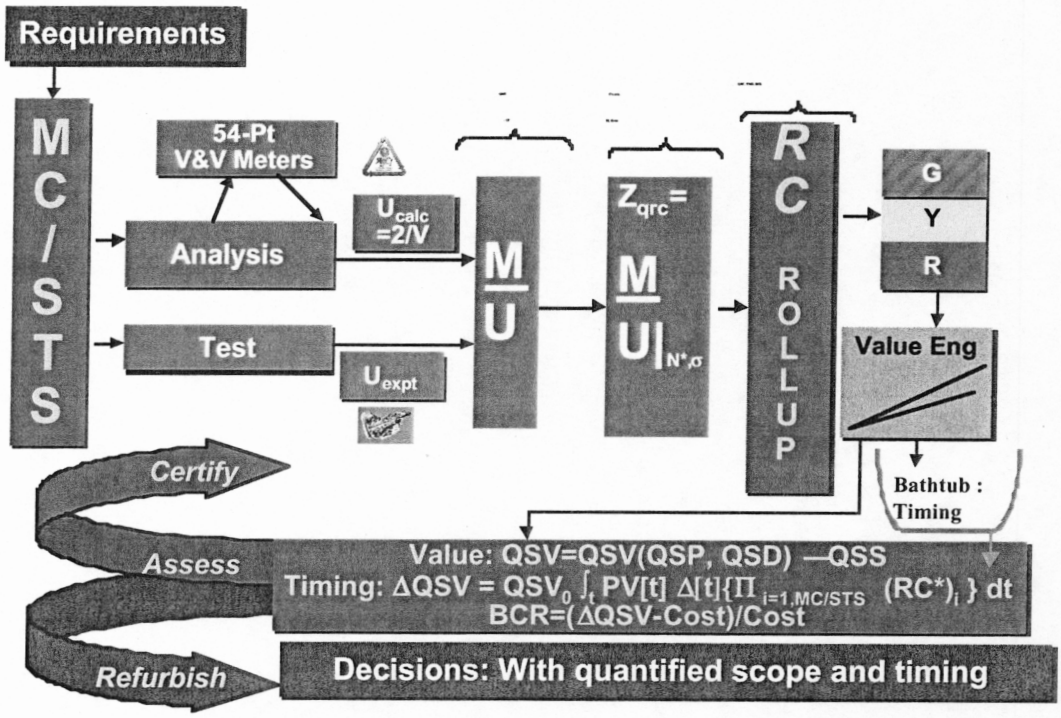
In Part II, we describe how validated models, with UQ at Confidence, are used to evaluate Margins for components and systems. Given uncertainty and margin values, we can then assess quantified measures of reliability at confidence (QRC).

In Part III, we briefly describe a conceptual model for quantifying business decisions based on inputs from V&V and QRC, which include consideration of priority, timing, deployment, and investment strategy. A Quantified Systems Value (QSV) is defined as a function of Reliability and Confidence in terms of Benefit (improvement in Value) and Benefit/Cost Ratios (BCR). We demonstrate the linkage of V&V level for particular simulation capabilities (including validation experiments) to the value of products and product decisions made under budget and schedule constraints. A concept of closure is introduced in the form of a simple equation that integrates UQ, QRC, and QSV quantities with the economic function of Present Value Factor (PV_F) in the time domain:

$$\Delta QSV = QSV_0 \int_t PV_F[t]\, \Delta[t] \{\Pi_{i=1,MC/STS}\ (RC^*)_i\}\, dt \qquad [1]$$

This equation enables quantification of Benefit/Cost tradeoffs and timing decisions (Logan and Nitta, 2000). **Although there is not a unique BCR, we should explore the bounds of its values for any given decision and we show its relationship to quantified V&V. These concepts are evaluated for particular requirements,** which in our case are represented by the MC (Military Characteristics) and STS (Stockpile to Target Sequence) requirements of our Department of Defense (DoD) colleagues. For general business application, the requirements would be those environments that determine product performance.

To work through the process that leads through V&V and eventually to our investment strategy, it will be useful to track our progress through the methods (and acronyms) by referring to the rather complex flow process in Figure S-1. We break this complex diagram into portions, and discuss each portion in turn in the 3-part discussion that follows.



.41

Figure S-1. Flow diagram from system Requirements (MC/STS) through V&V, through uncertainty quantification and margins; onward through QRC, then QSV. The concepts (and the acronyms) are easier to grasp as we address them in the 3 Parts below. See Section 7 for a list of symbols and acronyms.

Beginning with Part I, the first step is to establish system Requirements. For our work in the nuclear community, these are the MC (Military Characteristics or a top-level system requirements document) and STS (Stockpile to Target Sequence, a logistics and mission document one level deeper). Based on these requirements for the system and its environments, the V&V

process begins, leading to validated models with uncertainties at confidence as our product of Part I. Part II takes these quantities to margins and then reliability equivalents, on through to a Quantified Reliability at Confidence (QRC) Rollup. Part III continues, making use of the QRC products for Value Engineering and decision processes.

Key to the investment strategy process, and its linkage back to V&V, is the Benefit/Cost Ratio (BCR). Quantified V&V will show us that there is not a unique BCR — we must explore its bounds for any given decision. Due to the non-unique fidelity of any given BCR, it will become apparent that our decisions fall into 3 basic bins:

1. High BCR within our V&V bounds: Positive decision indicator [i.e. do it]
2. Low BCR within our V&V bounds: Negative decision indicator [i.e. don t do it]
3. BCR varies high to low depending on V&V bounds: more quantification is needed

The following 3-part discussion will lead us through Figure S-1 to our compact QSV equation and use of the BCR. **Part 1 (V&V), and Part 2 (Certification) and Part 3 (Value Engineering / QSV) are fields with evolving methodologies. The end product methodology and dollar benefit can be explained using a standard Risk=Likelihood*Consequence Matrix as shown in Figure S-2. This work represents the authors view of a method of closure for the 3 topics. We are not aware of any universally accepted method with such closure, so in this work we present our own work and progress in this area.**

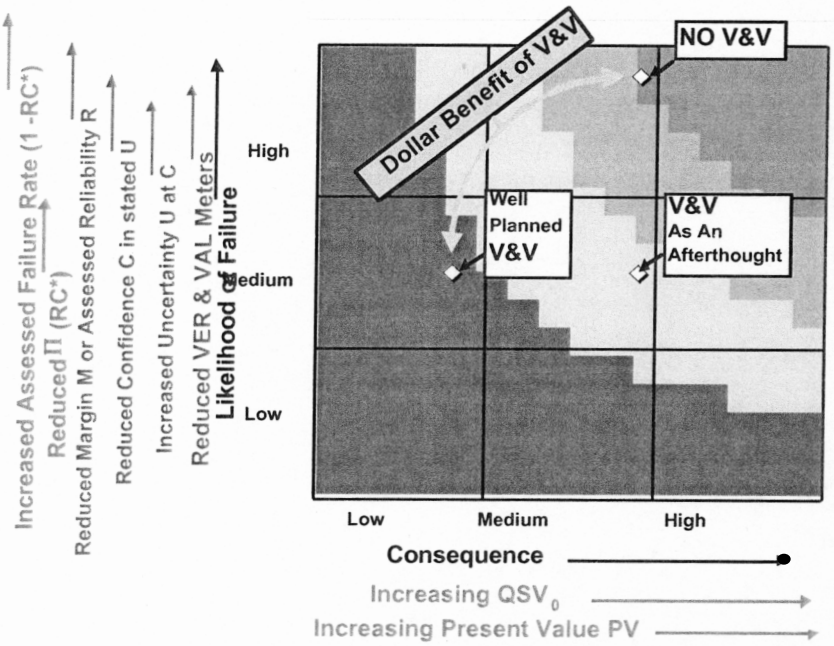
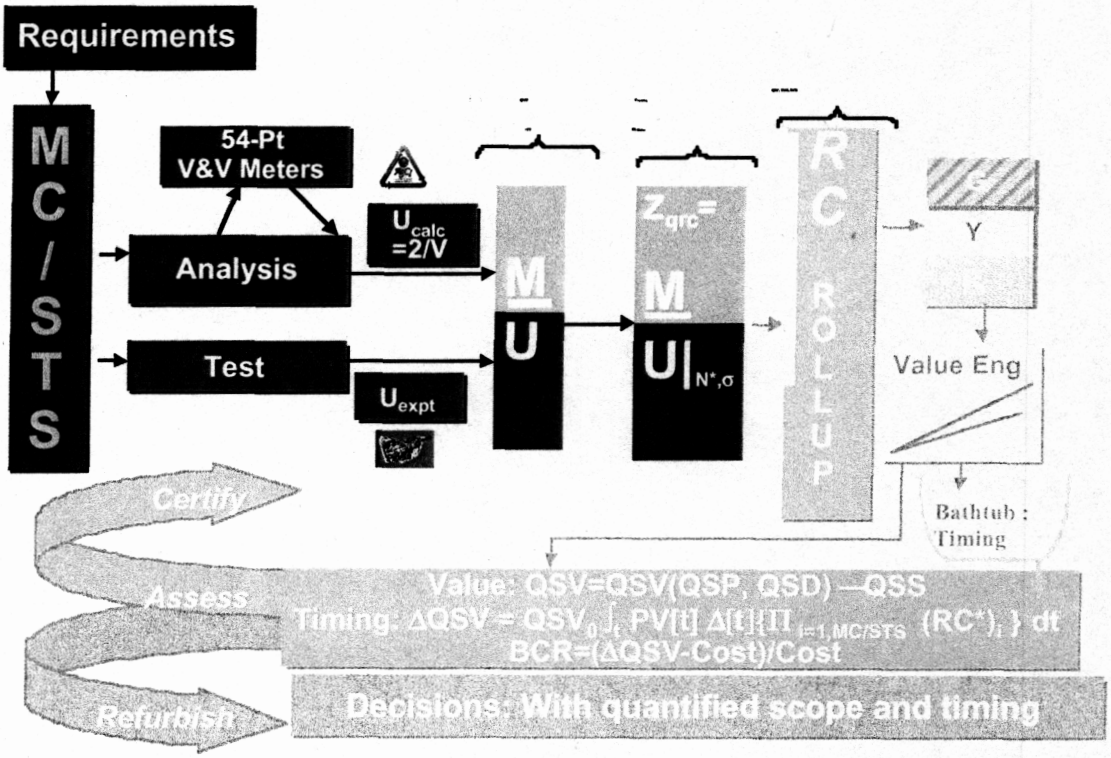


Figure S-2. Dollar Benefit of V&V and Quantitative Certification, expressed as a standard Risk=Likelihood*Consequence Matrix. The analogies are built step by step in the 3-part discussion that follows. Likelihood becomes analogous to assessed (1-QRC); Consequence is expressed in Value Engineering / Earned Value [ie dollars] terms.

1 PART I: V&V definitions, methods, metrics, guidelines, and checklists

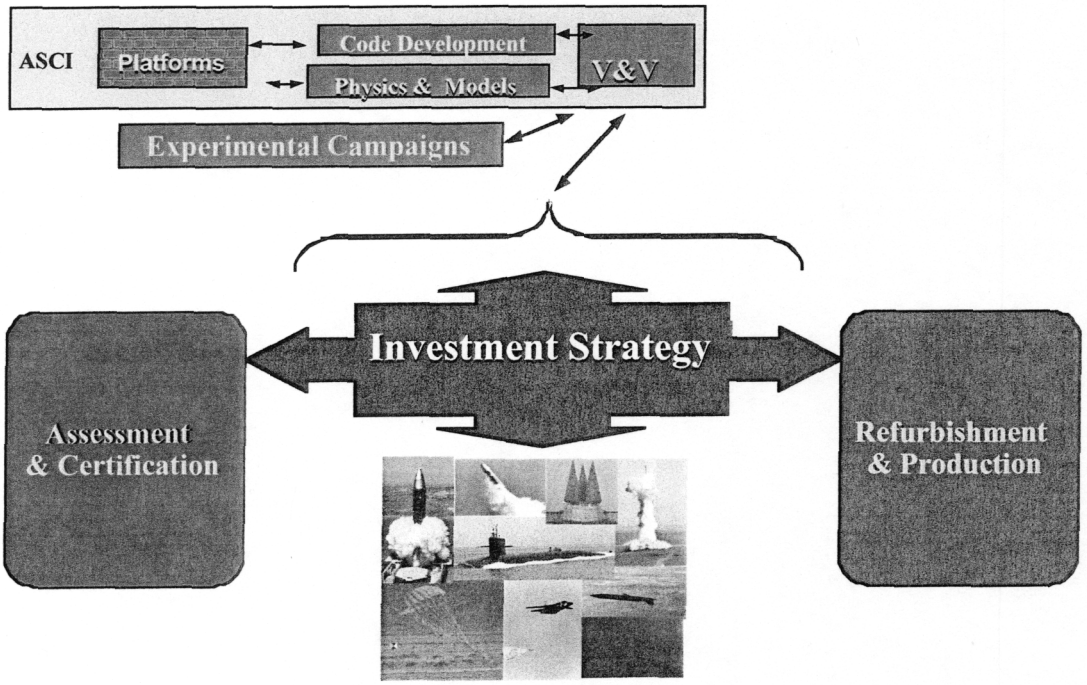


A2

Figure I-1. Flow diagram for Part I: From system Requirements (MC/STS) through V&V, through uncertainty quantification dependent on sample sizes (N*) and equivalent Confidence level e.g. at σ .

1.1 Introduction: V&V, Stockpile Stewardship, and the Decision Process

Figure I-1 shows the highlighted portion of our flow diagram relevant to Part I. The first step is to establish system Requirements. For our work in the nuclear community, these are the MC (Military Characteristics or a top-level system requirements document) and STS (Stockpile to Target Sequence, a logistics and mission document one level deeper). Based on these requirements for the system and its environments, the V&V process begins, leading to validated models with uncertainties at confidence as our product of Part I. We then continue with comments on specific factors that should be considered as part of a major V&V Milestone. In addition, we include a few comments on the mechanisms regarding how we have done this recently, with examples that emphasize our STS work. The example methodologies have been used recently in our nuclear performance arena as well, if one draws the appropriate analogies — not all of those can be discussed coherently in an unclassified setting.



2

Figure I-2: The role of V&V in Stockpile Stewardship today. At the risk of oversimplifying, V&V has to Verify the codes and physics/material algorithms, and Validate the continuum models that use these verified code features.

Figure I-2 shows a simplified layout of stockpile stewardship today: **ASCI and V&V have a key role in stewardship: Stockpile activities, ASCI with Validated Models, and Experimental Campaigns must be integrated for credible, quantified, investment strategy tradeoffs for optimal and cost effective Stockpile Stewardship.** ASCI must leverage the rapidly increasing computing power to enable us to offset our loss in confidence due to lack of testing; and Campaigns provide theory and validation data so that ASCI provides credible results.

A major goal of SBSS has been to enable us to go on into the future and continue to certify the stockpile without nuclear testing. The reality, imposed in part by a shrinking production complex, losses of DoD assets and partnerships, and sometimes by safety and security constraints, is that there are also lot of non-nuclear tests that we either can t or don t do. This makes SBSS a challenge even for non-nuclear STS environments, and makes quality standards for ASCI Validation all the more important.

The entire process can only have the credibility supported by its scientific basis on the ASCI and Campaign side. Certainly not every major action [or inaction] involves ASCI or work with ASCI tools. However, the big decisions do tend to involve ASCI (Modeling & Simulation) tools. Since

ASCI demands over 10% of the Stewardship budget that is not surprising and should even be expected. Stockpile decisions, backed by Validated ASCI analyses, will mean billions of dollars spent. The priority must be carefully determined by the entire stewardship partnership for best-cost effectiveness. And the decisions will either mean claims that we must return to underground testing and new weapon designs and/or production, or that we need not.

So ASCI tools and simulations bear a heavy burden. Platform funding must be adequate because our speed and fidelity and capacity are still too low. Heavy development funding continues to keep pace with faster, bigger platforms. But 7 years into ASCI, V&V has become the doorway to use of the products for stockpile decisions. This is because the credibility of ASCI is not related to the amount of dollars spent or the number of colors on a movie palette. The credibility is provided only by documented, quantified V&V performed by documented, qualified experts in a given area. Much is said about certifying the certifiers . We already have a quality people process — in a discipline-oriented culture of science and engineering, where credibility is built over time and pride in the curriculum vitae of our people. In all of the key certification Program elements at LLNL, we have a documented quality people process comprised of people s resumes, their position, their job description, and their reputation built over years of experience. There is a quality process for our stewardship people, it is documented, and it is quantified. With the large amounts of ASCI investment involved, should we do less to document and quantify the credibility of the ASCI products and simulations used to make recommendations for the stockpile? Similarly, the experimental validation campaigns, and the production complex, should partner with validated ASCI simulations leading to conclusions be backed with quantitative measures of quality.

We have entered an era of unprecedented computing power and speed. However, generating more numbers faster is only of value if the numbers have credibility. Hence the quote, Ultimately, the success of the whole of ASCI will depend on how well we accomplish our goals in V&V (Oden, 2001). There are some significant points about the benefit of V&V that we should note:

Robust, thorough, quantitative V&V:

- Can let us leverage our stockpile job and V&V job.
- Can mitigate the need for excessive Software Quality Assurance
- Our TriLab and LLNL SQA guidelines are built assuming extensive verification, validation, and uncertainty quantification as the other major elements of V&V

We are developing a continuously evolving 19-point VERification and 35-point VALidation checklist, with suggested criteria to consider when performing V&V analyses:

- Including VER and VAL Meter ratings, 0-10
- Including horsepower — elt-steps/msec — or inverse grind time .
- VER and VAL meters are needed for each code feature, each environment, domain, and quantity of interest: We address this need on a graded scale.
- Horsepower measures really need to be expanded to total solution time , including model building, pre and post processing, and model iteration process

These factors are summarized in an overall 10-point summary.

One goal for the use of V&V Meters is that they should reflect our qualitative expert judgment of Verification or Validation status at the component or system level for a given code feature or analysis model. Validation metrics and their relation to uncertainty will never be totally objective, but we have to proceed and evolve.

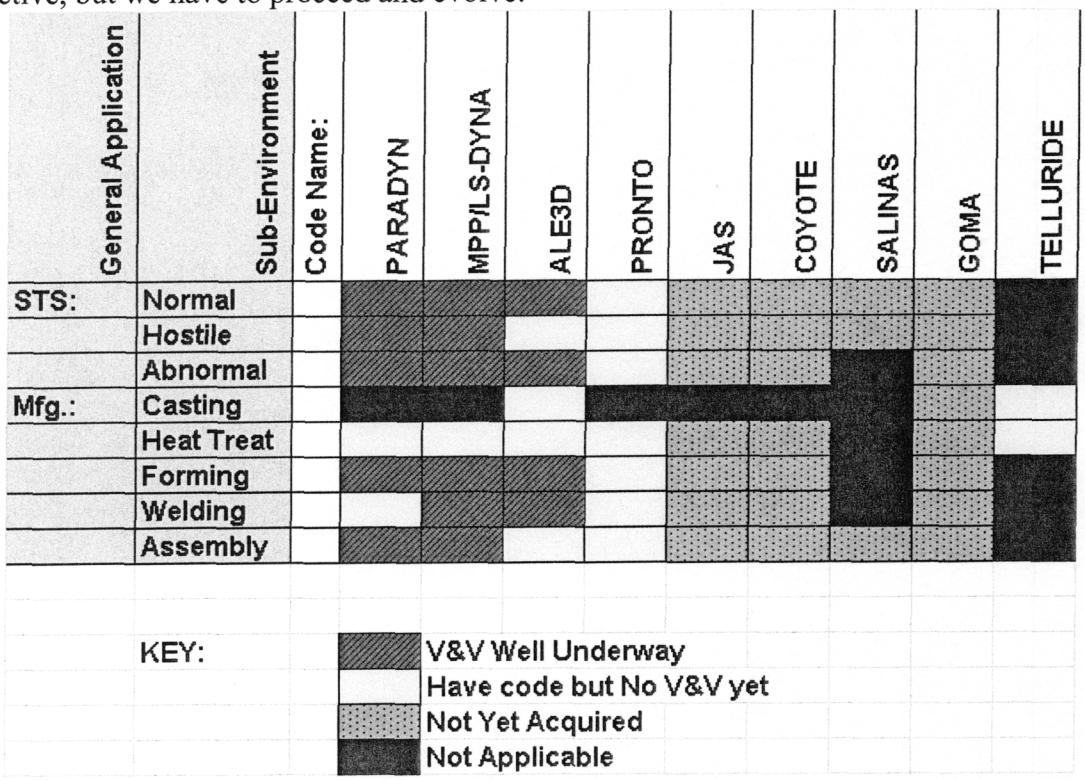


Figure I-3: V&V is a big mission: The scope of the effort and our progress is easier to show in the Engineering / STS environment area than in the weapon physics area. This figure, made circa April 1999, shows our evaluation status of V&V of the TriLab Codes for the Complex code efforts.

It is true that the loss of testing we face is most critical in the nuclear arena, due to the de facto Comprehensive Test Ban Treaty (CTBT) world we live in. Therefore, most of the power and funding of ASCI is keyed toward offsetting our losses in that area. However, a similar effort exists, on a smaller scale, in the weapon engineering arena. We have lost many opportunities for testing in that area as well, and yet are constrained from excursions or changes by the linkages to CTBT constraints just mentioned. Therefore it is important to harness the growing compute power of ASCI in weapon engineering and manufacturing environments as well. Figure I-3 shows the challenge we face in V&V given the large variety of engineering computational tools being developed under ASCI. We cannot simply wait until all these codes are done . A code is never done, and furthermore the stockpile will not wait — we have to make credible decisions now. Therefore, **a graded, but quantified, method for V&V is needed.** A description of our status in developing this quantitative, graded method is the scope of this work.

1.2 Definitions: Verification and Validation:

We begin by describing our favorite V&V definitions and a V&V process that leads to qualitative measures for V&V in Part I. Following the *qualitative* process of Part I enables *quantitative* Validation Statements expressed with Uncertainties (U) at Confidence [C] as described and used in Part II. A compendium of definitions proposed from the community is given in Appendix A. Our preferences include the following:

VERIFICATION:

Verification: (Roache, 1998): Verification ~ solving the equations right.

Verification: (Cafeo and Roache, 2002): Verification of a CODE: The process that determines that the computer code accurately represents the mathematical equations.

Verification: (Cafeo and Roache, 2002): Verification of a CALCULATION: The process that determines that the computer calculation for a particular problem of interest accurately represents the solutions of the mathematical model equations .

These definitions should enable a *quantitative* Verification Statement, such as the following:

> *This material model feature has demonstrated 99.3% accuracy on elastic plastic deviatoric stresses and strains [documentation cited], and it has shown this accuracy in combination with 8 element types with aspect ratios as high as 5 and angles as low as 50 degrees. The model is known not to work well with values of bulk-to-shear modulus higher than 10. The model has shown over 96% accuracy with contact bulk stiffness ratios as high as 1000. We give this code feature an overall qualitative VER rating of about 4-5 out of 10.*

(We will discuss the qualitative 0-10 VER and VAL ratings for V&V Meters below).

Verification assures not only that correct answers can be obtained from codes, but also that *users* can build inputs and obtain those same correct answers. When codes or platforms change, these verification assessments must be repeated. Therefore, verification is an automated and prioritized effort. Verification problems are built into a regression suite that is automatically run as new code versions and platforms appear, and the features needed most often and earliest are verified first and most extensively.

VALIDATION:

Validation: (Roache, 1998): Validation ~ solving the right equations.

Validation: (AIAA, 1998): The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model. *Quoted from (Oberkampf and Trucano, 2002) pg 14.*

Validation: (Cafeo and Roache, 2002): Validated Model: A model that has confidence bounds on the output. A validated model output has the following characteristics:

1. The quantity of interest
2. An estimate of the bias
3. A set of confidence bounds.

A validated model is one where we can make a formal statement after running the model similar to:

> *I am 90% confident that if I build and measure the quantity of interest, that it will fall within the confidence bands (of uncertainty) shown around the model output.*

1.3 General Statement of Methodology:

We view the V&V mission as starting with the code development products (i.e. the codes), supplied to the V&V evaluators, verified to the developers informal satisfaction during development. This implies that the codes will run single element level problems correctly, and that tests to assure that the algorithms show solution convergence as discretization in the spatial, temporal, and interactive domains is refined.

Our V&V methodology is based on a philosophy that emphasizes degrees of independent evaluation of the code capabilities by the code users, in our case by the designers and analysts who are outside the code development groups. Measures of independence of the V&V process are shown in Figure I-4. Our goal is to have the code evaluators outside the code development group, at the third or fourth level of independence from the left. We have chosen this degree of independence to strive for, even though in reality there is a great deal of verification done by the code development groups themselves, and often the code evaluators are inside the same organization as the code developers. This enables our designer and analysts V&V not only the code and model, but themselves, and the platform/operating system, while retaining independence from the development team.

Our V&V methodology includes evaluation of the code simulation results at three layers of comparison as suggested in (AIAA, 1998), as well as a Software Quality Engineering (SQE) Layer 0 that supports the confidence in the software engineering underlying all of the other V&V activities.

All of these layers, including Layer 0 SQE, must be done using a graded scale proportional to the risk, consequence, and investment in the codes and models anticipated.

LAYER 0:

The V&V Program must first address evaluation of the SQE Practices and improvement for each code effort. These evaluations have occurred to some degree throughout the ASCI Program life as well as during legacy code development. We are starting to require more formalization and documentation of these processes, just as we require greater predictive capability from the codes, as the models are used for successively higher risk-consequence events (see Figure I-8).

LAYER 1:

The first layer compares simulation results to known analytic and semi-analytic solutions for specific verification test problems (Verification).

LAYER 2:

The second layer compares simulation results of the ASCI codes against Legacy codes in a regime of interest that is known to be verified and validated to a given degree.

This is one example of the principle of **Validation by Similarity**, where we claim a validated capability because it is sufficiently similar in nature and domain, that in our opinion, we can claim a level of validation without an actual validation process. Code vs. code Validation by Similarity is certainly not rigorous, but it may be quite cost effective, allowing us to save precious resources for higher risk V&V activities. Many have noted the cautions of optimistic results from code-to-code comparison. We contend that code-vs-code validation can *never prove that the pair of code/model combination is right* — but it can prove that *at least one of them is wrong* — and by how much. Often this can be done with orders of magnitude less effort than physical validation testing (especially when such testing involves radioactive or toxic or expensive materials — or all of the above!) In our stewardship of the enduring nuclear stockpile, code-to-code comparisons will never of course equate to validation statements in and of themselves: However, they do offer partial remedies to the following:

- Some tests cannot be done due to treaty, cost, lack of parts, funds, health and safety
- Comparing to legacy data is indeed essential, but often not as convincing as we would like: Our number of data points N* is small for a given system, and the detailed data is usually lacking — and indeed if it had been obtained we would encounter at length the dilemma that you cannot take a measurement without perturbing the result of your measurement
- Code-to-code validation, while risking interdependency, allows comparison of nearly any desired quantity in any desired level of detail — at nil additional cost or fidelity risk
- Legacy vs. modern codes can often be compared — we can make statements about validation compared to the legacy validation level — if we consciously include any new degrees of freedom (K)
- In an era where computation cost is falling and test cost is rising, a strong V&V code-to-code element is compelling from an efficiency standpoint — before proposing expensive tests with scarce resources and competing priorities and liability risks.

LAYER 3:

The third layer compares simulation results of the ASCI codes against legacy and modern Above-Ground Experiment (AGEX) and nuclear data (Validation). The legacy comparisons can be properly termed either **Calibration** (involving tuning of the model), or **Postdictive Validation** (using the model with no more tuning, as if the data were new). The modern AGEX comparisons can be properly termed **Predictive Validation**, if we take advantage of these rare and expensive opportunities to make a prediction before our new tests are run. Of course, the Layer 3 process especially is iterative: We can repeat our Calibration process, or expand it at the expense of the old data remaining for Postdictive Validation. However, in doing so, we raise the **CAL/VAL Ratio, or the ratio of test data we have used to tune our models to the test data we use without tuning**). We can then sharpen our predictions and obtain, with yet more new tests, an improved assessment of Predictive Validation ability.

Of necessity intertwined with the three layers for code comparison within the V&V Program, Uncertainty Quantification (UQ) is essential as an integral focus area for evaluating the many

sources of error in simulation of stockpile devices, including uncertainties from the physical database, AGEX and nuclear databases, the code algorithms, software implementation, and physical and material models. Rather than view UQ as a separate activity with separate milestones, we feel it is compelling, due to our preferred definitions of Validation above, to view V&V with ensemble computing (Oberkampf and Trucano, 2002), where the UQ process is integral with quantitative validation statements that must accompany any claims of a level of validation.

1.3.1 Methodology and Independence Metrics

This section describes our V&V effort as it provides coordination and integration of the Verification and Validation of the ASCI code product developments with the needs and activities of the product design, analysis, assessment, qualification, and certification teams. V&V is essential to the credibility of the tremendous effort being put into the ASCI Code Team projects. This is because it is the V&V process, performed *independently of the Code Teams*, as depicted in Figure I-4, which provides the most direct evaluation of success and need for future work of the ASCI Code Team products.

Code development efforts are not described within the purview of V&V, since that activity occurs elsewhere in ASCI and in the core efforts. We view the V&V mission as starting with the code development products [ie the codes], supplied to the V&V teams, **verified to the developers informal satisfaction during development** and so in a condition such that the codes will run one-element level problems correctly, and tests to assure that the algorithms show solution convergence as discretization in the spatial, temporal, and iterative domains is refined.

Ideally V&V would be a one-pass process, where codes and code features would be passed to the independent design teams, V&V d, and cleared for production use. We postulate this for simplicity — but of course in reality V&V is an iterative process with the code teams themselves, as both the developers and the users verify both the codes and the familiarity with them. We enable that iterative process by running combinations of the code-team supplied regression or example or verification suite — *but for formal V&V even some of these verification runs are made by the V&V team — the people outside the code development team — who are also the eventual team of designer / analysts.* In this way, we assure a Validation product capability delivered as declared, because it is the eventual designer/analyst users making the declaration i.e. the customer, and not the supplier. This is not a slight to the code developers whatsoever — most of those in V&V have been developers at one time and recall the dilemma that when one develops a piece of code, it is common that the developer can successfully use the capability — but other users cannot. ***Verification and Validation cannot be formally declared until a user base (the designer/analyst groups) declare it so.***

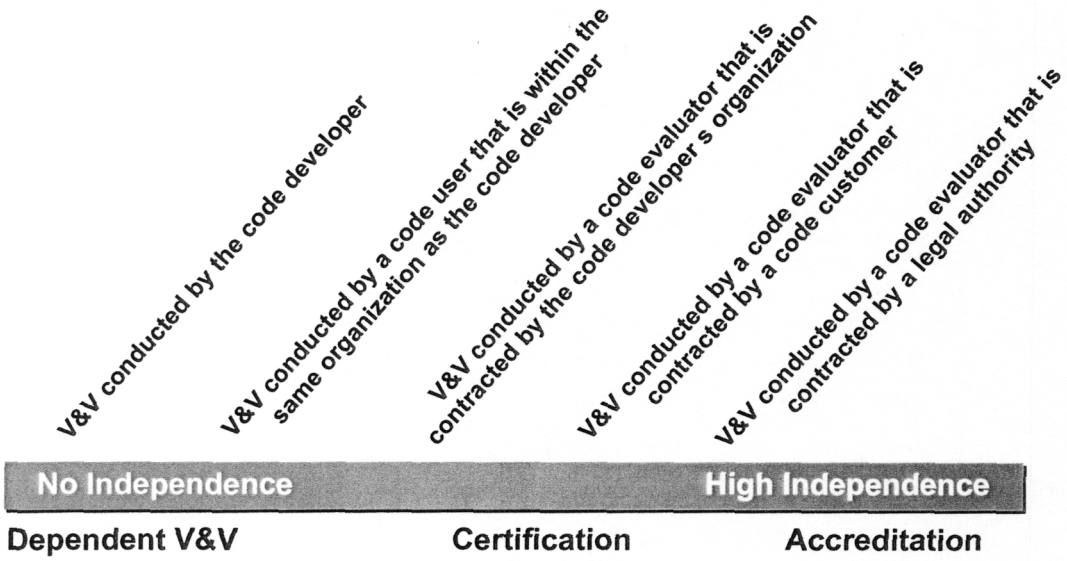


Figure I-4. Measures of independence of the V&V process (Oberkampf, 2001). Of the 5 descriptors, our goal is to be about 4th from the left. This enables our designer / analysts to get at least two for one in their efforts, because they are V&V-ing not only the code and model, but themselves, and the platform/operating system, while retaining independence from the development team. Metrics for V&V can be the same whether independence is stressed or not; but since the goal is to provide a V&V level claimed in design and assessment, the claim can only be credibly quantified when the V&V is performed within the whole system; platform, code, model, and [independent] user base.

Regarding the candidate codes, in the nuclear design arena these are by their very nature codes developed internally, or certainly within the design lab community. For many of the engineering analyses, it would be an expensive omission not to consider as well the use [or coupling] of commercial code products, if they are able to perform adequately, and in the ASCI scenario perform on the ASCI platform machines. In fact, it is our experience that the *world of commercial vendors and user conferences provides an excellent example of the code team vendor and V&V/designer customer relationship. That is, the most credible user meetings involve V&V testimonials presented by the customer [companies that buy or lease the commercial code and use it — and present their feelings about V&V and beyond.* This adds inherent credibility to the code and developers, and confidence to potential users and customers. V&V by the code team statements and presentations are helpful and informative, but do not constitute a true V&V evaluation for products intended for design, assessment, and qualification use. ***During our formal V&V Milestone presentations and evaluations, our methodology dictates that this formal V&V will be done and presented by the designer/analyst users, not the code teams themselves.***

Verification is ideally the first step for the V&V team. There are time and budget pressures to do minimal verification of the platform/code/user/model system; but it is essential to address this topic for at least a few key or new code/model features as they are used. Verification will consist in large part of fairly simple problems at the material or component level, in other words not

entire weapons. Examples of verification problems may include tension tests, closed-form heat transfer and momentum transfer analyses, modal analyses of materials, etc. In particular, verification of a wide spectrum of material models is needed to assure they perform as stated in the relevant manuals and that input decks will match relevant test data.

Validation [of the codes that pass verification on chosen classes of problems] will involve application of the verified codes to component-level and full-up warhead analyses. Examples of weapon engineering component-level validation problems may include modal analyses of partial assemblies, neutron heating of components, thermal reaction behavior of High Explosive pieces and simple assemblies, and impact safety behavior of buildups of energetic and other materials. These problems can be used [most in an unclassified setting] for benchmark competitions among codes, competing with each other and with real test data to obtain quantitatively credible results in a pre-selected timeframe. This concept has proven to work well in the industrial and academic settings. Legacy codes, on legacy and in some cases ASCI platforms, will be compared with ASCI codes, that is, those codes capable of producing quantitative answers to the STS problems and making use of the ASCI parallel platforms in the process. Eventually, during V&V, the speed and utility of all available code suites must be considered in light of the tradeoffs between product performance and cost to make or buy a given code capability. Speed may involve purely run time as measured in LLNL s Muscle Code Shootout (MCS) series (Sam et al, 2001), or scalability, or Teraflop efficiency [which is probably better addressed outside formal V&V by the code teams], and last but not least, ease and speed of setup, modification, and viewing of results. Make or buy applies more directly to the engineering aspects of our work — and in some cases to pre and post processing, etc. It is understood that purchase of *nuclear physics* simulation capability is not generally feasible or desirable.

1.3.2 Metrics: VERIFICATION and VALIDATION

Verification and Validation criteria documents applicable to our specific work have been fairly terse and qualitative until recently. Some aspects are addressed in LLNL s Engineering Policies and Procedures document (LLNL-ENG, 2001) section on design. We will note this reference below as one of the 35-point criteria, to be applied as appropriate. However, like any of the points in the checklists, the rigor of adherence depends on the cost/benefit/risk associated with the particular analysis. A more comprehensive but still *qualitative* set of Validation Metrics is described in the next section. A total of 35 Metrics for a Validation are depicted, and this document is our latest effort at these metrics. The goal of these 35 Metrics are for completeness — not for acceptability — of the model for design or assessment use and conclusions.

1.4 Verification and Validation Meters and Checklists

We can envision (at least) 4 methods for expressing the pedigree of V&V for a given code feature or model. Consider the following 4 methods:

A. Use of words like Fully Validated , Unvalidated , Validated Code , etc.

- *(We do not recommend this method.)*

B. Use of a simple one-number scale e.g. 0-1, 0-5, or 0-10 as we suggest in our following description of our VER and VAL Meters. The VER and VAL Meters described (and shown below as introduced at the January 2001 ASCI PI Meeting) provide some measure of acceptability and caution — the meters are quantitative but they are subjectively set.

- *(Part I of this work.)*

C. A multi-point checklist of thought, procedure, and documentation — including partial completion of (D) as well.

- *(Part I of this work.)*

D. Quantitative error and uncertainty bounds at confidence (we will show in Part II that even a simple example of this is complicated).

- *(Part II of this work)*

What, of (A-D), should be a minimum set? How can we express the level of V&V most efficiently and yet adequately, and as quantitatively as possible yet in a compact form? Full use of method (C) or (D) may not be warranted. However, use of the descriptors in (A) conveys very little meaning. We suggest that as a minimum, method (B) — a simple scale — can convey an expert opinion rating of the V&V level of a simulation in a condensed form. Certainly this is not a complete method, yet it conveys a more accurate picture than (A), and is thought provoking enough to lead us to pursue methods (C) and (D).

The goal of ***Verification*** is to assure not only that correct answers can be had from the ASCI codes, but that the end users can build the inputs, hit the enter key, and get those correct answers themselves. Of course, it is an endless task to verify every feature in conjunction with every model it might be used in — let alone to repeat this as new versions and platforms appear. Therefore, Verification is a prioritized effort — the features needed the most often and the soonest are verified first and most extensively. In addition, Verification is an automated effort — verification problems are built, over time, into a regression suite that is automatically run as new versions and platforms appear. To this end, in a *qualitative* sense, we describe below a set of 19 Verification metrics for ASCI/STS code products. In a quantitative sense, we must still get back to what is acceptable as a Verification answer. For problem classes with stable solutions, we can usually make one metric that of *Convergence*, or at least a demonstrated approach to it. Mesh refinement studies are in essence Finite Elements 101 . It is true that we may not end up using the verified code feature in a fully verified state as we move to Validation and design work. However, if we have to back off on mesh fidelity, having taken the problem to convergence tells

us how far we have backed off. For a few examples that illustrate the need and use for metrics in Verification and Validation, see (Logan and Nitta, 2001), and for more detail see (Trucano et al, 2001). A desired product of verification is a body of analysis and documentation to support a *quantitative* Verification Statement such as the one given above in the DEFINITIONS subsection.

As we will show and see in the discussion below, we can quantify metrics but to have much meaning, and to specify an acceptance level, it is usually necessary to also answer how much difference it will make if we are inside, at, near, just outside, or way outside a given level. ***We assert that we can and should use methods consistent with a statistical quantification of reliability at confidence,*** *even though we cannot commonly use the frequentist approach but must rely on expert system, fuzzy logic, or Bayesian sets as described in (Oberkampf and Trucano, 2002) and (Booker et al, 2002).* Methods described in (Logan, 2001a) and (Logan, 2001b), and in Part II of this work, can eventually help us with acceptance levels for Validation. Uncertainties and Sensitivities in the simulations can be quantitatively related to *assessed levels* of Confidence and Reliability numbers for the stockpile during this era of no nuclear testing.

The goal of ***Validation*** is to take the ASCI products to a quantified state where they can be evaluated for acceptance for assessment and certification work at a demonstrated level of uncertainty and hence confidence. Our analysis timelines span many orders of magnitude — sometimes a hero fidelity simulation is best, but may require 2 years to build and run. Other times, an answer just incrementally better than back-of-the-envelope is needed, but that answer is needed in 2 hours.

It is important to understand this time critical aspect of stockpile work. That is, in an ideal world of stewardship with floating deadlines , we might postulate that no ASCI codes will be used for stockpile models until the codes are completely verified and until the models in those stockpile STS (system environment) regimes are completely validated . Even if we could wait that long, the word completely is a boundless task.

We might then open with a Validation Metric that says, any run is acceptable . This is true enough — but the next and obvious question is, acceptable for what purpose? At this point, we feel that there are precious few instances where a pass/fail metric can be used for validation.

Rather, balance is the key — balance of funding, timelines, priority, and credibility. Tradeoffs are always necessary between these. We simply must know — and express — what tradeoffs we have accepted.

Verification and Validation criteria documents applicable to our specific work have been fairly terse and qualitative until recently. Some aspects are addressed in LLNL s Engineering Policies and Procedures document (LLNL-ENG, 2001) section on design. However, like any of the points in our guidance checklists, the rigor of adherence depends on the cost/benefit/risk associated with the particular analysis. A more comprehensive but still *qualitative* set of Verification and Validation Metrics is described in the next section. A total of 19 Metrics for Verification and 35 Metrics for a Validation are depicted, and this document is our latest effort at these metrics. The goal of these Metrics is for completeness — not for acceptability or mandate — of the model for design or assessment use and conclusions.

1.4.1 The Qualitative VER and VAL Meters

One basic dilemma is to express the V&V pedigree of a simulation result or conclusion, in a way that goes beyond yes or no , but remains a fairly simple quantitative expression of the V&V status of a simulation. Simplicity is essential to the decision making process, because calculation results and movies are often shown at fast paced meetings where numerous topics are covered in a few hours — with only cursory detail and never enough time for the audience to evaluate the credibility of the detail being shown. Typically only a few seconds are available to describe the pedigree of a given part of the simulation. And yet, impressions are formed at such meetings and can lead to misunderstandings and regrettable decisions, unless some kind of graded scale V&V measure is used.

Often, the highly trained and skilled finite element analyst/designer is unfamiliar to this high level review setting. We develop an impression during our education and professional society interactions that the nature of the game is to spend 20 minutes or so to make and justify a couple technical points. In a rigorous academic or professional environment, this focused presentation of viewgraphs is accompanied by a pre-published work of prose that the audience has to augment the verbal and visual presentation. Typically, terse visual aids like the VER and VAL meters described herein are not needed or desired; they are implicit to a qualified audience that has the chance to both read and listen to a detailed explanation and justification of the methods and conclusions of the work. However, high level meetings that pervade our business and indeed most corporate entities do not have this luxury of time and text. Often they do not even have the breadth and depth of background in a given area. This is no slight to the high level audience; nobody can be that deep of an expert in every area. These review and decision groups see hundreds of viewgraphs and movies fly across the screen during the day. Nevertheless, these high level audiences assume and deserve presentations that have some measure of quality with them. Ideally, every slide they see would be Thesis-Defense quality. In a world of procedural rigor and constraints on time and personnel coupled with challenging, sometimes arbitrary schedules for milestones and meetings and decisions, the concept of Thesis-Defense Quality for each analysis is not credible.

There is nothing wrong with using a lower quality or conceptual analysis [e.g. Hollywood Movie] to make a point or point out an area of risk. However, to avoid having the audience take such examples with verbatim precision, a VER and VAL meter or equivalent *as a minimum* should be used. The ASCI V&V Program, with analyses done in an independent environment by experienced designer/analysts, is ready made to produce and support the setting of such meters.

But, the meters are of course relevant for more than just a quick indicator at fast paced review meetings. The Meter readings (or any such rolled-up number rating for V&V) can, in addition:

- Enhance the capability of "designer-centric" or expert judgment based V&V;
- Firm up the credibility of conclusions that are drawn using any historic methodology;
- Make more scientific the V&V process;
- Make more scientific the decision process;
- Provide fundamentals for rational discourse on this subject;
- Provide a rational basis for common understanding and expression of V&V level.
- Provide an expression of *relative* information and level regarding V&V

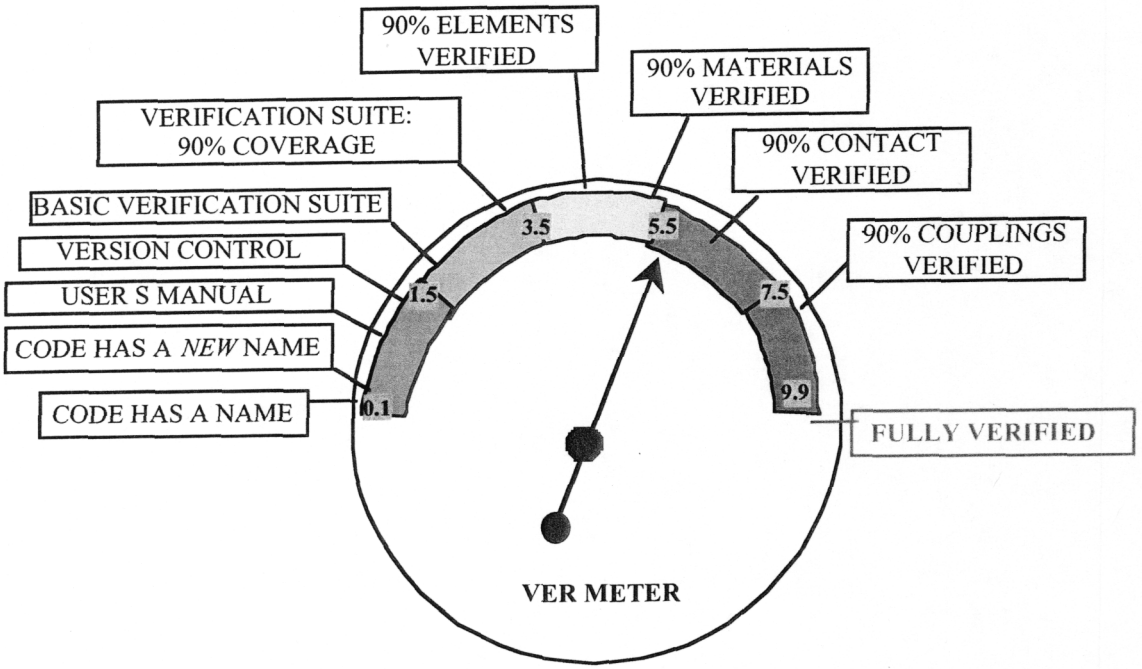


Figure I-5: The VER Meter: One or more needed for each code feature

The concept in Figure I-5 of a Verification Meter , that reads 0-10, is simple but it should be fairly clear in intent. Obviously such a Meter is still subjective, still qualitative in how we set the meter. However, it is likely that a consensus of veterans in a given area of mechanics or physics will not be too different in placement of the Meter Reading, even given the simple narrative on the slide, but especially with general criteria for guidance. More is needed but we start with this meter and improve it as we refine the V&V process and guidance. Ideally, several such meters would appear next to **each feature** in a code manual (for code verification). Several would be needed to cover the **combinations** of options within e.g. a material model, and to cover use of the code feature alone or in coupled mechanics settings. Use of the VER meter is also needed for Solution Verification, or Model Verification as noted above. This would typically indicate that even though the VER Meters in the code manual might all read 9 , we may know (and should know!) that we have chosen, for expediency, to use a solution model in space and/or time and/or convergence tolerance that gives the accuracy of our solution only a 5 or 6 rating. We should of course strive for more than a 1-10 scale. More desirable are quantitative Verification Statements such as the one given above in the DEFINITIONS subsection.

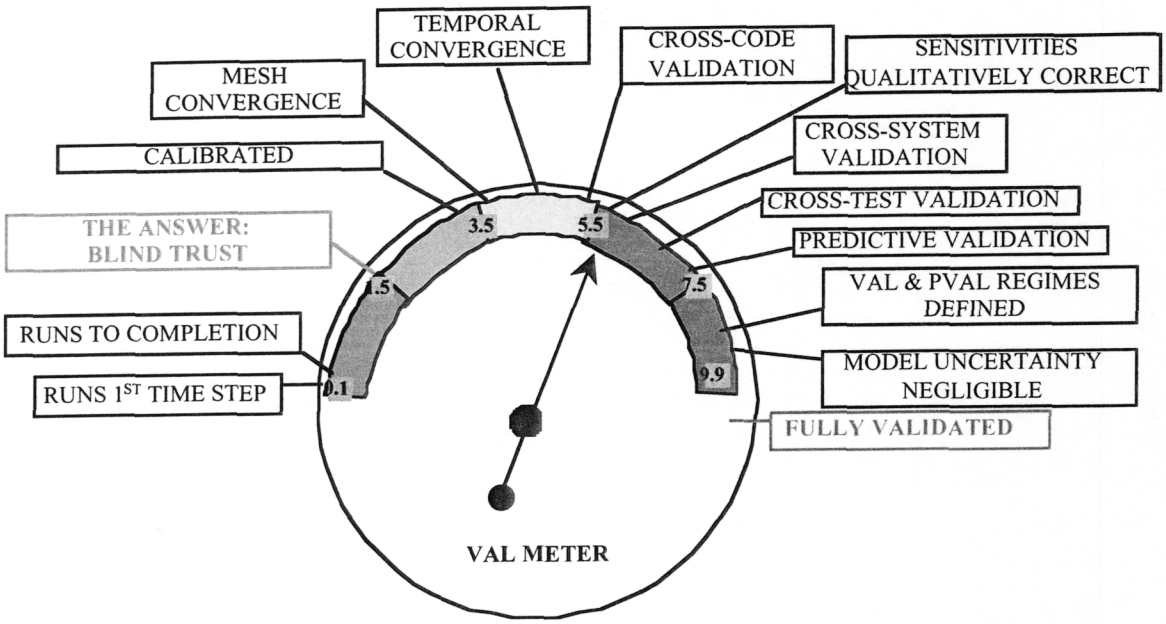


Figure I-6: The VAL Meter

The VAL Meter reading in Figure I-6 is as subjective as the VER Meter of the previous figure. However, it is again more likely than one might imagine that the placement of the meter and consistency across analyses, systems, and codes, is fairly easy to do in a consistent way. Given consistent qualitative criteria, we have found that knowledgeable observers tend to come to similar conclusions within the V&V community. The V&V process is much deeper and more quantitative than any single summary number can depict. But, the meters go beyond a yes/no V&V statement for communicating fidelity quickly.

To begin to take us beyond the simplistic meter summary, we have tried to capture four of the key levels of Validation contained in the following nomenclature:

DEMO: Demonstration, i.e. run to completion
CAL: Calibration, i.e. satisfactory agreement with accepted legacy metrics
VAL: Validation, agreement to predefined metrics, without [further] calibration
PVAL: Predictive Validation, i.e. agreement of a pre-test prediction with the test result within the pre-test confidence bounds of uncertainty established through V&V

For further refinement of the VAL Meter, we might imagine covering issues such as those in the 35 point checklist below. In fact, we might someday choose to assign in our mind a certain number of points to each of the criteria below. Those point scores would then ideally go a layer deeper still, e.g. they would express agree in terms of a statistical validation statement like the ones given in our definitions section, and express converged in terms of quantified

discretization and solution error studies. The difficulty and lengthy nature of the answers to these questions is why V&V is a continual effort that needs a graded scale (e.g. Meter Readings) to depict progress in any given realm. But in the meantime we cannot certify on the basis of, here s our Certification Movie and maybe in 5 years we ll tell you its VAL Meter reading . We have to make decisions based on such evaluations *now*. And in the CTBT era of no nuclear testing and precious little system-level testing of any kind, we should take advantage of the luxury of a large — unprecedented — ASCI effort and *as a beginning* use measures like the VAL Meter quantification that we may not have used in the past.

Uncertainty Quantification (UQ):
Since we endorse the concept that validation should be a quantitative statement, it is essential to consider validation and Uncertainty Quantification (UQ) as a set. As we describe in Part II of this work, uncertainty is inseparable from the confidence at which it is stated. Therefore, validation, UQ, and confidence become inseparable.

We now introduce definitions of the components of uncertainty we will be discussing qualitatively in the VER and VAL criteria, which we will later address quantitatively in Part II of this paper. Consider, after (Oberkampf and Trucano, 2002), these four components of Error or Uncertainty , as defined here as the differences between quantities q of interest:

For a measurement of the quantity q of interest, let (ibid.)

$$\Delta = (q_{nature}-q_{exp})+(q_{exp}-q_{exact})+(q_{exact}-q_{discrete}) \quad [2a]$$

or alternately (ibid.):

$$\Delta = E_1+ E_2+ E_3+ E_4 \quad [2b]$$

We might consider adding to Oberkampf and Trucano s set an E_0, and characterizing the E_i loosely with words, so that now:

$\mathbf{E_0=(q_{nature}-q_{nature})}$: Variability / Aleatory Uncertainty [3a]

$\mathbf{E_1=(q_{nature}-q_{exp})}$: Uncertainty / Epistemic Uncertainty [3b]

$\mathbf{E_2=(q_{exp}-q_{exact})}$: Error, in model [3c]

$\mathbf{E_3=(q_{exact}-q_{h,t>0})}$: Error, due to formulation or weak form [3d]

$\mathbf{E_4=(q_{h,t>0}- q_{h,t,I,C})}$: Error, due to discretization or solution error [3e]

The above E_i are a useful notional linear combination of error / uncertainty contributions in a validated analysis. Our quantitative method as detailed in Part II will require that we express variability, uncertainty, and error as the generalized U_{Ci}, which becomes an input to the Reliability at Confidence method in Part II. To generate each uncertainty term U_{Ci} on an

environmental condition (subscript C) due to a model contribution (subscript i), we combine *independent* uncertainty contributors (subscript j on the Parameter uncertainty or change U_{Pj} or Δ_{Pj}) and the sensitivity S_{Cij} to them as:

$$U_{Ci}=RSS(S_{Cij}U_{Pj}) \qquad [3f]$$

The sources, numerical values, and nature [i.e. assumed or known form of PDF, Probability Distribution Function] for all these notional linear combination E_i and the more rigorously combined Eqn. 3f should be stated as part of validation with uncertainty quantification. That is, enough information should be provided as part of the Validation / UQ to enable the QRC methods of Part II of this work to be accomplished quantitatively.

Since, in the opening of Part III of this work, the VER and VAL Meters will appear again, let us once again stress their intent, limitations, and ideal use. The methods described in Part II as Certification Methodology; that is assessed Reliability R at Confidence C, depend on Uncertainty Quantification. This is intimately tied to Validation Level. Linkage between Uncertainty and Validation Level have been proposed by others as well, such as the one shown in Figure I-7, taken from (Oberkampf and Trucano, 2000). Overlaid on their Validation Metric is the simple quantity V=2/U; the curves are similar. Obviously there is no true closed-form relation between V and U — but somehow we must link one to the other. This is because, as we just discussed, we must quantify V&V to show progress in ASCI, and we must quantify U to proceed with the QRC into QSV methodology of closure. In cases where we cannot rigorously show a path enabling statistical statements about U , we suggest for consideration *as a screening tool only* the use of U~2/V , where V is the VER/VAL Meter reading — set as merely a composite of qualitative expert judgment - of the model of the environment of interest. This is not the ideal process, but we must evaluate these terms constantly; we cannot wait for a statistical statement of validation to proceed. In fact, the simple method of setting U equal to an expert-opinion chosen V may help us focus our near term investment! Naturally, the only truly credible way to use the VAL meter is to use the form V=2/U (or perhaps V=n/U where n=number of sigma-*equivalents* of confidence). The uncertainty U should be determined by the method described in Part II of this work, or some similar, rigorous way. In this work, we will not even enter into the realm of E_3 and E_4 (weak form, approximation, and discretization error and uncertainty) in this work. These terms are the focus of many other extensive works by other authors in the V&V and finite element method field.

Let us stress that this simple sounding scale or meter concept for Validation is not unique to our thinking. At least two other works have used such a scale to relate in a summary way a Validation Rating to the overall quantification of a code/model for given assessments. These now include:

- The V=0-to-1 scale (Oberkampf and Trucano, 2000) reproduced in Fig. I-7
- A verbally quantified V=0-to-5 scale (Trucano et al, 2002)
- Our own V=0-to-10 scale working in a similar way.

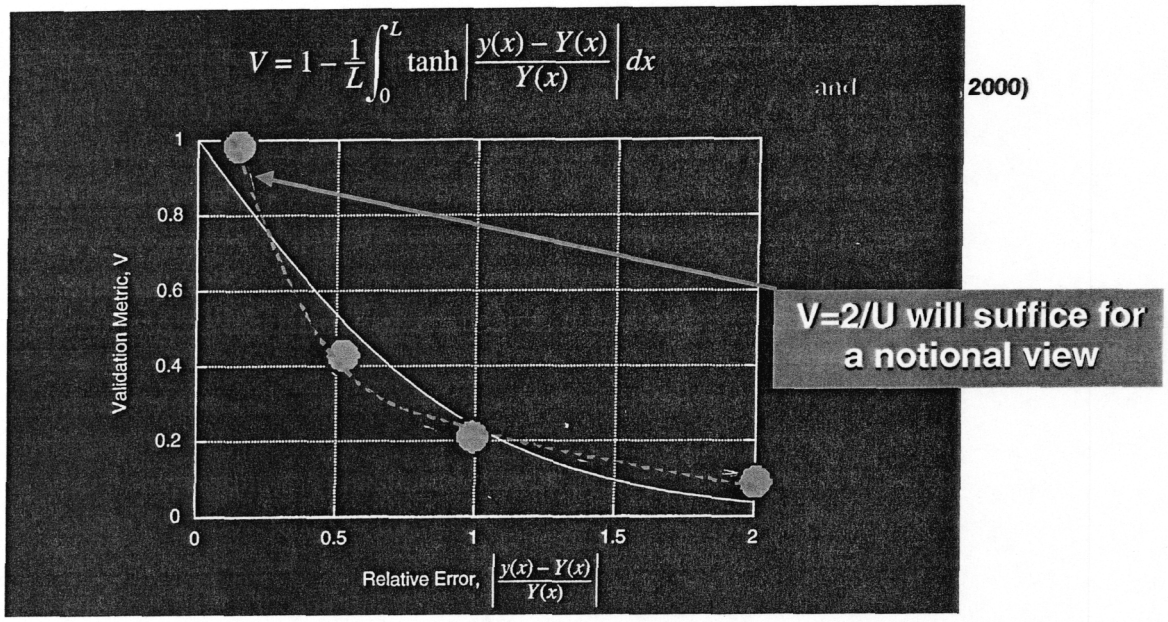


Figure I-7. VAL=0-to-10 Meter Readings are not a concept unique to our thinking! For example: Validation metrics proposed to relate V to U, giving a VAL=0-to-1 scale. V=2/U drawn over (Oberkampf and Trucano, 2000). Both are useful and more valuable than either no indicator at all or an oxymoron like fully validated or validated code. Both expressions for V should be used with great judgment and caution.

1.4.2 10 (or 30) steps to a V&V process: Summary Level Considerations

Our 84 point V&V process can be summarized in 10 steps 1-10, with a 30-step expanded procedure in detail counting the indices e.g. a-h. These reference embedded detail like the 19 VER steps, 35 VAL steps, and lead to various methods to determine and use acceptance metrics. The 10-step condensed form summary is as follows:

1. A Program Plan should exist with scope and timeline to balance the ability to build, verify, and validate code and model capability against the assessment needs of the product line.
2. A Code Capability Plan (expanded version of Figure I-8) should exist, with a simple method to track the V&V status of each capability. Figure I-8 is only a top-level matrix — more detail is needed.
3. A Risk Assessment methodology should exist and be documented. For example, in the area of engineering mechanics, a major part of our risk assessment mitigation is the ability to use multiple codes for any given analysis or planned analysis.
4. Verification (code feature e.g. Figure I-8) and Validation (prioritized system requirements) listings should exist and be prioritized. For Validation, an example of such a target list

would be Figure II-12 The Matrix of R_i and C_i terms for all i environments in our Requirements, shown in Part II of this work. The Validation list is of necessity tied to the prioritized product line assessment and certification needs and timing.

5. The long term plan from code features through V&V to assessment capability should address the sequence of SQE (Software Quality Engineering), VER (Verification), CAL (Calibration), and VAL (Validation):

   a. SQE: Software Quality Engineering: The software quality engineering practices may be tailored for each individual code, but should conform to a standard accepted by the developers and users organization. There are various Software Quality Engineering documents and standards in the community. The ones we use include the recent TriLab ASCI/V&V/SQE [Hodges et al, 2001], and the LLNL ASCI/V&V/SQE supplement (Storch et al, 2001). Both of these go into more detail than either the general references to SQA in (Levine and Twining, 1995) or even in our own Program (Miller, 1995) and Division (Clough, 1997) Quality policy documents.

   b. VER_{code}: Code Verification: In the maximum state of temporal, spatial, and iterative convergence achievable, we address the remaining error in the answers provided by the code for the feature being verified.

   c. VER_{soln}: Solution Verification: We assess the components of model error change as a function of discretization refinement in the temporal, spatial, and iterative domains. Code and model speed (Horsepower =element-steps per millisecond or inverse grind time) should be reported as it is key to determining tradeoffs of platform usage, time to solution, and quality of solution.

   d. CAL: Calibration of a Model [shows existence of a fit but not uniqueness]

   e. VAL_{cvc}: (Risk Mitigation) Validation, Code Vs. Code (CVC). This method is often time and cost effective but fraught with dangers of misinterpretation and misuse.

   f. VAL_{suv}: (Sensitivity and Uncertainty) Validation, with Sensitivity plus Uncertainty plus Variability (SUV). This may be appropriate when for example only one integral test is available. The uncertainty bounds obtained may be quite wide, especially when several terms are rolled up (multiplied) in succession. In this VAL_{suv} method we use model derivatives (tangent or secant) to represent the sensitivities S_{Cij}. We then multiply these computed S_{ij} by what may well be large and estimated material, environment, tolerancing, or other parameter uncertainties U_{Pj}. Adequate statistical quantities of test or measurement data on the U_{Pj} will help tighten our uncertainty bounds to meaningful levels on rolling up several terms; of course this information about each U_{Pj} must be obtained at a cost justified by its reduction in a given U_{Pj} and the importance of that U_{Pj} in total system performance. This importance is determined by knowing the values of the sensitivities S_{Cij}. We must therefore have some verification and experimental

component or modular validation [see the step below and (AIAA, 1998) on Phased V&V] to V&V the model sensitivities S_{ij}. This helps mitigate the interpolation and extrapolation dangers depicted in Figure I-8. These derivatives S_{ij} form the basis of our integral i^{th} environment uncertainty estimates for this method. For example, for the ith environment, our Uncertainty in Capability of the system may be:

$$U_{Ci}=RSS(S_{Cij}U_{Pj}) \qquad [3f]$$

with RSS=Root-Sum-Squares as appropriate. As with Verification, reporting model speed (e.g. Horsepower =element-steps per millisecond or inverse grind time, total model run time, and total user time) is key to determining tradeoffs of platform usage, time to solution, and quality of solution.

g. $VAL_{\underline{mlv}}$: Validation at the integral level, across the rolled up set of environments of application, can be done using a Maximum Likelihood Validation (mlv) fit to the integral data. In other words, if we are fortunate to have several integral tests, the demand that our model match over all of them will more readily quantify tighter uncertainty bounds. In other words, it helps us solve the dilemma that we usually do not have enough data (or model validation) to do VAL_{suv} without the subsequent rollup of terms leading to huge uncertainty bounds. And yet, with an adequate number of integral tests, we have a body of evidence that says that the uncertainties are not so boundless as our (incomplete) VAL_{suv} would contend. VAL_{mlv} is a way of showing our likely uncertainty embedded in the integral model vs. the integral data. The Validation, if done cross-test and cross-system (several systems with a similar mission) will be more robust and with tighter quantified uncertainty bounds.

h. $VAL_{\underline{pre}}$: Validation using Prediction (PRE). In principle we can predict tests that have already been done if we use a calibrated, validated model, with no further degrees of freedom adjusted, and then predict one additional pre-existing test N=N+1. However, this will always leave some doubt as to a subconscious bias of our fitting process, since we may have been influenced by that existing data even though it was not directly used to develop our validated model fit. A measure of VAL_{pre} can be obtained by using a low CAL/VAL ratio in the VAL_{mlv} process, so that if we do no additional tuning after our VAL_{mlv} fit to part of the data, we can predict (really post-predict) the rest of our existing data and see how good our predictive fit would be for the data we have.

We will discuss methods to quantify our model uncertainty and predictive confidence in Part II. Performing this process, that is, estimating a predictive confidence bound and then seeing if subsequent validation tests fall within our predicted bounds, requires extrapolation outside the V&V Cloud shown in Figure I-9, into a region not yet validated over the quantity of interest. If we have used even a few model degrees of freedom to fit the data for a Maximum Likelihood

Validation fit (VAL_{mlv}), we should then ask: How well would our model fit the *next* data point?

This is analogous to modeling a cross-country drive from say Washington, DC to San Francisco for example. If we have any model degrees-of-freedom (DOF) to adjust, we can choose to match the elapsed time to say Salt Lake City exactly. This should not inspire high confidence in our model predicted elapsed time to San Francisco. To quantify the confidence in our model-predicted elapsed time from Salt Lake City to San Francisco, we need to establish confidence bounds on our model sensitivity to elapsed time over distance, but also altitude, grade, weather, etc. The exact match to the elapsed time to Salt Lake City, if any DOF s were adjusted to get it, tells us nothing in itself about confidence we should have in the Salt Lake City to San Francisco elapsed time.

An automotive crashworthiness design study provides a similar example within the realm of our concern of nonlinear finite-element codes and models. During a design study of an aluminum spaceframe vehicle for crashworthiness (Logan et al, 1995), two types of extrapolations were performed. In both instances, lower tier (component) validation data was available (Logan et al, 1995) and (Logan and McMichael, 1995), but behavior at the system level required extrapolation. One of the extrapolations contained some measure of *system level* Validation by Similarity, and the other contained no *system level* validation or estimate of predictive confidence. (Given the minimal cost of this preliminary design study, such unvalidated extrapolations were acceptable, but we use them here only as an example, because more validation would be required before using these extrapolations for product qualification). The example chosen from (Logan et al, 1995) shows analyses of the front rail crashworthiness of aluminum extrusions in the spaceframe chassis. Energy Absorption, in Kilojoules per Kilogram (KJ/Kg) of rail mass, during the FMVSS 208 frontal impact standard (and variants) was modeled as a function of weight of the aluminum rail. The only validation available for this case was Validation By Similarity: That is, other aluminum designs have been modeled in similar ways, and experimentally show similar values and trends in KJ/Kg Energy Absorption. In contrast, consider the intriguing increase in KJ/Kg of foam filled aluminum extrusion rails. These numbers and their associated increase in KJ/Kg are completely unvalidated as used in this preliminary design study. In Part II, we will discuss the procedure for quantitative validation statements and for quantifying predictive confidence. That is, we seek to address how we would place bounds on an extrapolation outside the validated region of rail mass (if a validated region existed in the first place).

6. Phased or Tiered V&V: After (AIAA, 1998) and (Trucano et al, 2002) we suggest that the previous step be denoted and tracked at one of four phases or tiers:
    a. Unit Verification or Validation: A single code feature or quantity of interest

b. Benchmark or simply coupled V&V: A single coupling of features or quantities of interest; likely still a single type of physics e.g. mechanics *or* thermal *or* fluids. Verification should still be possible for this class. Validation at the component level.

c. Complex coupling V&V: Validation at the component or integral level, with multi-physics couplings e.g. mechanics *with* thermal *with* fluids. Verification will be difficult if not impossible; some Method of Manufactured Solutions (MMS) verification may be possible (Roache, 1998). This Tier is analogous to VAL_{suv} as in Step 5f.

d. Integral V&V: A V&V level suitable for integral system assessment, qualification, and certification. This Tier is analogous to VAL_{mlv} as in Step 5g.

7. Metrics: Quantitative V&V, Independent of the Code Development Teams:

   a. Specification of a set of metrics for assessing the execution of the above activities. In the case of a Milestone obligation, such metrics should be specified when the code capability is acquired or declared by code development; with a subsequent 1-2 year period to enable the quantification of V&V and UQ for the capability. Core components of this set of metrics include quantitative criteria for specific problems and acceptable thresholds of code performance on specific problems as determined by the ultimate users of the code product, specifically V&V and/or design and assessment teams not part of the development team.

   b. Assessment of code and model performance for the selected activities via the defined metrics. Is the validated model adequate to be used with design department funding for assessment or qualification decisions? If not, what is the path forward to achieve this?

8. Path Forward: Generation of future V&V [and code development] actions based on the outcomes of this assessment.

   a. What more is needed from code teams based on the V&V done?

   b. What additional V&V is needed?

   c. What additional experimental validation theory and data is needed to reduce undue uncertainty? This can only be justified if we can show a substantial sensitivity S_{Cij} to that material, environment, tolerance, or other uncertainty U_{Pj} (see Eqn. [3f]).

9. Documentation and archiving of experimental data and model results sufficient for future traceability and reproducibility of V&V activities.

   a. Each V&V analysis report should address an item on a prioritized Watch List ; such a list is called The Matrix in Part II of this work (Figure II-14); these matrices may be generated using QFD or Quality Function Deployment (Kogure and Akao, 1983), the PIRT or Phenomena Identification and Ranking Table (Pilch et al, 2000), or other methods.

b. Each report should address this Point Criteria (this point and the preceding points) and the detailed 19-point Verification and 35-Point Validation bullets to follow. This requires a graded approach for a balance of adequacy with efficiency.

c. Where appropriate, the report should choose and discuss quantitative measures and metrics for the quantities of interest being verified or validated in that particular work.

d. Each report should summarize the overall status of V&V for a feature or model quantity of interest the VER or VAL Meter reading. We suggest as a first cut that the designer/analyst evaluate uncertainties U as best possible and then set the meter reading V as V=2/U. We realize that these Meters have a subjective nature, and yet due to the pace of use of simulation summary results at quite high levels, some simple measure of fidelity is needed to flag the audience to the amount of overall caveats.

e. Quantitative information about the most important sensitivities S_{Cij}, parameter uncertainties U_{Pj} and environment Uncertainties U_{Ci} should be provided. The level should enable an analysis of these uncertainties, using the methods of Part II and Part III of this work, to justify further investment and iteration in improving the codes, models, data, or all of the above.

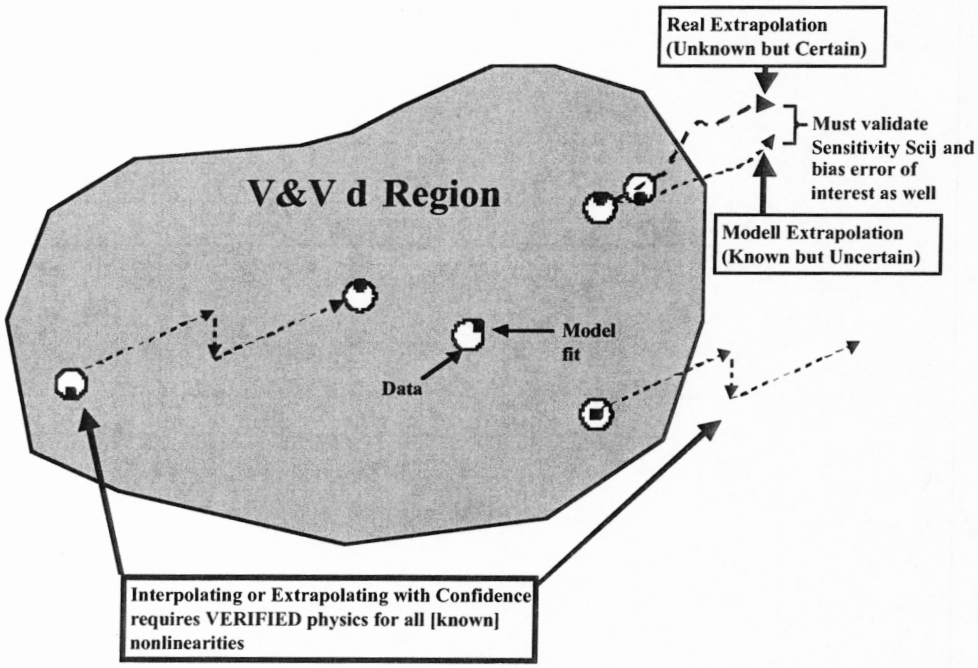
10. Integration into the larger V&V Community: This involves at least four steps, and offers the chance to maximize leverage and scientific credibility:

    a. Presentation and review of the documented V&V analysis in the community or external audiences.

    b. Review of the analysis and documentation.

    c. Establishment and augmentation of a common (e.g. TriLab nuclear weapons or community at large) Verification Problem set. Selections of the problems from this common set can and should be used to test new code capabilities.

Establishment and augmentation of a common (e.g. TriLab nuclear weapons or community at large) Validation Problem set. Selections of the problems from this common set can and should be used to test new model capabilities. Data available from documented, archived validation studies at the component level can be leveraged across a larger external community and offer more chances for inclusion as a common across agencies validation suite. Further, sharing of data and analysis results at the modular or component level can enable others to better plan their own validation activities without unnecessary duplication. A rich data base of modular validation tests and models can also enable the community to do more Validation by Similarity as described above; where, depending on risk and consequence, less formal validation may be needed if the community can leverage a large body of documentation already existing on a given code, model, and quantity of interest.



24

Figure I-8. Even if we define a cloud or region of V&V for a particular code/model combination, both Verification and Validation inside the cloud are essential to mitigate the risks that although the validation fit (VAL_{mlv}) may be quite good, the validation of a given sensitivity S_{Cij} (for VAL_{suv}) may be poor or unknown.

We now provide a numbered list of criteria to set the VER and VAL Meters for a given analysis. These points should serve as reminders to the designer/analyst to use their own judgment as to which and how many of these points apply, depending on the risk and consequence of the particular analysis. In other words, we feel that judgments about adequacy are embedded in any execution of V&V on a particular model.

1.4.3 19 Verification Points to help set the VER Meter:

1. Code features being verified: Point to them in the manual if possible.
2. Under what conditions? If a material model is being verified, with which element types? Solvers? Contact? Coupled physics?
3. What level of verification accuracy is necessary for the intended application?
4. Is the mesh used for verification typical of Validation / stockpile analyses?
5. What is the average VER Meter reading for features alone and in combination?
6. Mesh convergence study? What is the order of accuracy of the formulations used? Was this order of convergence achieved? How much accuracy will be lost as we coarsen to the mesh of a typical production 3D simulation?
7. Time-domain Solution convergence? What is the order of accuracy of the formulations used? Was this order of convergence achieved? How much accuracy is lost (or gained) as time step is varied within e.g. Courant (Explicit) limits or other limits (Implicit)?
8. Solution convergence in the iteration / stability domain? How do hourglassing schemes or differencing (forward, central, order of advection) affect the solution accuracy?
9. Based on the above what is the estimate of Uncertainty in the model as used in a production run?
10. Horsepower study? (element-steps/millisecond or other equivalent measure?)
11. What code development, new features, or Verification work would you recommend next?
12. If the VER was linked to a prior code capability delivery or Milestone, does the VER Meter reading reflect how well that Milestone was met?
13. Is remedial work needed by the code team to meet the claimed capability as promised?
14. Range for VER Meter readings — range of inputs where we trust the reading
15. Range and parameters where you feel the VER Meter reading is least trusted
16. Which items in the (LLNL-ENG, 2001) Table 7.7 checklist did you address?
17. A Verification Draft report should exist (addressing in fact this list)
18. Verification Draft report should be reviewed as appropriate (e.g. (LLNL-ENG, 2001)).
19. Is the Verification planned for external presentation / publication? If not, why not?

1.4.4 35 Validation Points to help in setting the VAL Meter

1. What question[s] is the analysis, when validated, trying to answer?
2. To what Uncertainty?
3. Concurrence Vs. 2nd code?
4. Concurrence Vs. Data, Post-Mortem?
5. Concurrence Vs. Data, Predicted?
6. Cross-test: Repeat steps 3-5, but with consistent model (no tuning) for several tests across the same system.
7. Cross-System: Repeat steps 3-5 but with consistent model, for different systems (design / geometry / materials).
8. Repeat Steps 3-7 and evaluate uncertainties in quantities of interest for consistent model — across multiple codes, tests, systems — and in predictive mode.

9. How many code features used have Verification documentation? How many do not?
10. What is the average VER Meter reading for features alone and in combination?
11. Mesh convergence study? Show mesh sensitivity and quantify it.
12. Solution convergence study in time domain: Show and quantify convergence [and the error accepted for the sake of expediency].
13. Solution convergence study in iterative [stability] domain: Show and quantify convergence, and again the error accepted for the sake of efficiency. For issues like radial return plasticity for example, Validation by Similarity may be appropriate and efficient as opposed to a time intensive numerical study.
14. Horsepower (model speed e.g. cell-steps/millisecond) study? (Relates to expediency)
15. Overall speed — including setup and post-processing time, and time to generate a new deck and make related / UQ runs? (Relates to expediency)
16. Complexity of e.g. material models: Effect on quality of the answer?
17. Complexity of e.g. material models: Effect on Horsepower? Number of runs per week?
18. Complexity: Was it worth it?
19. Sensitivity: Pick *at least* 2 key model parameters. Vary them by say 2σ uncertainty on the inputs. How much difference does it make? E.g. (Pao et al, 2000).
20. Based on the Uncertainty / Sensitivity question, what series of experimental material/component or system level Validation tests would you recommend?
21. What measures did you use to express agreement [i.e. a measure of UQ is needed to define agreement]. Was it the viewgraph norm ? Peak load? Peak strain? Stress as $\int|\sigma_i - \sigma_i|dt$? $\int|\sigma_i - \sigma_i|^2 dt$? Etc. For the $|\sigma_i - \sigma_i|$ type terms, compare code vs. code but also model vs. test at either component or system level.
22. Based on overall VAL, what VER work would you recommend next?
23. Estimates of calculational (model) U? (E_2, E_3, E_4 in Eqn. [3])
24. Compare the experimental (E_0, E_1) and calculational uncertainties. Which dominates? In other words, address the 5 Error or Uncertainty measures E_i in Eqn. [3].
25. CAL vs. VAL: Of the available N (Frequentist or Fuzzy) data sets, how many were used for CALibration of the model vs. VALidation? What is the CAL/VAL ratio?
26. If the VAL was linked to a prior code or material development Milestone, does the VAL Meter reading reflect how well that Milestone was met?
27. What remedial work if any is needed for the code capability to be as promised?
28. Domain of Validation for the VAL Meter readings — where we trust the reading. In other words, make a quantitative Validation Statement inside the Validation Box such as: within the box of element size < 2mm and peak G s <2000, the quantities of interest (stress and plastic strain) carry a model uncertainty less than 10% at a confidence level of 2σ, assuming a Gaussian PDF.
29. Range and parameters where you feel the VAL Meter reading is least reliable. In other words, make a similar statement of where the outside of the validation box is, and estimates of prediction error as [say] extrapolation distance outside that box.
30. Which items in the (LLNL-ENG, 2001) Table 7.7 checklist did you address?
31. A Validation Draft report should exist (addressing in fact this list!)

32. Validation Draft report should be reviewed as appropriate, e.g. per (LLNL-ENG, 2001).
33. Validation Final report with appropriate distribution.
34. Is the Validation planned for presentation within the community? If not, why not?
35. Is the Validation planned for external publication? If not, why not? (Obviously there are constraints, however there are and should be times [e.g. a container] where unclassified portions of even a validation analysis can be published externally).

1.4.5 Relating the "84 point" criteria to VER and VAL Meters and other criteria

How does one relate the preceding 84 point criteria to the VER and VAL Meter readings, and to the ability to make quantitative statistical validation statements such as those quoted as goals in the definition section?

Certainly if all 84 points are addressed in depth, we will achieve VER>7 and VAL>7, and be able to make statistical V&V statements for the quantities of interest. The details of this process are the topic of Part II of this work — with many complementary details in the works of others in references noted therein. One might even dare to call such a work fully validated at say VAL>8, even though we prefer to think of fully validated as an oxymoron.

If none of the 84 points are addressed, we will almost certainly have VER<3 and VAL<3, and statistical V&V statements [with any credibility or support] will be impossible. This condition might be termed unvalidated .

It is our experience that most code/model combinations in actual use lie in the range 3<V<7, in other words in the subjective, middle ground. Some *weak* statistical assessments can be made; quantitative statements indeed but not complete ones. Some of the 84 points above will have been addressed, but not all or even most. This is the reality, and can lead to another dangerous situation that leads us to advocate the meters even in the presence of a [nearly always incomplete] statistical validation statement. Even when such a statement is made for a quantity and regime of interest, there are always a list of model, code, and data caveats to accompany the statement. We have seen many times when the quantitative statement about Margin and Uncertainty is shown, but the caveats are dropped for expediency. This leads us to favor the V&V Meter reading as a simple summary number that hopefully will remain on the summary form presentation of an analysis; as a cautionary note regarding the fidelity and unstated caveats about the analysis being presented.

Recently, since our 69-point criteria described in (Logan and Nitta, 2001), we have expanded our point criteria from 69 to 84. We have found that our 84-point criteria encompasses the bullet-form criteria measures found in other V&V literature as shown in Table I-1 as follows:

Table I-1. Comparison of bullet-form criteria addressed in various V&V guidelines.

Reference	Number of V&V Criteria Bullets	Number Covered By Our 69 (now 84) —Points
J. Heat Transfer Ed. Board (ASME, 1994)	7	7
IJNMF Ed. Board (IJNMF, 1994)	6	6
AIAA Ed. Board (AIAA, 1994)	3	3
J. Fluids Eng. Ed. Board (JFE, 1993)	10	9, now 10
AIAA Guide (AIAA, 1998)	6	5, now 6
Sandia VAL Metrics (Trucano et al, 2001)	7	7
Sandia V&V Guide V2 (Pilch et al, 2000)	26	25, now 26
Sandia Exptl Val** (Trucano et al, 2002)	41	28
Los Alamos V&V Plan (Heath et al, 2002)	52	35, now 48
TOTAL:	158	141

***The lack of near-unanimous agreement with (Trucano et al, 2002) stems in large part from that document s role as an experimental guideline; there is a lot of overlap with code verification and model validation, the focus of this work; but by definition the overlap is not complete.*

Certainly we are not suggesting that a criterion for V&V credibility have anything to do with the number of bullets — but it is reassuring to know that the guidelines written over nearly the last decade are attempting to address essentially the same issues. The struggle seems to be more with exact definitions, requirements vs. suggestions, and the balance of cost effectiveness, credibility, and timeliness.

We suggest only that when determining the V&V status of a given capability, that some of these criteria be addressed and commented upon explicitly. The number of criteria addressed and the (judgment based) quality of the capability relative to the criteria will then determine the V&V Meter readings.

1.4.6 V&V Meters and Checklist Steps: How many is enough?

We might well end by envisioning a series of VER Meters for each code, all measured both in absolutes but also relative to the demonstrated abilities of other such codes:

1. Overall VER Meter reading for a given code s STS-Relevant features
2. VER Meter reading for a given code in the different areas of mechanics and physics relevant to STS analyses
3. VER Meter reading for each code feature / material model itself
4. Additional VER Meters for each code feature / material for Serial vs. SMP vs. MPP Verification status, and for VER of the feature alone vs. in conjunction with other features e.g. non-square elements, contact, coupled physics.

Similarly, a set of VAL Meter readings across codes should exist for a given type of STS Environment analysis. Naturally, in the ASCI (A for Accelerated) world, it is important to note and track whether these V&V Meter readings are on the old serial compute platforms, or whether they are truly MPP capabilities. Either way, the speed of analysis to a given level of accuracy and validation is a key measure, and indeed speed has been noted as a measure of Quality in codes striving for Quality Assurance using ISO 9000 methods (Keane, 2001).

ASCI is well into its 7th or so year of existence. ASCI V&V is in its 4th or so of existence. With the theme of *delivery* of capability growing, it is important to provide measures that summarize the successes we have had in ASCI capabilities that are ready for design and assessment use — and the areas where we have more to go.

Figure I-9 is a way of assembling high level VER and VAL Meter readings into an overall status summary. It reflects this summary by translating the VER and VAL number readings into Extended Stoplite colors as shown under the VER and VAL number readings. V&V Meter reading ranges are provided for each of the ASCI-engineering code suites we currently use; MPP/LS-Dyna, ParaDyn, Ale3d, and Sierra respectively. The readings are provided for the 1st V (VER) and the 2nd V (VAL), and for Horsepower (HP), defined currently as element-steps per millisecond on representative STS analyses in the V&V Sanctioned Muscle Code Shootout series. This set of readings, V, V, and HP, is repeated for both the Serial / SMP mode and for MPP mode. Note that although we have made great strides in ASCI-STS, most of the V&V-HP ratings under MPP are in the Explicit Lagrangian mode. This is telling regarding where the frontiers remain.

V&V Meter readings are also provided as a function of the nature of the STS mechanics and couplings being used. For example, are the mechanics Implicit (I) or Explicit (E)? Are the couplings Mechanical Explicit (ME) to Thermal (T)? Transport (X) to Thermal (T)? Thermal (T) to Chemical (C)? These are basic capabilities that tend to be topics of ASCI Milestones. We have postulated a set of V&V Milestones that allow a 2-year or so interval between ASCI Apps demo and ASCI V&V development of that capability up to a VAL Meter quality that is ready for design, assessment, and certification use. This interval allows time for interaction, feedback, and validation tests provided by [hopefully predictive] experimental validation opportunities.

KEY:
- GREEN
- # YELLOW
- ORANGE
- RED

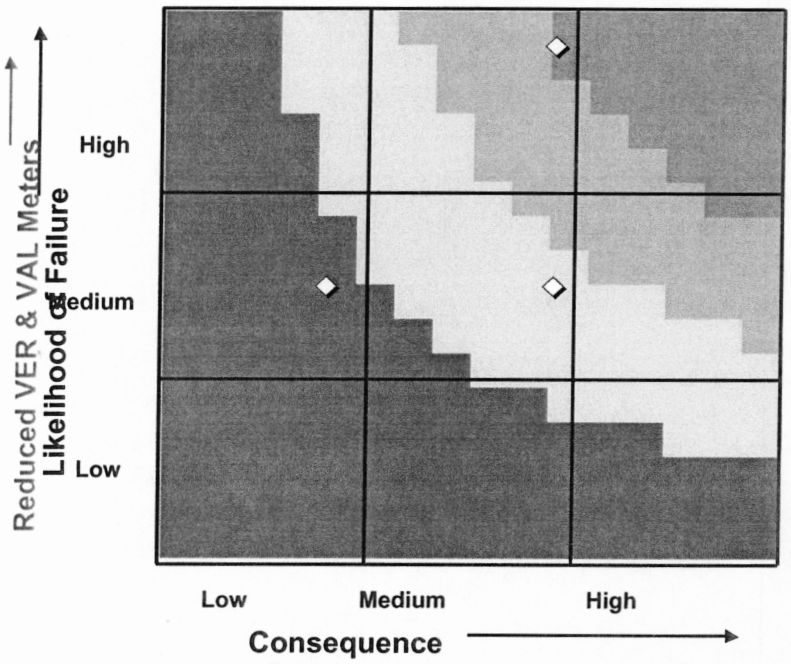
	MPP/LS-DYNA						PARADYN						ALE3D						SIERRA					
	Serial			MPP			Serial			MPP			Serial			MPP			Serial			MPP		
	VER	VAL	HP	VER	VAL	HP	VER	VAL	HP	VER	VAL	HP	VER	VAL	HP	VER	VAL	HP	VER	VAL	HP	VER	VAL	HP
Mechanics-(I)-Linear	#	#	3						3						1									
Mechanics-(I)-Nonlinear	#	#	3						3						1									
M+I+Preload, offset, etc	#	#	3						3						1									
Mechanics-(E)			60	#	#	4200			60	#	#	2500			20			1200			70			70
Thermal (T)	#	#	7				#	#	9						3									
Chemical [C]									1						1									
Diffusion																								
Advection, ALE																								
X-Transport																								
ME+T		#	7												1									
MI+T			1												1									
MI+T / Threaded			1												1									
M(E)+M(I)																								
T+C									1						1									
X+T																								
X+T+M																								
Aero+T+M																								
Fire+T+M																								

Figure I-9. The VER and VAL numbers translate into Red-Yellow-Green stoplight charts, enabling a good visual grasp of the state of maturity of our V&V capability from year to year. Both serial and MPP columns are shown; the only significant MPP capability with significant V&V is Lagrangian Explicit (E). Implicit=(I); Mechanics=(M); Thermal=(T); Chemical=; Transport=(T).

We should remember that our meter reading , even for each cell as shown in Figure I-9, is really only an attempt to roll up the code/model capability of generating acceptable analyses with acceptably (and demonstrably) low uncertainties for Equations [2] and [3]. It would be better to retain and report the application and quantity specific V&V statements as in the definition section — but often abbreviated time scales for the presentation of information prevent this and a qualitative meter reading is needed.

*1.5 Relation of V&V to Risk=Likelihood*Consequence*

Qualitatively, the V&V Meter readings can be easily related to the oft-used Risk=Likelihood*Consequence diagram for a system and series of environmental events as in Figure I-10. In Part II, we will show a way to quantify V&V in terms used directly in such a diagram, as well as in our own QRC/QSV expression.



55

Figure I-10. Relation of V&V efforts to the Risk=Likelihood*Consequence matrix. Reduced VER and VAL meter readings denote higher Likelihood of failure due to model error.

2 PART II: Relation of V&V to Adequacy, Acceptance, Qualification, and Certification

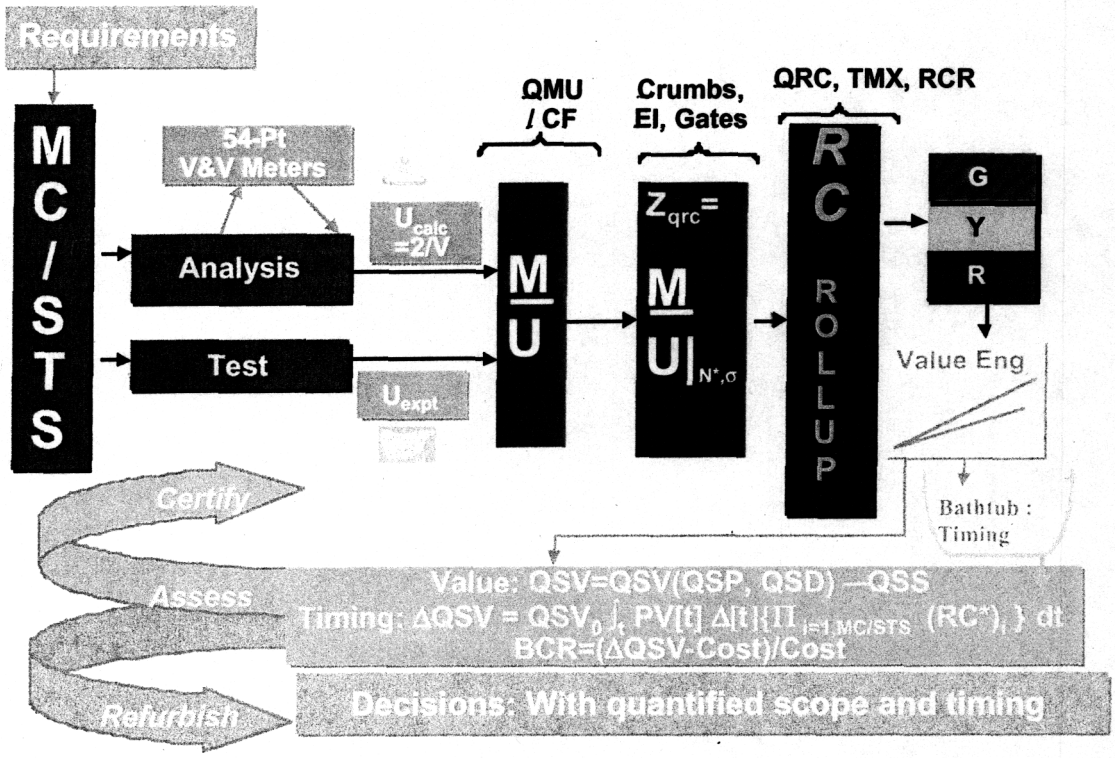


Figure II-1. Flow diagram for Part II: From system requirements (MC/STS) translated into Margins M with validated models from Part I. Uncertainty Quantification (UQ) of Part I then enables M and U to combine into Reliability at Confidence (R at C) terms and then Rollup those terms.

2.1 Relation to Certification Methodology, Metrics, and Prioritization

In order to set true acceptability Metrics for Validation, we must know where we are going — what the model tells us, and how much it matters.

Our goal in Part II will be to show the use of a path from V&V (with UQ i.e. Uncertainties at stated Confidence) through the use of validated models to get component and integral system margins (i.e. Factor of Safety = Margin+1). With these quantities, we proceed to an ***assessment*** *of Quantitative Reliability at Confidence equivalent for conditions of little or no full system testing. Figure II-1 shows the highlights of our journey in Part II.*

In Part II, we describe how to take the V&V and UQ of Part I and proceed to use these models to assess margins and then reliability at confidence equivalents, on through to an RC Rollup, a total performance measure for the nuclear package. (In Part III we will use this RC Rollup for a subsequent Value Engineering and decision process methodology). Along the way in Part II, as in Part I, exists the quantitative, albeit judgment based, information to take numerical derivatives that enable inputs to cost/benefit tradeoffs later in Part III.

An important theme of the Part II to Part III coupling is that these are the areas from which we derive acceptance or adequacy criteria. What is good enough V&V is a function not only of Requirements (MC/STS for us) but of the implications regarding Benefit / Cost tradeoffs with time and the rest of the mission and product line. We feel that to be ready to address acceptance in V&V, it is essential to have a process that leads, in the end, to a Benefit / Cost closure. Otherwise, we will be doing UQ in V&V forever, never knowing when we should stop.

For VERIFICATION, we can more simply postulate the exact solution as the quantitative acceptance goal, and measure and document our ability to approximate an exact or manufactured solution on a design-effective time scale. That is, we must quantify our verification of a solution, but one that we can afford in the context of computer and time to generate, run the model, and post-process the results. This quantified verification should be related to the quantity of interest, with a statement that the Eqn. 3 uncertainty errors E_i (specifically E_3 and E_4) are, e.g., less than 5% in displacement at a confidence of 2σ or 95% , and we show or assume a Gaussian Probability Distribution Function or PDF.

For VALIDATION, we can set some qualitative measures [e.g. the 35 measures described above], and we can set some quantitative acceptance *goals* by inference from similar problem classes over the years. Of course, we have to consider, for Validation, the entire suite of uncertainty errors E_i, E_0 through E_4 in Eqn. 3 above. We can state some *typical* quantitative goals for Uncertainty Quantification [UQ] metrics for simulation fidelity with examples:

Mode Frequencies: Within 2% Linear, 5% Nonlinear

Strain: Within 20%

Stress: Static, within 20%

Deflection: Within 20%
Peak Accelerations, dynamics: Within 20%

We would suggest that in most cases (especially for explicit dynamics models) these and other quantities be expressed as an averaged integral over some appropriate time interval as is common in, for example, vehicle crashworthiness analyses and standards. Peaks obtained on a single time

step can be misleading with undue noise; the rolling integrals over time have proven to correlate well with reality in passenger safety [e.g. the Head Injury Criterion or HIC] and other fields involving time-dependent mechanics. Of course, in doing so, we must be careful to use enough fidelity in our integral measures to avoid ambiguity. For example, we relate a comment from (Schwer, 2002). Consider the HIC definition for a moment, with criterion HIC<1000, acceleration a in multiples of acceleration of gravity g, and interval (t_2-t_1)<36 milliseconds:

$$\mathrm{HIC}=\left\{\frac{1}{(t_2-t_1)}\int_{t_1}^{t_2} a\,dt\right\}^{2.5}(t_2-t_1) \qquad (4)$$

Now, consider this moving window integral for occupant head acceleration. It is true that two nearly identical HIC numbers can be obtained, say, from a light vehicle with an airbag or a heavy vehicle with only a shoulder harness should they collide with each other. Could we then say that since HIC is the same, we have validated that the light vehicle model is identical to the heavy one? No — but we must be clear about the comparison we are making. The integral measure is appropriate for many quantities of interest, e.g. accumulated damage over short intervals of time, hence HIC. But before we can say that two models are the same because they give the same HIC, we must also look at the rest of the model — it must be a consistent model of one vehicle or another, i.e. light vehicle mass, stiffness, geometry, containing an airbag versus only shoulder harness, etc. If all these geometric and environmental inputs to the model are the same, and we obtain the same HIC (i.e. within acceptable uncertainty at confidence) as defined over any interval (t_2-t_1), then we might indeed claim that our model is well validated (to the quantitative levels implicit in this paragraph) for HIC.

Some of the values described above are *typical* uncertainties or agreement of models vs. data. However, that does not address at all what may be *an adequate* level of uncertainty. If our Margin (M) is low, we should require [and hence pay to determine via validation, analysis, and test] a lower uncertainty (U); and vice-versa. These are in fact quite close to the range of experiment minus calculation values we have tended to regard over the years as goals. However — what if we are outside them? Further, what if we can do better — get the 20% quantities down to 10%? The value of this reduction in uncertainty is not a step-function — it is a continuous scale. We use these numbers to make priority decisions, and implicitly our desire is to maximize the value of [in this case] the stockpile as a deterrent.

A metric for Uncertainty Quantification UQ is by its very nature tied to *Sensitivity* to that *Uncertainty*. The level of rigor in the UQ metric must therefore be proportional in some way to the value, priority, effort, etc we are willing to invest into reducing the uncertainty in that STS scenario. In other words, we can and perhaps *should* specify a higher acceptable uncertainty at a given confidence level for things that don t matter much. That way, we can concentrate our

resources on achieving a smaller assessed uncertainty for things that matter a lot. Others (Trucano et al, 2001) have as well noted the need to address quantification, but also address sufficiency for both Verification (ibid., p. 22) and Validation (ibid, p. 62).

We must prioritize our resources and determine the acceptable level of uncertainty. For product acceptance [in our case stockpile stewardship], it is our responsibility to assure, with high confidence C in a high assessed reliability R, that the stockpile system can and will accommodate its STS through its logistics and delivery lifetime, with high C in a high R handoff through the nuclear performance regime. We are developing and quantifying such a methodology to lead us to values of C and assessed R to fold into a total performance number or non-frequentist RC-Equivalent for the nuclear package. We emphasize the wording, *assessed* Reliability at Confidence, because although we may never disprove an assertion that Reliability is in fact Unity, we can only quantify what we can assess with positive numerical evidence.

In the following section we will take the next step toward closure of our Value Engineering / Investment Strategy method by developing Quantitative Reliability at Confidence (QRC). QRC depends on the notion that Uncertainty (U) and Confidence [C] are statistically linked; whether we are coin flipping with a frequentist number of coins N or using an inferred number of information points fuzzy N*. With an (albeit assumed) PDF, we can use our V&V-obtained Uncertainty at Confidence *statistical validation statement* to obtain an assessed Reliability at Confidence.

Success will depend on the existence and fidelity of our statistical validation statement for the quantity and condition of interest. Hence, we must now go beyond the VER and VAL meter expressions for V&V, and note that in order to move into QRC:

- V&V is now a statistical process, whether with a frequentist N or inferential, relevance based N*
- V&V must provide uncertainty at a stated [and quantified] statistical confidence
- V&V must show the origins of its N or N* and the [perhaps expert judgment] weightings used
- V&V must allow us to assess a Reliability measure [R] from Margin (M) and assumed or known PDF, whether normal distribution or other
- V&V must consider adequacy, before stating whether a given model is validated for its application or not

These factors will allow the V&V process (and resulting validated models) to contribute to the Value Engineering Investment Strategy that has a mathematical closure (in dollars for example).

Of late, there has been much interest in quantitative certification, a quest for confidence that is more than just low , medium , or high . There is much more to do. So it is essential to clarify what methods can and cannot be credibly used under given circumstances, because of the importance of the topic and the methods, and above all, the emerging desire to use them as **business decision and investment strategy tools.**

As indicated in Figure II-1, there are several methods that have the potential to lead to quantitative certification, denoted as CRUMBS (Logan, 2001a) and (Logan, 2001b), Gates (Juzaitis and Sharp, 2001), TRB (Hamada, 2001), EI (Booker et al, 2001), the latter ones denoting work at Los Alamos. In addition, there are some noteworthy early works [sans acronyms] published by Los Alamos including (Adams et al, 1999) and work described in (Flicker et al, 1999). These works have helped to motivate our methodology as described in this paper.

In some cases, the uniqueness of our CTBT situation is giving whole new meanings to terms like performance, probability, reliability, and confidence. Just like V&V, the terms do not yet have unique meaning in the community. The eventual result of all these methods is, in our own nomenclature, the goal of an *assessed lower bound* Reliability-Confidence Equivalent for components and conditions in our portion of the weapon system, Rolled-Up over the entire system and its lifetime environments, leading to the RC Rollup nomenclature.

2.2 Moving from V&V and UQ to QRC: Quantified Reliability at Confidence

We will describe a general process for moving from V&V and UQ of Part I to quantified certification as we view it and have discussed in recent documents (Logan and Nitta, 2000) and (Logan and Nitta, 2002). We present the basic equations for FOM/QMU, CF, and QRC used by various groups (ours and others) in analysis of margins, uncertainties, and in our case the extension to measures of quantified reliability at confidence (QRC), for the nuclear package at the component, environment, or system level. We attempt to show, with a quantitative example, our non-frequentist statistical struggle in trying to quantify comfort , or conservatism , or confidence . We have built as much rigor and [quantified] judgment as we can into QRC, and employed statistical terms, be they frequentist or inferential in nature. We have balanced this against our desire to keep our overall equation [1] modular in nature; meaning that we can successively turn off or set to unity features and terms until the equations and methods reduce to the binomial coin-flipping situation. We show that **methods like QRC enable closure to an Investment Strategy** method like QSV. Methods like Quantitative Margins & Uncertainties, QMU, (Goodwin, 2001) and Confidence Factor, CF, (Krauser, 2001) deal in their documentation with margins and uncertainties, and are important *steps along the path* toward quantification with closure — but QRC into QSV completes the closure. The path of V&V with UQ, then QRC, and then QSV is one way of expressing a methodology from V&V to Value Engineering with traceable, quantified closure.

We showed, in Part I, the process for moving from Requirements to V&V and Uncertainty Quantification at Confidence. We showed the use of a quantitative, yet judgment based, 0-10 Meter Reading for V&V status. We provided a simplistic U~2/V qualitative correlation, with a suggestion that we go further to quantify terms such as sensitivities (S_{Cij}) and parameter variabilities, changes or uncertainties (U_{Pj}). In the portion of our flow diagram highlighted for Part II (Figure II-1) we will need these terms.

Moving toward the middle of Figure II-1, we can now express a quantity called Figure of Merit (designated B here for Bont , or goodness); some of the research groups at LANL and LLNL have begun to express B=FoM=M/U in a notional sense. In this equation, U is a global U for a given environment scenario — it comes from an assembly of values of S to all the component, material, and numerical simulation values of U for system environments involving the nuclear package. Because of this, we have used the acronym CRUMBS (Confidence, Reliability, Uncertainty, Margin, Bont , Sensitivity) to describe the system for closure to Quantified Reliability at Confidence of all the quantities represented by the alphabet characters involved in this picture. The CRUMBS acronym is only meant to bring attention to the fact that in any such certification methodology, we must use enough letters of the alphabet to reach closure and avoid ambiguity in the numbers we generate. Similarly, the comment about inclusion of

Reliability R at Confidence C leads us into the next point. The key is not the *acronym*: Rather, the key is a *process* that provides closure to the terms.

FOM=M/U is a simple concept and indeed an old one, but it cannot be left open ended; for closure, U requires specification of confidence level C . Of note is the subscript on the *second* (M/U) term at the right of center. We must go beyond the first QMU (Quantified Margin and Uncertainty) in order to proceed into the realm of Quantified Reliability at Confidence (QRC). In the second (M/U) box of Fig. II-1, the Z_{qrc}=M/U at σ denotes some measure of standard deviation in a Gaussian or other Probability Distribution Function (PDF). This is an important factor required to later derive the reliability equivalent [R] quantity for the RC Rollup . Note the presence of N*, the number of data elements in the set being evaluated. The quantity N* has an analogy to, for example, a binomial N in binomial analyses for R and C. However, N* is a weighted N — weighted with the relevance of number of tests, relevance of tests, number and relevance (e.g. VER and VAL settings) of analyses, and in some cases an expert opinion weighting (Booker et al, 2002).

2.2.1 Measures of "M" and "U": As used in FOM/QMU, CF, and QRC

Our comfort at the system level comes from the sense that Margins are large . What does that mean? Figure II-2 was one of our first attempts to depict STS Margins , and note our nomenclature for uncertainty, factor of safety, and eventually inclusion of uncertainties. We also noted at that time that for investment tradeoffs over time, we must consider the real, assessable, and assessed change in these values over time. However, the depiction of Figure II-2 is not enough to take us into the realm of Quantified Reliability at Confidence. To do so, we consider Figure II-3. All three pictures in Figure II-3 define different ways of expressing the comfort of a Large Margin.

The pictorial in Figure II-3(a) depicts the Quantitative Margins and Uncertainties method (QMU). We can choose some predetermined [even implicit, e.g. Unity] definition for Reliability [R] and Confidence [C] and express our comfort using the value of the FOM=M/U. This method leads to the QMU Figure Of Merit (FOM) goal of M/U _FOM>1 or FOM>2, or a higher is better edict but one that does not close into reliability or direct benefit/cost tradeoffs. For this method we have, assuming as in QMU s FOM that U is a *total uncertainty measure*:

1<FOM: We are sure our M exceeds uncertainty (in other words, the system will function) within the limits of our declared U.

0<FOM<1: M is positive, but may be swamped by uncertainty in some proportion of the cases. We expect the system to function but there in an increasing chance it might not.

-1<FOM<0: M is negative, so we do not expect the system to function, but uncertainty in our favor may mean the system will function anyway.

FOM<-1: M is negative and our declared U is not large enough to overcome this even in the rare instances where we are lucky . We expect the system to fail each and every time it is called upon.

As an alternative to the QMU method of Figure II-3(a), we consider the method of Figure II-3(c). Here, we can express our uncertainty in terms of Normal or other Probability Distribution Function (PDF) defining in essence the point where a Standard Normal Distribution variable Z (Moore and McCabe, 1989) becomes Z=1. With our definition of mid-point margin M as M=1-FOS (Factor of Safety), there is a direct analogy between the standard statistical Z and our Z for QRC, $Z_{qrc}=(M/U|_{N^*,\sigma})$. Expanding U [and de facto raising R and C as M→0] may be expressed another way as: you have to express margin in terms of the probability it will be used up (Wood, 2000). The tails overlapping this closure then give us measures of R; and the method we obtain this information (equivalent sample size N*, our fuzzy N) gives us the value of C. For these methods, e.g. CRUMBS, TRB, MOM, QRC, we extend the preceding acceptance cutoff of |B|>1, a simple go or no-go criterion, to a statement that $|Z|>n_\sigma \sigma$. , where n_σ is the number of standard deviations over which we define our uncertainty (for the normal PDF assumption). This way, $n_\sigma \sigma$, and the corrections due to our fuzzy N*, give us *lower bound assessed* numerical equivalents for Reliability R at Confidence C that can be quantitatively defended from our V&V and assessment process. If the resulting R at C are not acceptable, we may be able to improve them by further investment in the codes, V&V, or validation testing. These are **then investment strategy choices** we can trade off in a quantitative way as described in Part III and in (Logan and Nitta, 2000).

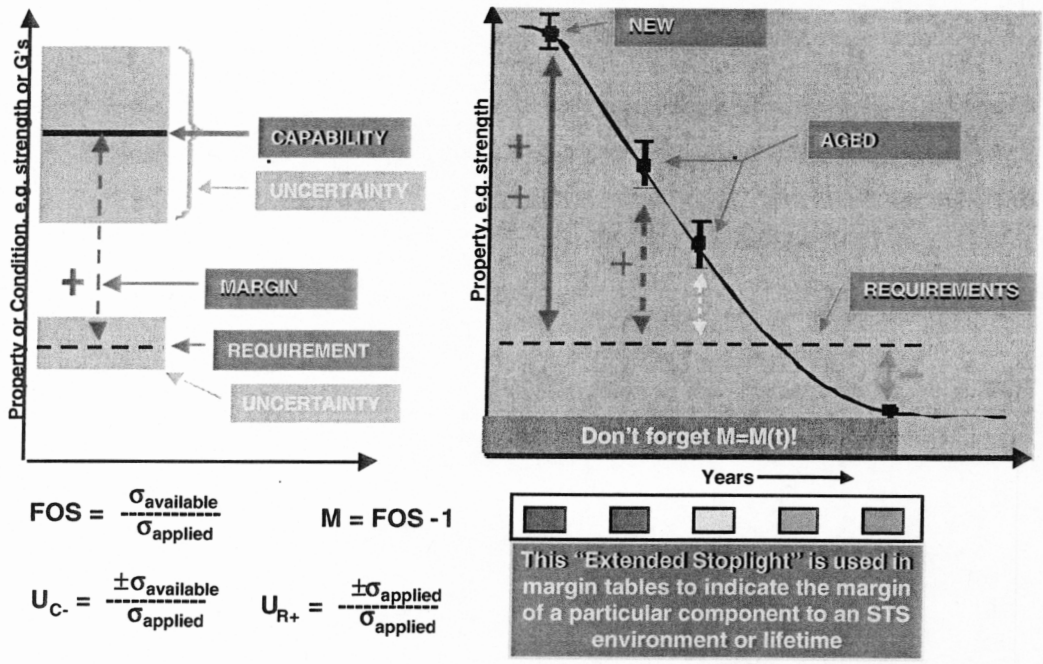


Figure II-2. Early depiction (circa 8 December 1999) of Margin, Uncertainty, and evolution with time: Vision of quantified evolution with time, enabled by today s QRC and EI.

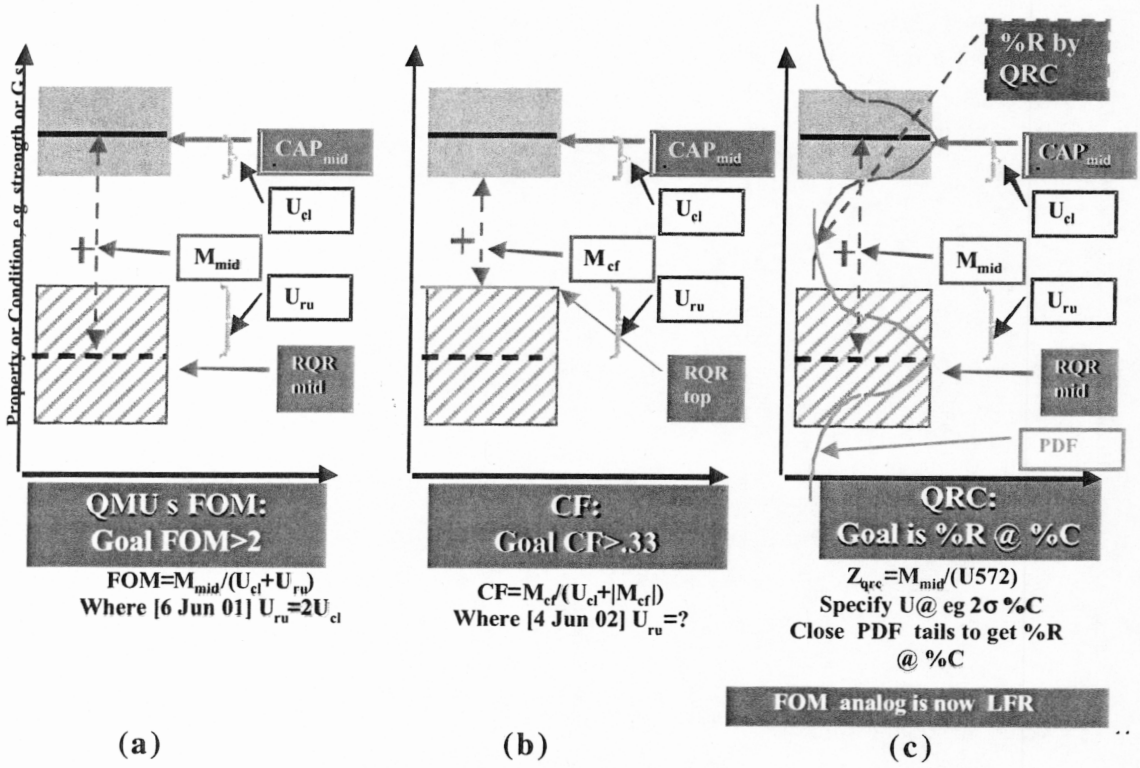


Figure II-3. Comparison of QMU's FOM, CF, and QRC's Z_{qrc} methods for Margins and Uncertainties. (Dates refer to first known presentation of FOM and CF.) Only QRC method allows quantitative extension to R(M,U) at C(N*), and then to QSV. U572 e.g. root-sum-squared quadrature defined in (Logan, 2001b). Similar methods to QRC are being developed at Los Alamos.

2.2.2 Contrasting "CF", "FOM/QMU" and "QRC"

Figure II-3 also compels a more lengthy explanation of CF, the method depicted in Figure II-3(b). This will enable us to contrast CF with FOM/QMU and with QRC/QSV. We take nominal M_{mid} as the distance between the mid-points of Capability-to-Requirement. The CF method uses M_{cf}, the margin or gate remaining after all the uncertainties of concern are accounted for. The differences in the methods come mainly in the definition of U . In the process control literature from decades past, U is often defined in six-sigma terms, that is, U is a certain size derived from statistics and process control data (Montgomery, 1985). This has been coined and used as the Capability Process Ratio, Lower [CPL] in work at Ford Motor Company (FoMoCo) (Kane, 1986) and others. As we entered the era of Margins and Uncertainties for the Stockpile, LLNL began to explore and use an open-ended, linear sum M/U as a Figure of Merit (FOM) at LLNL (Horrillo, 2001). At this time, the FOM U did not contain the rigor of definition of the CPL=M/U at a given statistical measure. Rather, U was simply defined as total of uncertainties . This is similar to the concept of gates introduced at Los Alamos during the same era as described in (Juzaitis and Sharp, 2001). The gate is defined as the region between the uncertainties, in other words as M_{cf}. We have described the dual-ended use of these methods in (Logan and Nitta, 2001) using Mesas , and insight into the relations between mesas and inverted mesas was provided by (Trujillo, 2001). One depiction of the relation between gates and Mesas is shown in Figure II-4. The width of the CF gate margin is taken to scale in some way similar to QMU s FOM M/U . We can see how this gate margin relates to M/U as follows. Using nomenclature similar to (Krauser, 2001) and in (Logan and Nitta, 2001) some relevant quantities are:

DP=Design Point — typically the middle of the Capability

OR_L=Operating Range, Lower half

U_{CL}=Uncertainty in Capability, Lower (half of the Capability uncertainty)

U_{RU}=Uncertainty in Requirement, Upper (half of the Capability uncertainty)

M_{CF}=Margin, Confidence ie Capacity Factor method. Distance in between the uncertainties of the lower requirement line and the capability lower band.

M_{1tail}=Margin, 1-tail method, special case to compare CF to QMU to QRC. Distance in between the uncertainties of the top of U_{RU} and the capability lower band.

$Z_{qrc}=M/U_{|\sigma, N*}$, after the Standard statistical Z, quantified uniquely and used in QRC.

These terms enable the definition of a Capability Factor or **Capacity Factor** [preference of these authors] or Confidence Factor (Krauser, 2001):

$$CF = \frac{M_{CF}}{|M_{CF}| + U_{CL} + OR_L} \qquad [5a]$$

CF takes on a range of acceptable values such that:

$$0 \leq CF \leq 1 \qquad [5b]$$

When $CF \rightarrow 0$, we have in effect a zero M_{CF}, although the mid-point M_{mid} used by QMU/FOM and by QRC/QSV may still be large and positive. As $CF \rightarrow 1$, this says that M_{CF} is large compared to uncertainties and operational range. The parameter EI (Booker et al, 2001), developed from and credited to industrial process control literature from decades past, has a form identical to CF as shown below; but like QRC, the EI documentation has recognized that, like the well established process control technology (e.g. six-sigma), uncertainties must be expressed *at confidence.*

How are the forms of EI, CF and QMU s FOM=M/U (and the $Z_{qrc}=M/U|_{\sigma,N}*$ used in QRC) related?

After (Booker et al, 2001) and (Krauser, 2001) and for this example, we note that $M_{CF}=M_{EI}$ and then let:

$OR_L \rightarrow 0$, then

$$M_{mid} = M_{CF} + U_{CL} + U_{RU} \qquad [5c]$$

$$M_{1tail} = M_{CF} + U_{CL} \qquad [5d]$$

Then, for a 1-tailed analysis where $U_{RU}=0$:

$$CF\ (\sim EI) = 1. - \frac{U_{CL}}{M_{1tail}} = 1. - \frac{U_{CL}}{M_{mid}} \qquad [5e]$$

Equivalent ranges are now:

$$0 \leq EI \leq 1 \qquad [5f]$$

$$0 \leq CF \leq 1 \qquad [5g]$$

$$1 \leq FOM \leq \infty \qquad [5h]$$

This discussion holds true for a 1-tailed analysis where the Requirement is a line, ie the comparison is only direct when $U_{ru}=0$.

For a 2-tailed situation with nonzero U_{CL} and U_{RU}, we must use

$$CF = 1. - \frac{U_{CL}}{M_{1tail}} \qquad [5i]$$

As the general quantity M/U is defined with more rigor in our own business, just as described by FoMoCo for their use in the automotive industry (Kane, 1986), we recognize that the 6-sigma process control notion can be extended to an n-sigma notion. That is, we can expand our uncertainty at some notional n-sigma, e.g. 2σ, 3σ to uncertainty at n-sigma, such that $M_{CF} \rightarrow 0$. These concepts, described in (Kane, 1986) and (Wood, 2000) and others are pulled together into QRC (Logan and Nitta, 2001) to lead to an RC Equivalent as we expand the tails of the uncertainties until the gate closes and meets. The N* that we use is in fact an Inferred or Fuzzy N*, rather than e.g. a binomial go / no-go or Frequentist N. Our inferred N* may be viewed as a test-plus-analysis-plus-judgment extension of Hamada s Bayesian N as described in TRB (Hamada, 2001) or Booker s Fuzzy Set N* (Booker et al, 2002). There is not a simple mathematical relation between the QMU/FOM, the CF, and the $Z_{qrc}=M/U|_{\sigma,N}*$ used in QRC. In the case where $U_{RU}=0$, the $Z_{qrc}= M/U|_{\sigma,N}*$ can be shown to be equal, assuming some level of $n\sigma$, to QMU s FOM. In the general case, due to the mathematical treatment of uncertainties in QRC versus the more ad-hoc treatment in QMU and CF, there is not so simple an equation for the relationship. We stress that these relations should be made as simple as possible, but not simpler.

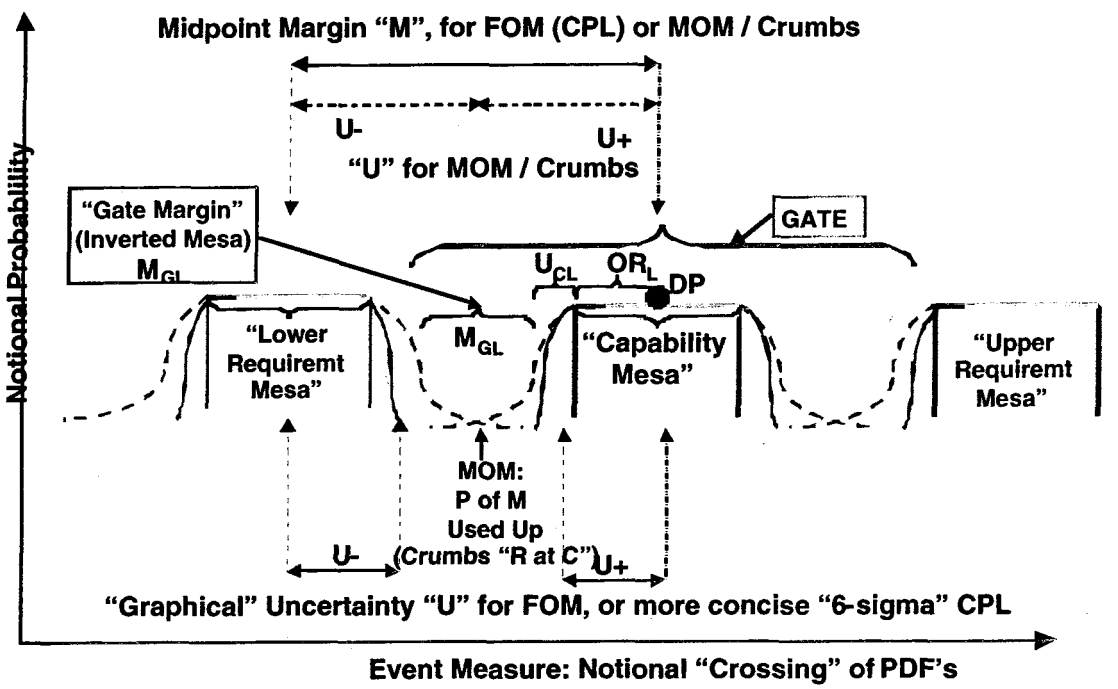


Figure II-4. Different methods and nomenclature for the use of V&V uncertainty quantification, processes that all can lead to an RC Rollup.

Although Figure II-4 is a more thorough depiction of the relations between the various groups in quantifying margin M, uncertainty U, and other quantities, we prefer Figure II-3 for its *relative* simplicity. Figure II-3 of course also graphically depicts the use of QMU s FOM on 6 June 01, the use of CF on 4 June 02, and the process used in QRC, with reference to its definition of uncertainty U (Logan, 2001b). We can use Z_{qrc} as depicted in Figure II-3 to obtain, using Bayesian Relevance Factor (BRF) and other methods, values for C and R entered into The Matrix, (Logan and Nitta, 2001), in the nuclear package for its logistic and operational environments. These values of C and R equivalents are our products — a way to numerically roll up the state of a system at the system level. Targets of opportunity for stewardship are then where R, C, Z_{qrc}, are lower than we would like, which may mean that U is higher than we would like. The phrase what we would like is another layer in our decision process that has begun to evolve and will continue to do so, with a measured evolution each year. This additional layer enables us to use Value Engineering to prioritize and Earned Value to evaluate (in Part III) our activities, and we have called this methodology part of our QSV (Quantitative Stockpile Value) method since circa 1997 and documented the process recently (Logan and Nitta, 2000). It is a circle that closes — we note the development of similar methods at Sandia (Trucano et al, 2001).

2.2.3 Features of QRC: The Good, the Bad, and the Ugly

We will now give a point by point narrative on the features of QRC, our preferred method for linking V&V and Uncertainty to Reliability and Confidence and ultimately value and investment measures. Table II-1 is a tabular summary of these points.
GOOD: Shows how to link Margin M and Uncertainty U , as the physicists may like to think, directly to Reliability R at Confidence C as engineers are more accustomed to.

GOOD: The $Z_{qrc}=M/U|_{\sigma,N}*$ term in QRC with a mid-point M is a good start as long as physical feel for the quantities we are used to is not lost; M/U has been used for decades as Signal-To-Noise and Process Control CPL as noted above. Since U at C is a direct output of our preferred validation statement, the $Z_{qrc}=M/U|_{\sigma,N}*$ term in QRC, as opposed to some other methods, enables closure; it is not open-ended.

GOOD: Most of the components of QRC are not really new. QRC certainly builds, in our business, on the works of (Wood, 2000), (Hamada, 2001) and others, just as it builds on the standard statistical definition of Z . We believe what is new about QRC is that it enables complete and quantified (in reliability, dollars, or both) linkage from V&V and UQ, through QRC, into QSV and Value Engineered Investment Strategy.

GOOD: The QRC/QSV method is as simple as connecting M and U to R at C can be — but not simpler .

BAD: While QRC is as simple as possible , it is not simpler. That means it is still complicated.

UGLY: The current usage and demonstrations of QRC into QSV have assumed Gaussian PDFs and independence of the various U terms. We apologize for perhaps premature numbers in that sense — and apologize to pioneers in this field (Pao et al, 2000), (Wood, 2000), (Adams et al, 1999) for using Gaussians, but we have nothing better to use thus far. Although binned as an UGLY , this is perhaps more of a GOOD because the challenge of a quest for better PDF s can now become a quantified mission for the relation between science, computations, and Value Engineering.

GOOD: In QRC/QSV, we attempt to realize and state and document our assumptions and form the method QRC+QSV so that it will accommodate this information [e.g. PDFs] as stewardship makes it available.

GOOD: The form M/U could be made proportional to the M we are used to seeing, if we defined a universal standard U for a given MC or STS event. Such a standard U was used for some of the BCR examples in (Logan and Nitta, 2000) regarding program planning priority quantifications.

GOOD: QRC does in fact address the definition of U as $U|_{\sigma,N}*$, or U at Confidence C.

GOOD: QRC does define Confidence C quantitatively, but with a simple analogy of our Inferred, Fuzzy N* to the frequentist N of coin-flipping.

UGLY: This quantified definition of U at C, if actually used, will negate the proportional relation between M_{mid} and $Z_{qrc}=M_{mid}/U_{qrc}$.

UGLY: For hard-core nuclear test advocates, we are going to be hard pressed in the CTBT era to find a large and defensible N*. We cannot think in terms of reliability and confidence as obtained in a binomial coin flipping exercise.

GOOD: QRC/QSV enables the concept first presented by EI (Booker et al 2001) of a time-domain Investment Strategy; we can quantify the brilliantly clear EI Stoplight in Time with the simple QRC/QSV investment strategy equation:

$$\Delta QSV = QSV_0 \int_t PV_F[t]\, \Delta[t]\{\Pi_{i=1,MC/STS}\ (RC^*)_i\}\, dt \qquad [1]$$

This equation, optionally using Present Value PV as we recommend, enables direct and quantified Benefit/Cost tradeoffs and timing decisions (Logan and Nitta, 2000). Note the term, $\{\Pi_{i=1,MC/STS}\ (RC^*)_i\}$, which in fact represents the RC Rollup of The Matrix (Logan and Nitta, 2001). Note that if there is only one system environment , then either the signal-to-noise like M/U , or R at C can be used to represent its status. However, if several environmental events must succeed

in series (often the case), R at C (or QRC) does enable a Rollup of these events; M/U in any form, does not to our knowledge allow such a Rollup .

QRC documented features	
Feature	QRC
Documentation	Series of reports
PDF	Gaussian or Other
As Simple as Possible	Yes
Still Complex	Yes
Ucl, Uru	Independent but not limited to that
Link from M,U to R,C	Clear and Easy
Standard Use of Math, Statistics	Yes
Use over Time Demonstrated	Yes, Integral of QSV*QRC
Definition of U includes C	Yes
Quantified Definition for C	Yes, Frequentist and Inferred
Can be made proportional to M	Yes
Direct term in Investment Strategy	Yes
Analogy in Engineering Literature	Yes, but QRC fold into QSV is new
Compatible with Sandia "R" Methods	Yes, but needs "Fuzzy N*"

Table II-1. Tabular features of QRC. EI has conceptualized integral over time: $EI(t)=EI(M_{EI}(t), U|_{N^*,\sigma}(t)$, as formalized in QRC as the term $\Delta[t]\{\Pi_{i=1,MC/STS}\ (RC^*)_i\}$. EI provides a quantitative early warning that links directly to the QRC analysis.

2.3 Example comparison of V&V methods with and without closure

The following example will illustrate in a quantitative way the limits of QMU and CF as defined in dealing with quantitative levels of comfort, conservatism, and confidence — and the ease with which complete methods like QRC lead to unique values for *assessed* reliability at confidence, and therefore enable an Investment Strategy closure in the same instances.

It will be helpful to refer to Figure II-3 containing the QMU, CF, and QRC method diagrams.

If U_{cl} and U_{ru} are fixed in the following examples, then the values of QMU/FOM, CF, and Z_{qrc} are unique.

If the "U" are really "total U" as both the QMU and CF methods assert, then all that we need is $Z_{qrc} \geq 1$ or FOM$\geq$1 or CF$\geq$0, and then perfect reliability and confidence are assured or at least alleged. However, we have seen instances where FOM=B_{qmu}>>1 or CF>>0 is demanded for extra comfort . We have seen no objective criteria for quantified benefit to FOM>1 or CF>0; hence our use of extra comfort for this unquantified [albeit intuitively true] benefit. Like the VER and VAL Meter Readings , desires for FOM>1 and CF>0 are vague unless quantitatively linked to a measure like QRC. In QRC, the Margin M and Uncertainty U are expressed as standard variables at statistical confidence — albeit at small sample size — but we can avoid vague terms like Excess Comfort . For example, we can examine how much 'Excess Comfort' is in each example value:

FOM= M/U =B_{qmu}=2
or
CF=.33

We will for this example comparison define Excess Comfort as multiples of M/U , and we will see that this is meaningful only in a qualitative intuitive sense. We will then define our Excess Comfort in terms of risk avoidance quantities.

Let s examine 2 cases:

1. Assume U_{ru}=0
[this may be effectively true if U_{cl} and U_{ru} are totally dependent.....]
Then FOM=B_{qmu}=2 shows Z_{qrc}=2, Excess Comfort is 2.0x

CF=.33 shows Z_{qrc}=1.493, Excess Comfort =1.493x

2. With U_{ru}=20
If Uru=20 as stated in the 6 Jun 01 FOM/QMU example, then:
FOM=B_{qmu}=2 shows Z_{qrc}=2.683, Excess Comfort =2.683x

CF=.33 shows Z_{qrc}=1.562, Excess Comfort =1.562x

Either way this is just an arbitrary amount of Excess Comfort . To get some insight into the meaning of this Excess Comfort , let us compare it to more common risk assessment

terminology. Let us use a conversion to QRC, then NeLoFaRA. (Negative Log Fail Rate *Assessment* - a common practice in industry and even with some of our partners e.g. (Trucano et al, 2001)). We will abbreviate NeLoFaRA as LFR for short, but remember that it really is the *negative* of LFR, and that this is an *assessed* failure rate, based on assumptions about the uncertainties and the PDFs, with an inferred and Fuzzy N*. Although we wish we were coin flipping many thousands of coins here, we unfortunately do not have that luxury.

The "A for Assessed" in NeLoFaRA should be key to easing our minds and to convey the correct message. QRC *is* a quantitative method with closure - and like standard Risk Assessment methods it enables a direct link into Value (QSV) for systems either in existence or those being contemplated for the future. However:

It is only a *computed* LFR, from a *computed* QRC. We do not have enough information to compile or even approach a *Frequentist* QRC either at the component or system level. This QRC - Reliability at Confidence - is *inferred* from *weighted relevant values* and those weights are based on *Subject Matter Expert judgment*.

We assume Gaussian PDFs for now; and we remember the challenging quote (Pao, 2001) "so, how will you get the PDF??" The QRC method can accommodate non-Gaussian PDFs if we ever get one that is quantified, even if it is quantified using Bayesian Likelihood (Booker et al, 2002).

Repeating the examples of Excess Comfort above but expressing them as an "Excess NeLoFaRA": (We will assume that when the QMU and CF research groups said 'Total U', they may have meant a more traditional '2σ U' - so the LFR values below are based on 2-sigma Gaussian; to further simplify we consider N* approaching infinity):

Assuming U_{ru}=0 [extreme of the unspecified treatment of U_{ru} in CF]:
As a base case, Z_{qrc}=1 gives QRC@2σ=.98, Failure Rate FR=.02, LFR=1.6

FOM=B_{qmu}=2 gives QRC@2σ=.999968, FR=.000032, LFR=4.5, so the Excess LFR=2.9 orders of magnitude [nearly **800x**] improvement LFR vs. base case.

CF=.33 gives QRC@2σ=.9986, FR=.0014, LFR=2.8 so the Excess LFR=1.2 orders of magnitude [about **16x**] improvement, LFR vs. base case.

Assuming U_{ru}=20 [analogy to LLNL 6 June 01 FOM/QMU], then:
As a base case, Z_{qrc}=1 again gives QRC@2σ=.98, FR=.02, LFR=1.6.

FOM=B_{qmu}=2 gives QRC@2σ=1.000000, FR=.000000 and LFR>7.4 so the Excess LFR is *nearly 6 orders of magnitude* comfort [about **1 million x**]. Any B_{qmu}>1.75 gives LFR>6 orders of magnitude from failure - about the maximum we ever consider as a system value.

CF=.33 gives QRC@2σ=.9991, FR=.0009, LFR=3.0, so Excess LFR=1.8 orders of magnitude [about **63x**] improvement in LFR vs. base case.

TO SUMMARIZE, relative to a classical $U|_{2\sigma}$ QRC-style analysis:

- QMU/FOM=2 in this example was conservative by 800 to 10^{13} times (3 to 13 digits) in LFR.
- CF=.33 in this example was conservative by 16 to 63 times in LFR (1.2 to 1.8 digits in LFR).
- The choice of QMU s FOM>2 led to extremely conservative LFR s, and the amount of conservatism is strongly dependent on requirement uncertainty U_{ru}.
- The choice of CF>.33 led to less conservatism in LFR, BUT:
- In either case, we didn t have any real world feel for the amount of conservatism until we did the QRC analysis. We suggest going directly to a QRC analysis to begin with.

The "padding" in the chosen "desired" FOM>2 and CF>.33 values show that an *arbitrary* level of excess comfort results during an adequacy assessment using only M and U terms. The only method we see to develop a physical feel for the excess comfort — and to proceed further — is by moving directly to Quantified Reliability at Confidence measures.

Even though QRC enables us to avoid such hunting or fishing due to its direct connection between V&V, UQ, margin, and R at C , we must first establish the gambit of component, environment, and system conditions under which we want to evaluate these quantities; and find a simple [graphical or tabular] way to express these terms and then roll them up to enable us to fold them into QSV, our value engineered investment strategy method.

Figures II-5 and II-6 graphically illustrate some of the advantages of moving directly from Margin M and Uncertainty U into Reliability (at Confidence) as in QRC. Certainly, if we are fortunate enough to have a six sigma assessable reliability, or even 3 or 4 sigma — we might then desire a way to express this excess comfort . The Engineering Index (EI) as proposed by Los Alamos (Booker et al, 2001) does this, and extends QRC into the range where R~1. Translating R into Log Fail Rate, or NeLoFaRa as denoted above and shown in Figure II-6, also conveys information with a physical feel as R approaches O-N-E .

Use of either EI in Figure II-5 or Log Fail Rate Assessment (~LFR) in Figure II-6 can also give us an early warning system in time. Recall equation [1]:

$$\Delta QSV = QSV_0 \int_t PV_F[t]\, \Delta[t]\{\Pi_{i=1,MC/STS}\ (RC^*)_i\}\, dt \qquad [1]$$

However, watching QSV (Value) deteriorate over time will do us little good; we need an early warning system so we know when a negative ΔQSV is *approaching* — with time to plan and do something about it. Since R is changing very little if any during this early warning period, we can make use of a change indicator — process control indicator like EI, or even LFR in Figure II-6, to enable us to plan in advance with an indicator that will more clearly show us when time is running out .

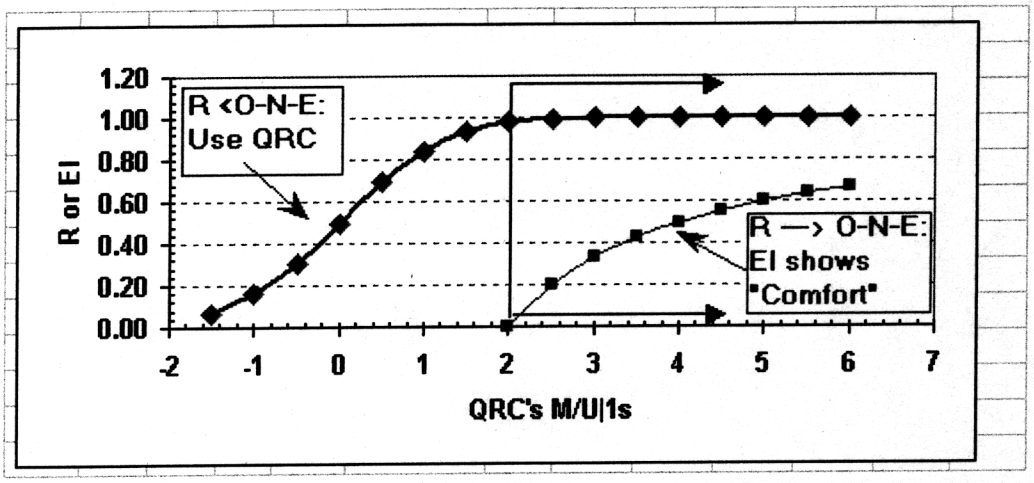


Figure II-5. Relation as used in QRC for Normal PDF s: Reliability R from $Z_{qrc}=M/U|_{N^*,\sigma}$. As R approaches 1, changes lose their meaning — a control parameter such as Los Alamos EI [shown here with a 1-tail $U|_{1\sigma}$] can take over.

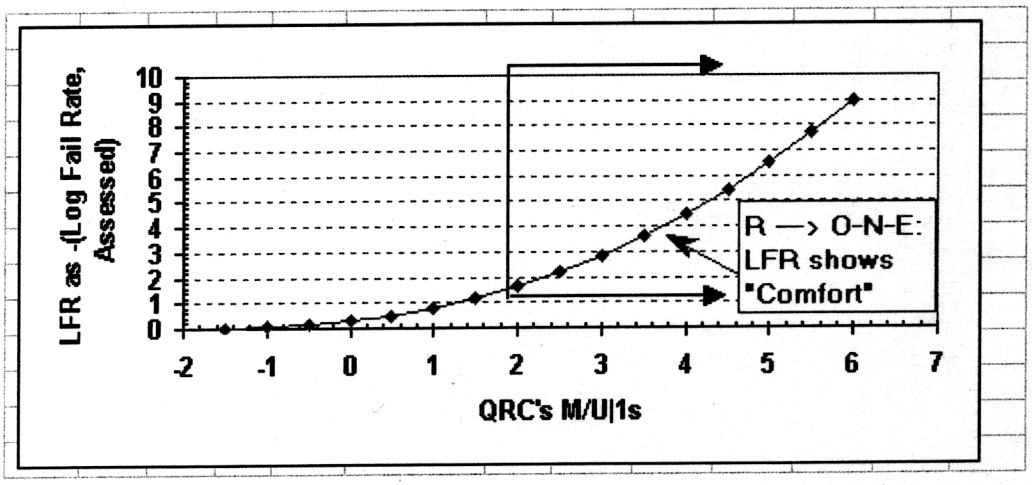


Figure II-6. Either QMU s FOM=M/U , or QRC s $Z_{qrc}=M/U|_{N^*,\sigma}$ — can express Excess Comfort . But, Log Fail Rate is very close in magnitude — and carries a well-known physical meaning.

2.4 Example of Process to a Quantitative Validation Statement

The preceding text and figures have laid the framework to take us through a Quantitative Reliability at Confidence (QRC) analysis. We now provide an example of the use of available data and model analysis results to generate the numbers necessary to make a Quantitative Validation Statement as expressed in our DEFINITIONS subsection of Part I. An example of internal combustion engine power output as a function of exhaust restriction is not exactly our every day mission in nuclear stockpile stewardship, but since the data is handy and the application is unclassified, a quick analysis of such data will show how we run through the methodology. Consider Figure II-7. Our desire is to take some available data regarding power output versus exhaust restriction (translate, noise reduction), and then find the exhaust restriction Margin ; that is, the amount of exhaust flow restriction we can tolerate and still produce the desired power output, say 300 or the value on the plot s x-axis.

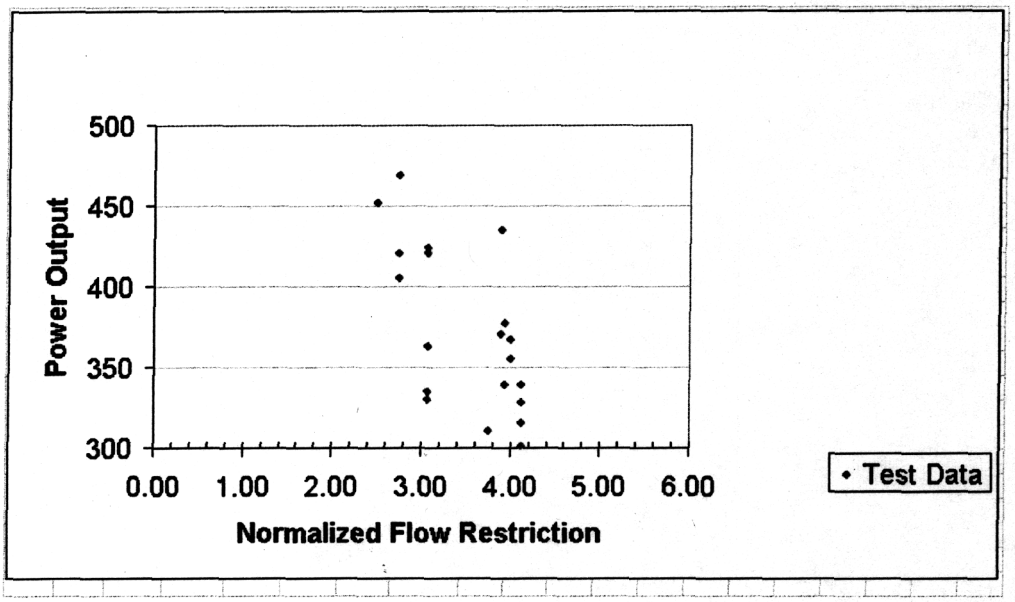


Figure II-7. Available data , taken under mixed conditions of environment and build, for power output as a function of exhaust restriction. Without a *model*, all we have is scatter.

2.4.1 *Model Fit and Uncertainty*

But, our available data just looks like a pattern full of scatter. This is because, as is most complex systems, power output is a function of many things — in this case not just exhaust restriction, but many other factors that were varied from test-to-test. This is not ideal in trying to isolate the effect of exhaust restriction, but it may be all the data we have. We are lucky to have as many as N=29 data points; this relatively large N will help reduce our epistemic uncertainty error E_1. Our goal is to use a suitable model to tell us what the power output would have been, under standard conditions with a standard production build , with exhaust restriction then being

the only variable remaining. Suffice it to say that there is such a model; in this case a closed form model, hence free of weak form, approximation, and spatial-temporal discretization errors, so we could claim $E_3=E_4=0$, but with many potential model errors or physics errors that we would lump into the E_2 term. In addition, this particular model provides a smooth fit (Figure II-8) on normalizing all N=29 data points, but it uses K=6 degrees of freedom in the model fit. These are mostly physically based degrees of freedom; that is, with enough component level test data we could get down to say K=1; nonetheless, since all we have is system level power output, we can only guess what values these component-level model input numbers would have been or might have been, and use them as degrees of freedom to fit out system level model of power output.

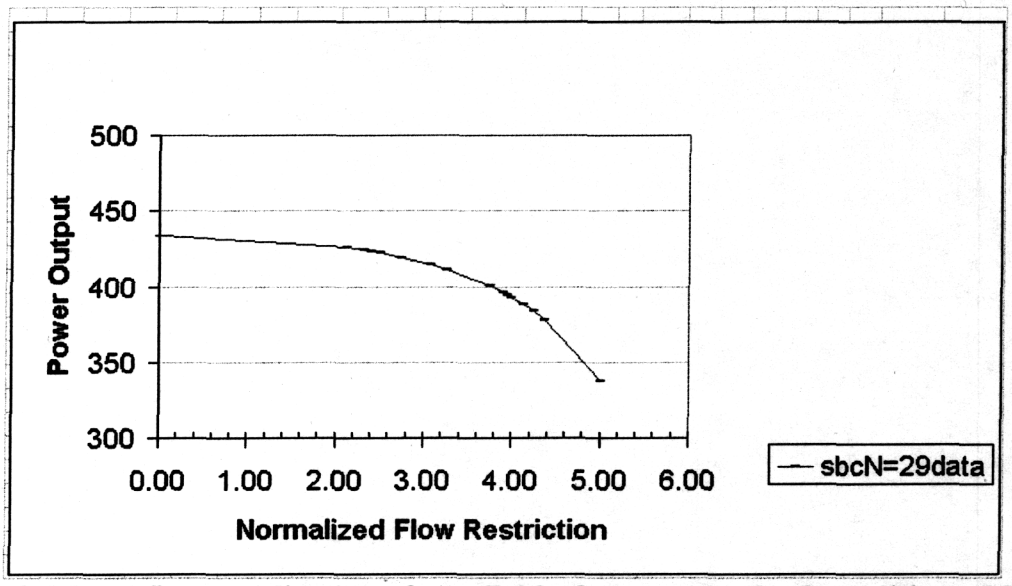


Figure II-8. Model fit normalization of Available data , to standard conditions. Model fit is quite smooth, but we must examine what this means.

Nonetheless, with N=29 and model degrees of freedom K=6, our closed-form model does quite well in providing a smooth curve of output versus restriction. But, is this just a smooth model that looks like it captures the physics well , or is it actually a good model? How do we tell? Some method for UQ, Uncertainty Quantification, must be used to tell us how well our model really did capture the data as measured, else we would have no confidence in using this model to provide a curve for standard condition performance. Figure II-9 provides such a comparison. By showing the difference between the actual mixed condition data available and our model fit to that data, we can establish model error (including bias error as the reader may notice) in our model. We can then, bearing in mind our data set N=29 and degrees of freedom K=6, establish confidence bounds (in this case 2σ) on our model. As we might expect, the confidence bound estimates are broad, in part due to our epistemic uncertainty E_1, and they become more broad as we get more distant from where the actual measurements are clustered.

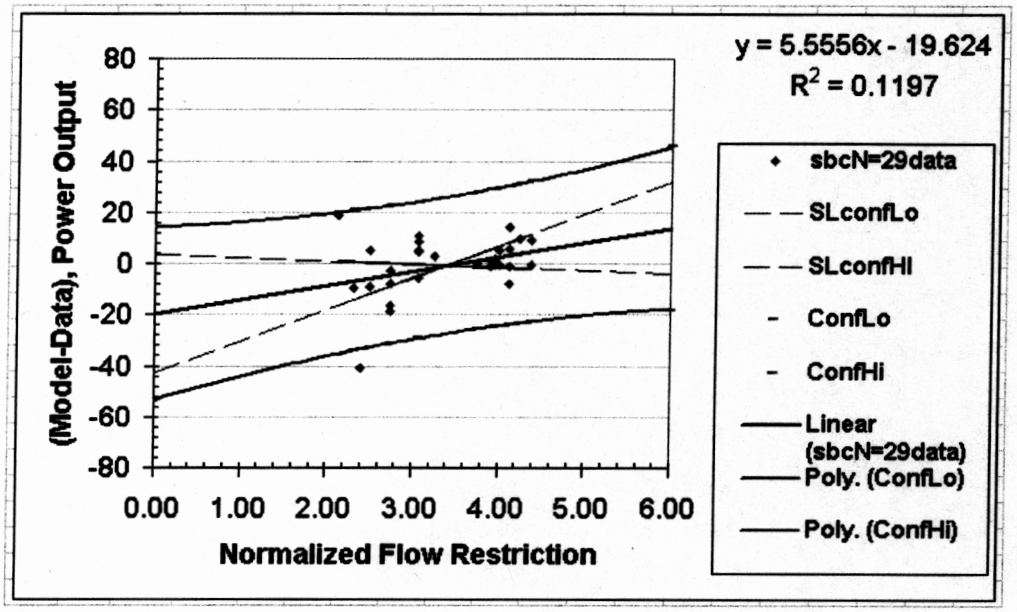


Figure II-9. Model vs. data error (E_2), with estimates of confidence bounds and model bias. Dashed lines are straight-line fits at low and high confidence slope; solid outer lines are confidence bounds [at chosen σ-level] on the model fit.

We now have to translate these confidence bounds back our perfect fit model from Figure II-8. When we do this, Figure II-10 results; this is the model, with model uncertainty error estimates (E_2) at specified confidence levels, for a moderately small (albeit still frequentist) data set.

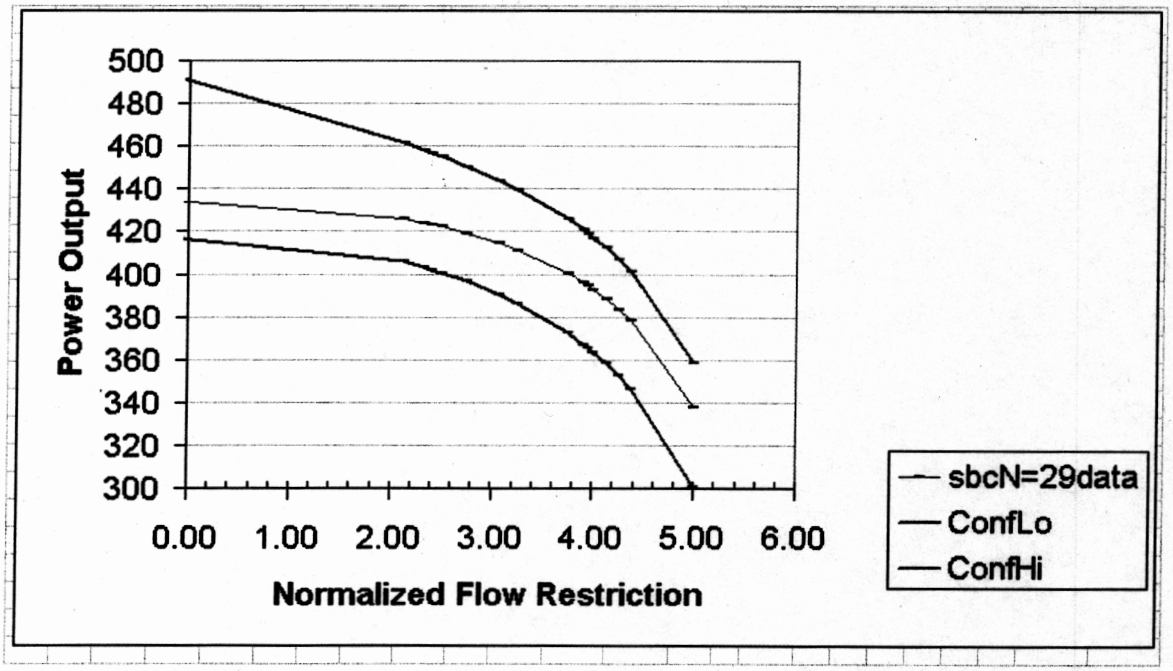


Figure II-10. Model for output versus restriction, with estimates of confidence bounds.

Of course, we can add to Figure II-10 all of our available data . At the risk of being redundant, we point out, as shown in Figure II-11, that even though there was a lot (N=29 points) of available data , it will not line up with our curve for standard condition output vs. restriction. It just means that our model was built from a lot of data (N=29 points). Our confidence in the model is expressed by the green bounding lines. The fact that the raw available data does not fit our smooth standard condition line is in fact an indication that we are extrapolating something — not restriction, but some other parameter affecting output at a given restriction — to get to standard conditions. Or, it may be that one of our raw data points happens to lie right on the standard condition model line; this may mean we did not extrapolate at all, or it may mean that we used our model to extrapolate two non-standard conditions and ended up back on the curve, so it *looks like* we did not extrapolate at all. Such a plot of standard condition model curves overlaid with raw mixed condition data can be very informative, but misleading if taken casually.

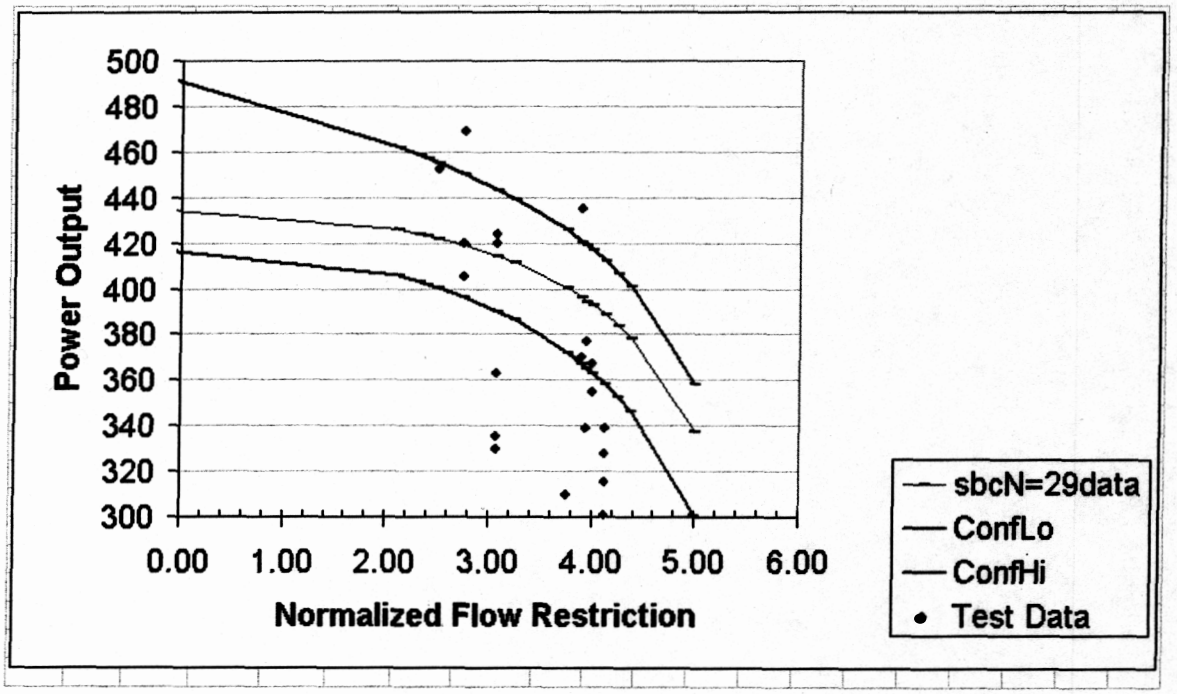


Figure II-11. Model prediction of output vs. restriction at standard conditions, and the mixed-condition raw data from which the model was built.

The methodology and numerics are a minimum desired set for our quantified performance era of stockpile stewardship; this type of analysis should be shown as justification for our claims of high confidence , etc., versus statements like the model is very smooth and behaves as physically expected, and our [sketched in] uncertainty estimate roughly bounds the data. Of course, we do not always have N=29 data points in our mission. Lower N can be offset by using lower K model degrees of freedom, by accepting higher epistemic uncertainty bounds E_1, or by using Inference methods to effectively increase our N* .

As a final step in preparation for a minimal fidelity QRC analysis, we sketch some quantities of interest on our model for output to capture a measure of Margin and Uncertainty, M/U . In this process, we show Figure II-12; on referring back to Figure II-3, we see the similarity in the quantities mid-point M (that is, power capability P_c minus power required P_r), and the relevant model uncertainties (E_1 and E_2), lower on capability U_{cl}, and upper on requirement U_{ru}; we assume U_{ru}=0 in this example (See Eqns. [3] and [5]). Now, if there is no further uncertainty, ie aleatoric uncertainty error E_1=0 and approximation / discretization / solution uncertainty errors E_3=E_4=0 (a good approximation in this closed-form case). Then our reliability R is the number of Confidence-corrected sigmas defined by this specific M/U.

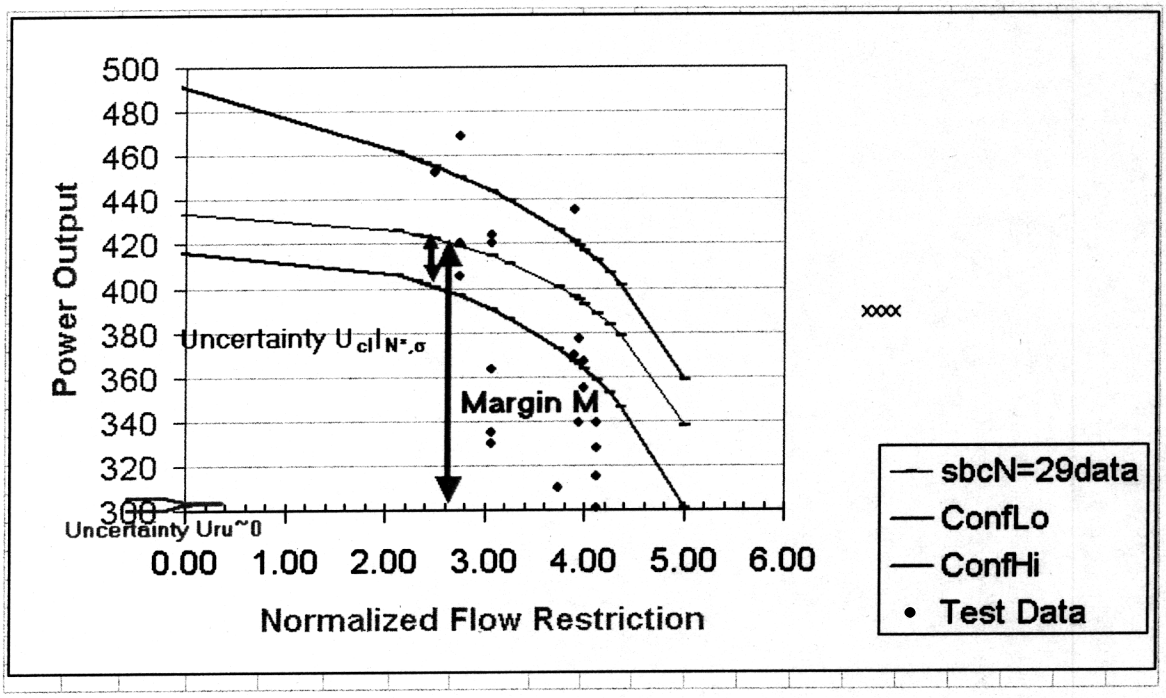


Figure II-12. Final steps on our model fit plot, illustrating the evolution toward a QRC input value.

Now that we have worked our example all the way through Figure II-12, are we ready to continue the QRC process? Are we certain of all our uncertainties? In fact, we have missed a whole set in this limited example. We did not look at the uncertainties in our certain model inputs (that is, the ones that we did not allow to float as degrees of freedom K to match the data.) Notice that we have yet to use the word Validated Model in the narrative of Figures II-7 to II-12. In fact, we will not even get to use that word fully in this example: Since we have used our full available data set together with the K=6 degrees of freedom to fit it, we are going to end this example with a calibrated model that is then validated to a Maximum Likelihood Validation (VAL_{mlv}). We have quite a bit of work to do before our model could claim VAL_{suv}, where each input uncertainty U_{Pj} and sensitivity S_{Cij} are methodically bounded regarding their contribution to

the overall system uncertainty of interest U_{Ci}. The calibration and VAL_{mlv} [to its credit, a fit to 29 data points] may well be the one that is the best fit to all the relevant data. In that sense, we have built the maximum likelihood model. We have quantified our uncertainties, margins, and can extend into QRC, but we did not discuss any instances of agreement [or prediction] of the next data point — or of extrapolation beyond the Validated — we should say here the *Calibrated* — regime. We can get some hint of what we would be up against by examining the sensitivities our model predicts to all the inputs into the power output model that might have uncertainties tied to them; and the bounds of those uncertainties. This issue of examining sensitivities — whether or not they are used as degrees of freedom in model matching — is an essential part of calibration or validation for that matter. Figure II-13 is from another industrial realm, that of sheet metal stamping — and yet it shows as clearly as would any other complex system problem how sensitivities, some strong, some weak, may influence the predictions of our model or the results in reality. Figure II-13 is an example from (Logan and Maker, 1997). The study involves a metal forming and stretching test where a flat sheet of metal is stretched to some height H*. The origins of the test can be found in (Keeler, 1968) and (Hecker, 1974). In this design study, we would use H* as a Margin measure, and explore sensitivity to material and process parameters. Clearly F and B would be first candidates for follow-on physical validation testing. For this discussion, the definitions of F and B are not important. What matters is that the Margin measure H* is far more sensitive to F and B than any other sensitivity in the table. Therefore, we would likely focus our test investment for sensitivity first on F and B . **A *model based* sensitivity table or pictorial of this type should be generated before significant funds are spent on validation component or system tests.** In fact, this same type of information should precede similar magnitude expenditures on the code development side. If we can t show that the lack of a code feature or model would make a difference, we cannot justify spending large amounts of funding be implementing the model. Of course, the issues raised in Figure II-13 lead toward our discussion of capacity issues, resource balancing, and prioritization, that is, to Part III of this work. There is a balance between effort spent on the 10-fold to 100-fold computational effort for complex models in order to find and quantify model sensitivities (with perhaps some crude code features to represent them), and the effort spent coding those physics properly if justified. These code and model efforts must also balance against effort in performing the physical validation testing to see if our model predictions of large sensitivity actually come true, and what range the sensitivity variables may take in production and in real world scenarios. (See the VAL_{suv} criterion in Part I s 10-Point Summary; we have to examine the ranges of

$$U_{Ci}=RSS(S_{Cij}U_{Pj}) \qquad [3f]$$

to go beyond our maximum likelihood fit and look at the bounds of less likely fits.)

2.4.2 *Confidence Bounds on Predictive Extrapolations*

We can only quantitatively assess our predictive confidence if we have quantified our validation statements. Even so, with model based estimates for uncertainties as in Eqn. [3f], we face another challenge as depicted in the V&V d Region Cloud illustrated notionally in Figure I-9. That is,

we may have a good VAL_{mlv} fit to the integral system level data set. And we have trusted our model s values for sensitivities S_{Cij}. But, as can be seen on the right hand side of the Cloud depiction in Figure I-9, we will face cases of extrapolation outside the V&V d region where our model VAL_{mlv} fit is good, but where in local areas we doubt our model inferred value for a given sensitivity S_{Cij}. Ideally, we would have (or obtain) component level validation data to validate (with quantified uncertainties) not just the integral fits as in Figs. II-7 to II-12, but to validate the S_{Cij}. In fact, some of this information is embedded in the integral data itself; that is, quantified trajectory estimates of any S_{Cij} and bias error due to model S_{Cij} vs. test-inferred S_{Cij} can be obtained by inverse analysis of the integral data or the integral model results. If the S_{Cij} are nonlinear, these local sensitivity trajectory estimates will be neither universally accurate nor unique. However, since our VAL_{mlv} model contends to match the same integral data points (and hence the same trajectory sensitivities), we can perform a quantitative validation on these model trajectory S_{Cij} s in the integral sense by comparing integral data to integral model results. This can lead to quantitative validation statements for the model sensitivities S_{Cij} of the form:

> *We are ___% confident that the model estimated prediction sensitivity S_{Cij} is equal to the observed S_{Cij}, +__% or -__%*

Whether the *system* and *quantity of interest* at hand is the power production from internal combustion, the forming ability of metal sheet, or the performance of strategic nuclear systems, it is important first of all to do two essential steps: First, we must break the system into components, functions, and environments. Then, we must determine (eventually in priority order) which quantities of interest we want to calibrate — and then validate. Only then can we carry through extensions of the simple example process outlined here. It is only possible to validate (and hence enter into a QRC analysis) the specific combination of a **code, model, user, platform, system conditions, and quantities of interest**. It is, in a word, **impossible to validate a code** per se, and nearly equally **impossible to fully validate a model** — specifics are needed. We can draw some inferences about other conditions, quantities, etc — but then again our process already involves more inferences than we would like.

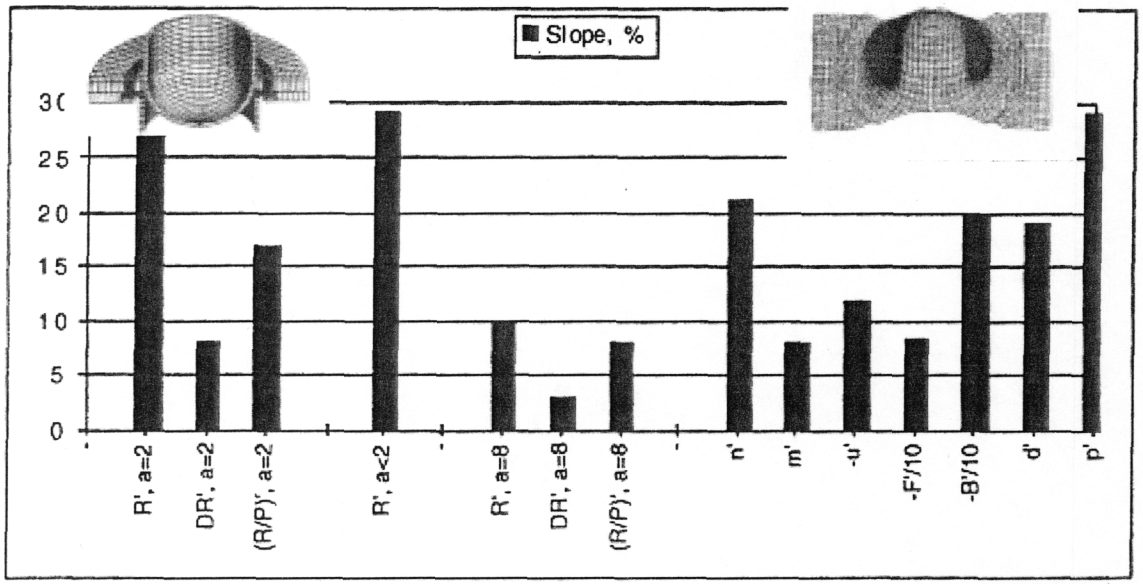
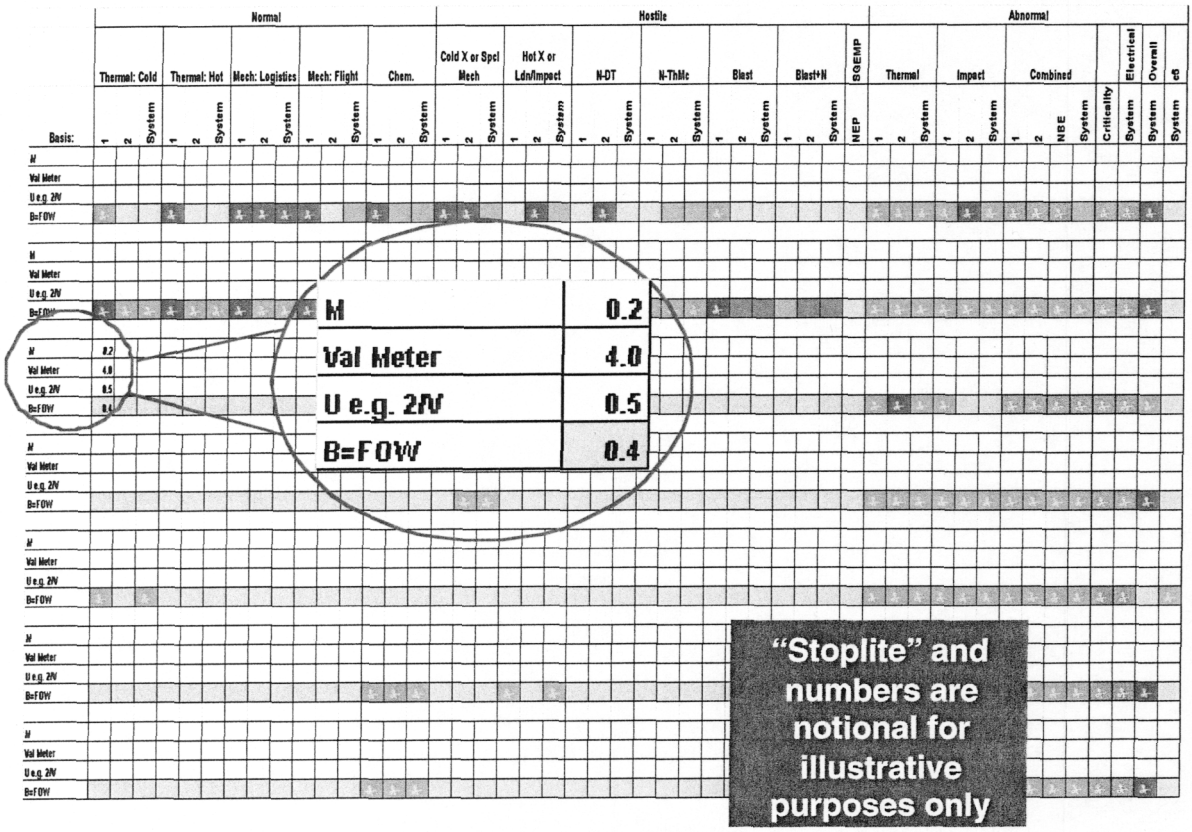


Figure II-13. Sensitivity study of maximum metal forming height H* to process and material variables, e.g. R, a, n, m, F, B (Logan and Maker, 1997). Comprehensive sensitivity studies, of the sort *not* considered in the simple UQ / QRC example of Figure II-7 to II-12, can increase 10-100 fold the capacity needs in the ASCI model realm.

2.5 The QRC Matrix: A first step toward *prioritization*

Before moving to Part III, it is instructive to ponder at high level the scope of the problem we face in deployment of the QRC method across all our systems and environments. Figure II-14 is an example of a notional first-level breakdown of the component and environment level RC* values that must be assessed (meaning validation, UQ, margin assessment, frequentist or inferred N* and confidence bound estimates, and QRC values) across the system environments for our weapons systems should they be called upon to do their job.

The advantage of such a matrix is that it lets us begin to choose, by the shade of red in each cell, where to focus our effort. However, timing will be an issue as noted in Part IIIa below; certain cells that are green may be turning yellow at a more rapid pace. And, the Benefit/Cost of our effort on those Red or Orange cells must be measured. In Part IIIb we will provide a method to quantitatively answer questions like what if we turn a cell from yellow to green; what is the *value* of doing so? And of course: what is the *cost* of doing so? We will measure both *value* (risk avoidance, consequence) and *cost* in dollars.



RWL and CKN in CODTU-2001-0266 pg 40

Figure II-14. The Matrix of system environments and values for M, Validation level, U at C and N*, Z_{qrc}, and ultimately R and C from Z_{qrc}. A similar diagram exists for physics performance. Priority is to work the red cells first, then orange, etc: But to work on the low-hanging fruit first, i.e. where Benefit/Cost Ratio is highest as computed in Part III.

We refer to Figure II-14 as The Matrix of System Environments for components: it is very easy to imagine being lost or disoriented in The Matrix. The Matrix for System Environments is a chart we generate with the following procedure. Notionally, it is simple. In reality, a credible matrix is a multi-year challenge for ASCI and stewardship as a whole, as it is for any assemblage of single or multiple complex systems (weapon systems in this instance). The Matrix as shown contains a series of 6 rows (4 shown) for each system under consideration. Across the rows, each column represents an environment to be considered as a requirement. Validated analyses provide a margin M for survival or margin to design standards. Validation level for that analysis is entered from the VAL Meter, e.g. V=0 to V=10. The uncertainty in that M is expressed as U ; U~2/V if no more rigorous estimate is known. We certainly do not recommend estimating U~2/V from a VAL meter reading V=0 to V=10 when that estimate of V is simply a judgment call anyway! However, for a screening study it is a good place to start; it *provokes improvement*. We certainly must go beyond U~2/V , because we have already stressed that U only has meaning if stated at a confidence C . Our QMU-CF-QRC Excess Comfort example above showed the pitfalls of not expressing U at C . The 4th row, the one with the stoplight colors,

is $Z_{qrc}=M/U|_C$, color coded in stoplight form. The Matrix allows us to focus on particular areas where our analysis shows behavior at or below some critical design criteria. In the layout shown here red does not mean Failure - it is simply a way to denote that we have passed below a design criterion so more attention is justified. The concept of The Matrix row and column meanings and stoplight colors was first introduced for the STS in (Nitta and Logan, 2000). The importance of tracking changes with time in this respect was noted there as well. The 5th and 6th rows, not shown, represent the quantified estimates of Reliability R at Confidence C, obtained from the V&V / UQ information that is shown. It is this R at C (the QRC method) that enables the RC Rollup that we must do to Figure II-14, as depicted in the overall QRC/QSV flow chart of Figure II-1. Such an RC Rollup is not unlike the reliability approach taken for the non-nuclear system components of our weapons systems (SAND, 1993); except that here the RC Rollup and the methods leading to it by no means resemble a frequentist approach.

We use QRC s $Z=M/U|_{x\sigma,N*}$ as depicted in Figure II-3 to obtain, using Frequentist, Bayesian Relevance Factor (BRF), or other methods, values for C and R entered into The Matrix in the nuclear package for STS environments. These values of C and R equivalents are products — a way to numerically roll up the state of the stockpile. Targets of opportunity for stewardship are then where the particular environment R_i, C_i, Z_i, are lower than we would like, which may mean that our *assessed* uncertainty U_i is higher than what we would like. The phrase what we would like is another layer in our decision process, that of adequacy assessment, resources, and investment strategy, the topics of Part III of this work. The use of V&V, UQ, QRC, and QSV (Part III of this work) together is what enables such a decision method with numerical closure; with many caveats to be sure, but with demonstrable closure. This additional layer — of how good is good enough enables us to get to the value of our activities. We have called this methodology part of our QSV (Quantitative Stockpile Value) method since circa 1997. It is, like the CRUMBS algorithm and QRC leading to the RC Rollup, a circle that closes — we note the development of similar methods at Sandia (Trucano et al, 2001). Others at Los Alamos have independently developed a similar way of thinking, with notable progress on expressing these quantities over the time domain from the EI group (Booker et al. 2001).

For the RC Rollup, many of the entries in Figure II-14 are multiplied in series, some in parallel. Some notionally represent safety factor to yield values — others leak before break type logic — but all things considered, we note first that the scope — especially cross-system — is quite large. As the time from our legacy evaluations increases, we must quantify the entries in this matrix with new tools. Then, we must optimize the balance of green — that is, where to put our computational and experimental effort to maximize our overall return on investment which is directly proportional to the Rollup of RC*.

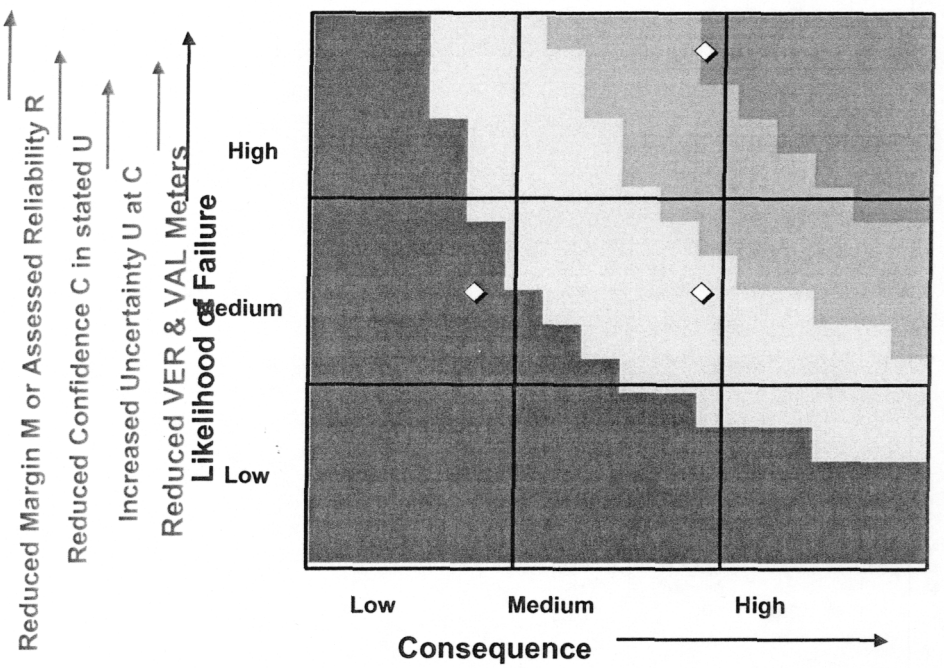
It is this RC Rollup that is used directly in our expression of value or change in value, which is described in Part III, the next part of this work. The most compact form we have found to express this value is:

$$\Delta QSV = QSV_0 \int_t PV_F[t]\, \Delta[t]\{\Pi_{i=1,MC/STS}\ (RC^*)_i\}\, dt \qquad [1]$$

This equation, optionally using Present Value PV as we recommend, enables direct and quantified Benefit/Cost tradeoffs and timing decisions (Logan and Nitta, 2000). Note the term, $\{\Pi_{i=1,MC/STS} (RC^*)_i\}$, which in fact represents the RC Rollup that we just described above.

2.6 V&V to QRC: Relation to Risk=Likelihood*Consequence

Figure II-16 adds some quantification to our relation of V&V-QRC-Investment Strategy to the Risk=Likelihood*Consequence diagram. Part I gave us a qualitative (meter reading) feel for V&V and hence inverse of risk likelihood. In Part II, we have outlined ways to test the validation status of our model (and relevant data) to quantify our uncertainty (U) at confidence [C] while calculating a given system or component margin (M). In Part III, we will add more terms to enable our entire quantitative method to be related to this Risk=Likelihood*Consequence diagram.

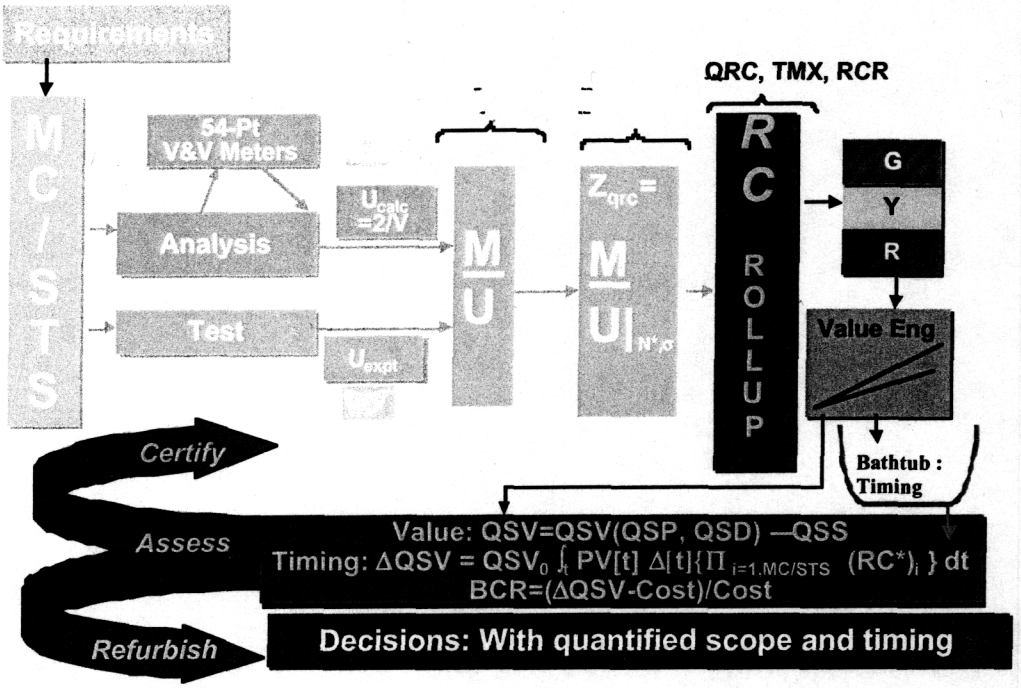


56

Figure II-16. Risk=Likelihood*Consequence Matrix, with relevance to Part II: From V&V to QRC.

3 PART III: Deployment, Resources, Value Engineering, and Investment Strategy

Better has become the enemy of Good Enough
- Anon, 17 May 1999



.44

Figure III-1. Flow diagram for Part III: The RC Rollup terms from Part II are used with time (t) at Present Value (PV) and cost inputs to develop a Quantitative Systems Value tradeoff equation ΔQSV. QSV is made up of QSP (P=Performance), QSD (D=Diversity), and QSS (S=Safety). The Benefit / Cost Ratio (BCR) is much like a Value Return On Investment ROI .

Part III offers a quantitative investment strategy that uses the V&V d products of Part I (Models with Uncertainty at Confidence) to determine factors of safety, or Margins as we refer to them, and use these in the QRC method of Part II. These terms fold into our QSV, Quantitative Systems Value (or Quantitative *Stockpile* Value for our application) equation:

$$\Delta QSV = QSV_0 \int_t PV_F[t]\, \Delta[t]\{\Pi_{i=1,MC/STS}\ (RC^*)_i\}\, dt \qquad [1]$$

This equation enables quantitative investment strategy inputs for timing and resource allocation tradeoffs. A qualitative example of the resource allocation struggle is given in Part IIIa, regarding ASCI compute platform needs for projected fulfillment of all the work we have thrust upon ourselves in Part I and Part II. The investment strategy we can use to justify the resources and timing for these platforms is given in Part IIIb.

3.1 Part IIIa: Priority and Timing: Deployment and Resources

Part IIIa shows the resource struggle we face in attempting to deploy the ability to develop and verify our ASCI codes, validate the models, and use these models to fill in The Matrix of Part II (Figure II-14), enabling the Quantitative Reliability at Confidence Rollup. Part IIIb shows us how the use of that QRC method couples into an Investment Strategy that can help us quantify tough choices of priority and timing.

A veritable mountain of work has been described in Part I and Part II. In any complex system business endeavor, the V&V, UQ, and QRC described above are of value for design, timeliness, and cost-effectiveness. For our unique task of stewardship of the enduring nuclear stockpile in an era with no nuclear testing and minimal designing and production activity, the task is also great. We face a further risk of our own kind of bathtub effect, as depicted in Figure III-2. The bathtub typically depicts both growing pains and aging to end of life of a production run of a complex system. Such a plot of potential defects, or investigations thereof, typically takes a bathtub shape; the development and production of a system has numerous growing pains early on, then a stable, low-trouble life, then indications of deterioration near its end of useful life. Even though our nuclear stockpile is enduring quite well and falls in the flat and stable bottom of the bathtub, we must demand not only that our stockpile be *doing* well, but that we have the people, codes, code platforms, models, and test validation capability to *assess quantitatively* its high state of wellness. The methods of rigorous V&V, UQ, and QRC are evidence of such an assessment; they demand not only product performance but performance of people and tools to assess and document that continued assured product performance. This leads us into a discussion of the continual balance of demands of resources and priorities — and of the use of perhaps the VER and VAL meters, not just to quantify increasing Meter readings, but perhaps decreasing meter readings — an indicator in a given area of a decline, not in the actual state of the stockpile, but in our state of quality or quantity of tools or even people expertise to do a quantitative assessment. If, as we were to show each of our analysis-based assessments, a small pair of VER and VAL meters were placed in the corner of the slide showing the analysis results, the viewer would have at least a judgment based feeling for not just the level of margin or uncertainty of the component undergoing the finite element analysis, but of the state of that analysis — the model, the analyst, the code, the platforms, the tools, the ability for validation testing. Such a meter might well be lower several years hence, should e.g. an individual retire, or a test capability be lost, etc; even though code capability per se might be stable or increasing. Such use of the V&V Meters might well help us to express the people and capability bathtub that is just as real as the actual product performance lifetime bathtub. In concept, Science Based Stockpile Stewardship, in achieving its Earned Value Milestones [perhaps quantified with statistical validation statements] will continually show the quantitative assessment of our ability to push back the far end of the bathtub and help us live in the CTBT era and maintain a credible assessment of our stockpile.

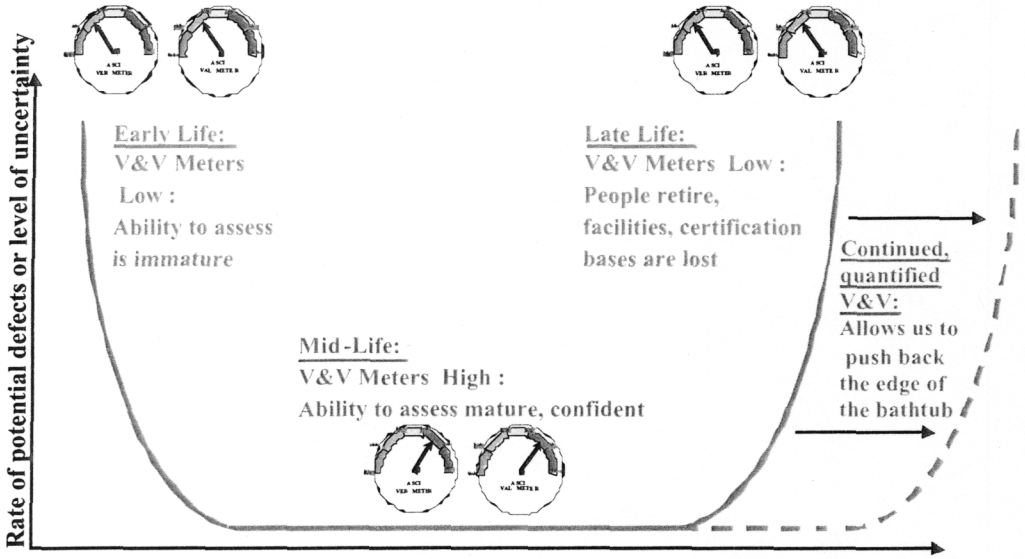


Figure III-2. Bathtub , for product performance over a lifetime, or for people / tools combined ability to *assess* that performance. The latter can be depicted by V&V Meter readings that might well fall as well as rise over time, or more exactly by color and value of the QRC components in the matrices like Figure II-14.

It is true that such a highly abbreviated indicator of V&V fidelity like a single pair of Meter readings can be misleading; and yet we would contend, having attended many fast-paced decision level meetings, that a quantitative V&V indicator is better than no indicator at all. The preceding text and referenced works tell a well-known story of many of the pitfalls that have always accompanied finite element analysis, and nonlinear analysis in particular. And yet, there is not time to address practically with absolute quality all of the issues mentioned above. We recognize that the pace of programmatic business decisions required do not allow full presentation of the rigor discussed in Part I and Part II of this work and the related references by others in the field. We do however feel that as a minimum, some indicator of the pedigree of both the code (Verification) and the model (Validation) are appropriate to be shown on any such color slide used for decision-level purposes. As an extra measure, the level of review obtained on the analysis viewgraph package and accompanying *document* (corporate report, reviewed literature, etc.) might be indicated on the V&V Meters. Many times it will be adequate and most of the time it will be quite accurate for an individual analyst with their typical modesty to estimate their own V&V Meter readings. It is likely they will not be too far off if scrutinized by successive review groups. In most instances an individual veteran analysts opinion of the Meter settings will be enough for this cursory but quantitative and communicative assessment of V&V status. One of the most thought-provoking means in reviewing and pondering the V&V Meter

level of a given work is to document it in a coherent manner. It is not a new revelation to state that an attempt to create coherent words to describe one s work and conclusions will in itself illustrate the level of confidence in the work and its shortcomings. In Figure III-3, we use the V&V Meter readings to attempt to show graphically but quantitatively the difference between demonstration and certification model fidelity. The Meters (or simply their VER or VAL readings — or more ideally the measure of Uncertainty U quantified as in Part II) can graphically illustrate a perception gap that may occur between the expectations of code development and use for product design, qualification, and certification. This perception gap, in our opinion, holds a higher risk to the system and decision process if a demonstration run has gratuitous resemblance to the real system . The result is that even though a code development capability demonstration is only [rightfully] demonstrating about V=2, the viewing audience, due to severe time constraints and brilliant color, may think they are seeing V=6 or even V=8!

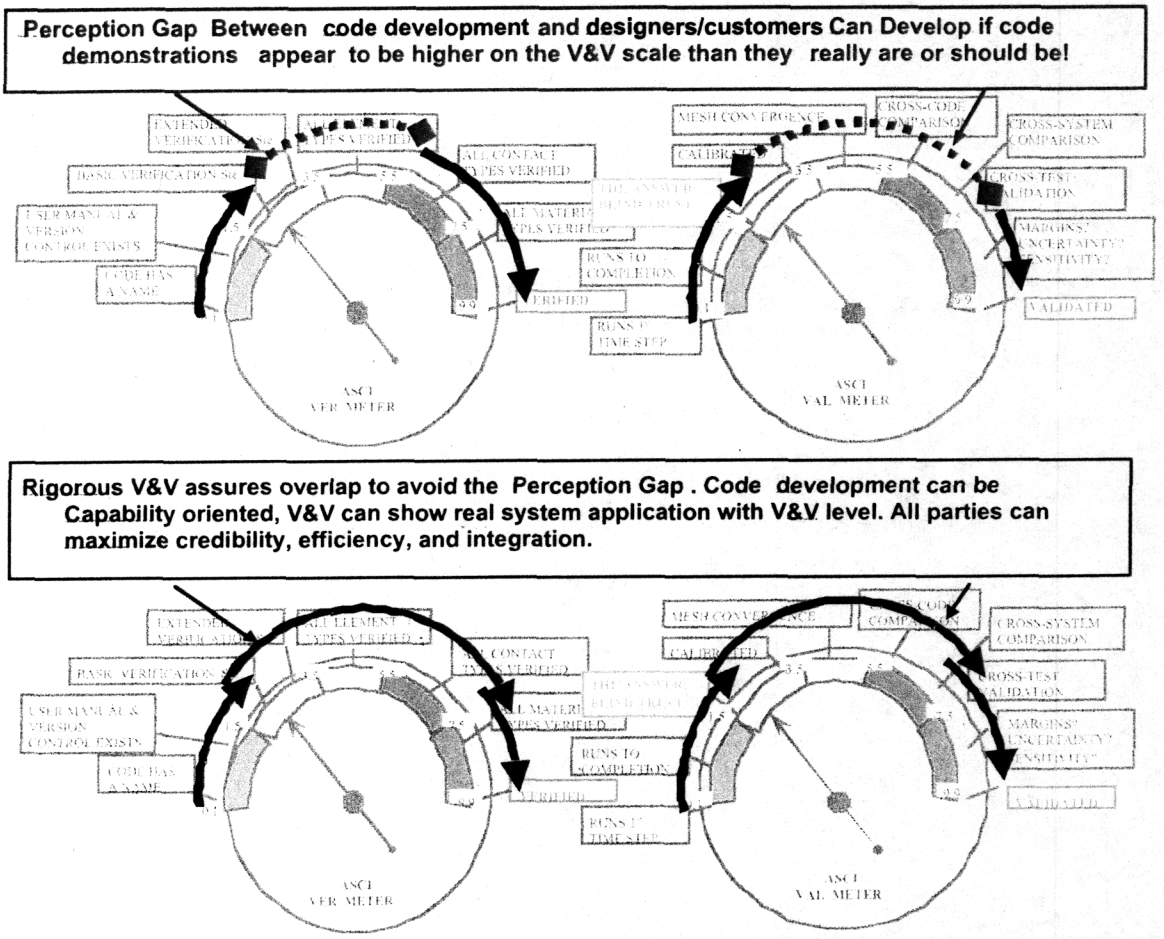


Figure III-3 Danger of perception disconnects when movies are made too soon on real systems versus demonstration and true *verification* of code and solution. The V&V Meters illustrate this gap, and show the role of the V&V process in bridging the gap.

On delivery and use, the design community will realize that much more time was needed for the level of VER and VAL they were anticipating from the realistic movie shown as the demonstration. But by that time, the planning process and timelines are already interdependent and difficult to remedy.

A robust V&V program should be intentionally set up so as to bridge, even overlap, the gap between code development delivery of a demonstration capability and designer/customer expectations of a model ready for certification quality analysis. It takes quantified, documented V&V to take code / model combinations from V=2 to V=8. In this manner, code development can be *capability* oriented, and V&V can show *real system application* with V&V level and quantification. Both entities can maximize credibility, efficiency, and integration.

3.1.1 Schedule and Timing: The Ladder Chart

How will we arrive at the day where each model and its results will not only exist but also have some measure of V&V Meter readings on the slide? The process we suggest is in the Ladder concept (Figure III-4) for linking the products from code development through V&V (with theory and validation tests) and then on to designer/customer use for conclusions regarding the actual systems in the product line.

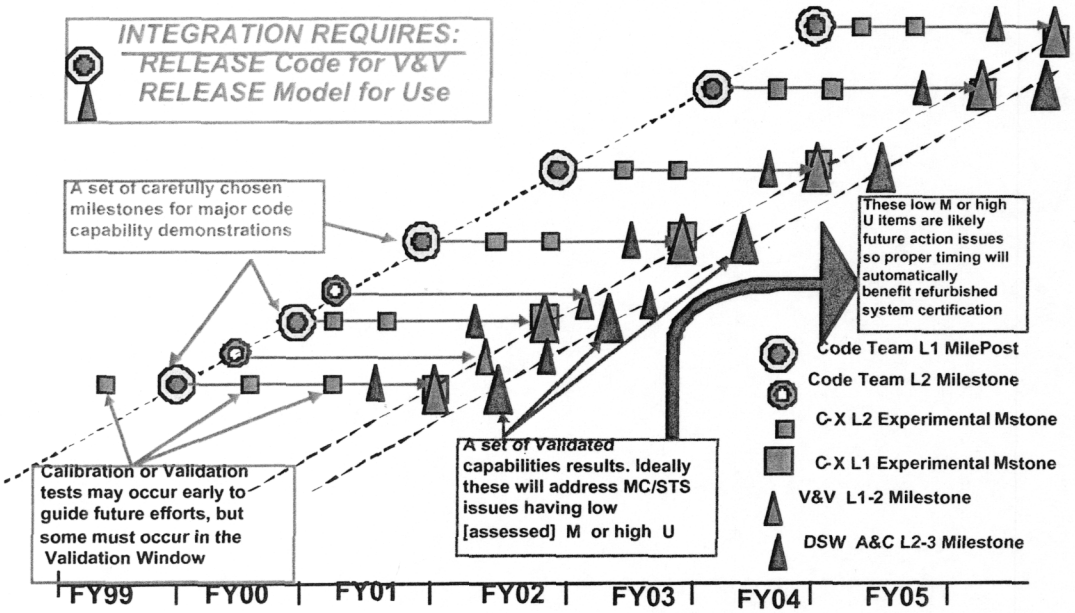


Figure III-4. Ladder Integration can be used with QFD or other methods for timing. QFD can be used within the context of this work s Part II and Part IIIb methods (QRC/QSV). Code development demonstration is followed approximately 1-2 years later by a V&V Milestone statement of quantification. The 1-2 year time frame will vary depending on the level of risk, consequence, and uncertainty needed to fill the mission of the code/model product. This time is needed for validation testing, theory, code feedback, and uncertainty quantification to establish the levels of V&V.

The interval between code development Demonstration Milestones and corresponding V&V Milestones has been discussed in the ASCI forum for quite some time. A graphical representation of this, known as the Ladder Chart, was first shown formally by one of the authors [ckn] two years ago. It allows alignment and timing between code and material model development, V&V, and delivery of V&V d products to the design community. The Ladder Chart concept aligns the Up and Right plans (Ladder Rails) of these entities with Ladder Rungs connecting them. Along a Ladder Rung connecting code development and V&V, there are typically validation tests provided by component and even full system test partnerships; the Ladder Rung activity is essential for Validation in quantitative and predictive mode. Usually, a short time after one gains the basic concepts of the ladder chart, attention turns to the rail of delivery; the system design Directed Stockpile Work (DSW) Rail. This brings a cold realization of just how much focus is needed to align the integrated milestones to meet the goals of stewardship. There is a constant negotiation and prioritization, either implicitly in our minds or explicitly quantified, of what it takes to maximize our confidence in stockpile certification. Continued certification also requires an assessment of how a given stockpile system is doing or may continue to do as requirements change, components are replaced, and as aging sets in. The situation is so complex and of necessity balanced that there is of necessity an equal partnership between ASCI, the Experimental Campaigns, and Stockpile System (WXX) design assessments (or in general the analogous elements in any business plan) that determines the most effective priority for maximizing the progress of ASCI and the Campaigns while also maximizing our efficiency in addressing the ever present and ever growing Watch List — the high demand yellow/orange/red cells derived from The Matrix of Figure II-14 - with maximum confidence. This Triad of Design Assessment, V&V, and the relevant theory and validation data campaigns leads to an expanded set which should encompass the entire business plan for research and development (R&D). This R&D set must of course be balanced with production and refurbishment investments and timelines — and of course ideally those timelines would be justified as a balance of quantified need (e.g. The Matrix of QRC and then value from QSV), and best use of available production plant capability and capacity. As these partnerships are expanded, they soon encompass, with the credibility of the simple V&V Meter Summary or the more complex V&V-UQ-QRC process, and the integration of the Ladder Chart, the ability to Assess, Certify, and as needed Refurbish the product line selectively to lead us credibly into the future. (In the case of nuclear stockpile stewardship, this credible combination provides a path into our CTBT future). How do we align the Ladder? We will give an example of this alignment below, with a notional example of its effect on maximizing the sharpness of our knowledge of reliability and confidence in the stockpile. The method we have chosen for this is a form of QFD, Quality Function Deployment (Kogure and Akao, 1983).

3.1.2 Examples of "Ladder Success"

We have already used the Ladder concept, from code development through V&V to full system application and stockpile certification, as depicted in the actual stockpile examples shown in Figure III-5. (Each WXX represents an actual weapon system in the U.S. stockpile and actual refurbishment and certification events). However, our goal today is to have a more coherent, quantified, defensible linkage to make the right tradeoffs as stewardship goes forward with pressures from many directions.

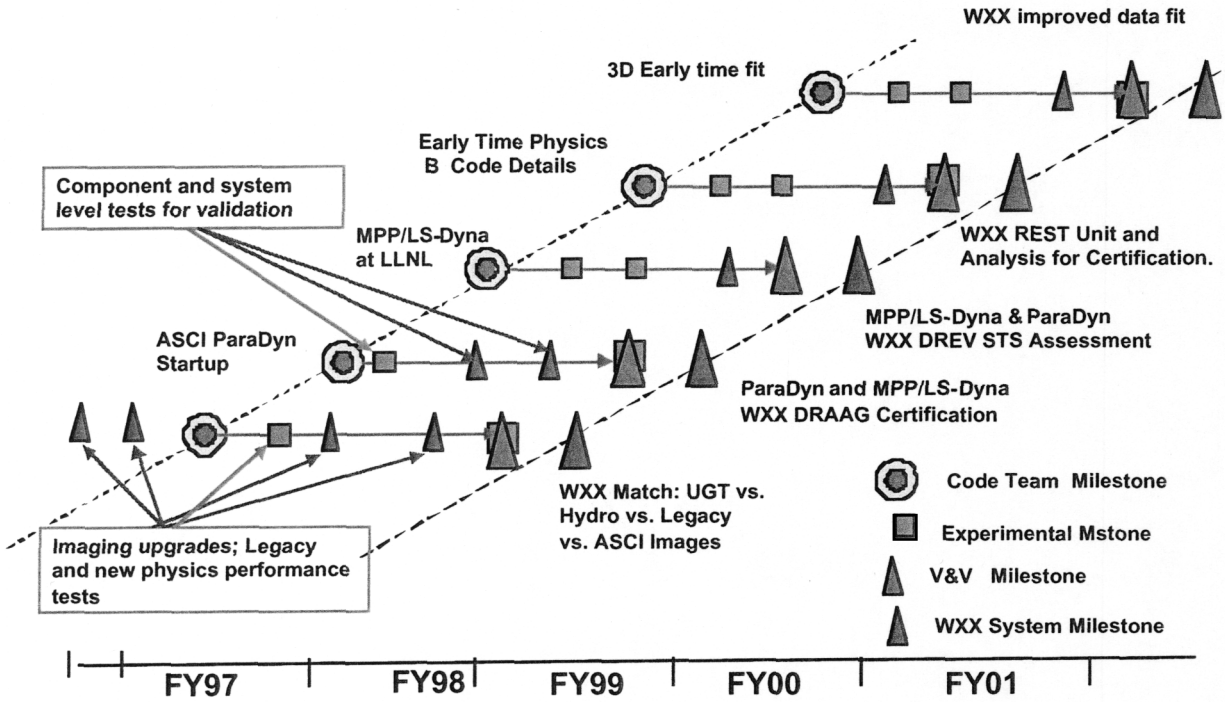


Figure III-5. Integration Ladder examples: Success stories from code development through V&V through [quantified!] product qualification certification.

3.1.3 Two Ladders – Branches – and a Trellis

One of the pressures in defending stewardship resource allocation demands is of course in timing of facilities and capabilities that seem years distant from directly impacting the stockpile. As an example method to quantify the defense of these items and determine their best timing, we have suggested the addition of a second ladder , going from Platforms (faster machines) to Development, next to the ladder going from Development to V&V. And of course, as we have already described, the impacts to stockpile reliability and confidence do not really come as Milestones , they come as sharpening of the numbers and colors in The Matrix, the huge array of responses across the systems and their components in all the different operational and logistics environments. Figure III-6 shows this additional ladder (the ladder rails between platforms — the long lead items so essential to the amount of V&V and UQ we can plan to accomplish), and the trail that branches into the trellis of cells in The Matrix of M, U, R, and C which must be quantified using a rigorous V&V-UQ method like an extension of the example shown in Part II (Figures II-7 to II-12) of this work. The credibility of the colors in the trellis matrix can be traced back all the way to judicious selection and timing of V&V, code development, and platform capacity and capability. These colors represent values of Z_{qrc} (Margin-per-Uncertainty-at-Confidence), and R@C Reliability at Confidence. Where rigorous validation is lacking, we must accept legacy estimates of these values of make judgment calls about them. With rigorous validation (and adequate codes, platforms, and data to validate), we can quantitatively show how we arrived at each of the colors (values) in the cells.

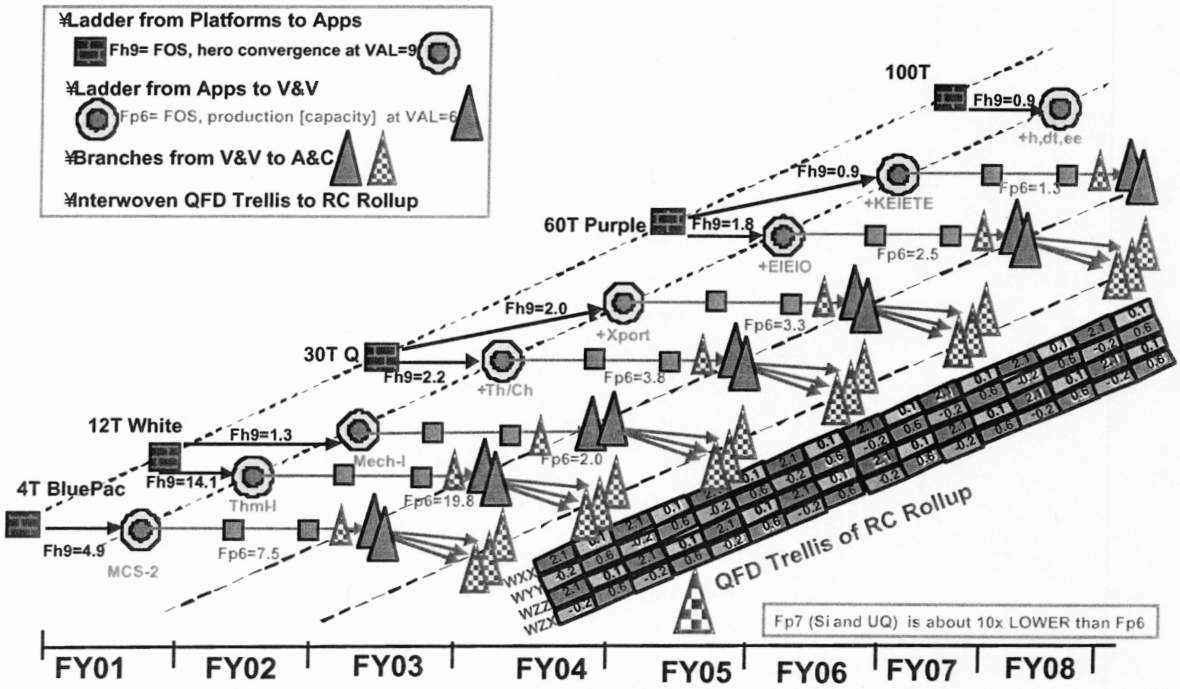


Figure III-6. Integration Ladder : Two Ladders, Branches, and a Trellis of cells of component and system environment responses, each cell having a QRC . Investment and timing in all these elements should be balanced carefully and quantitatively; hence the need for rigorous, clear and reasonable V&V and UQ. This balance involves evaluations of the complex nomenclature describing the linkage of each term to the others; see Table III-1.

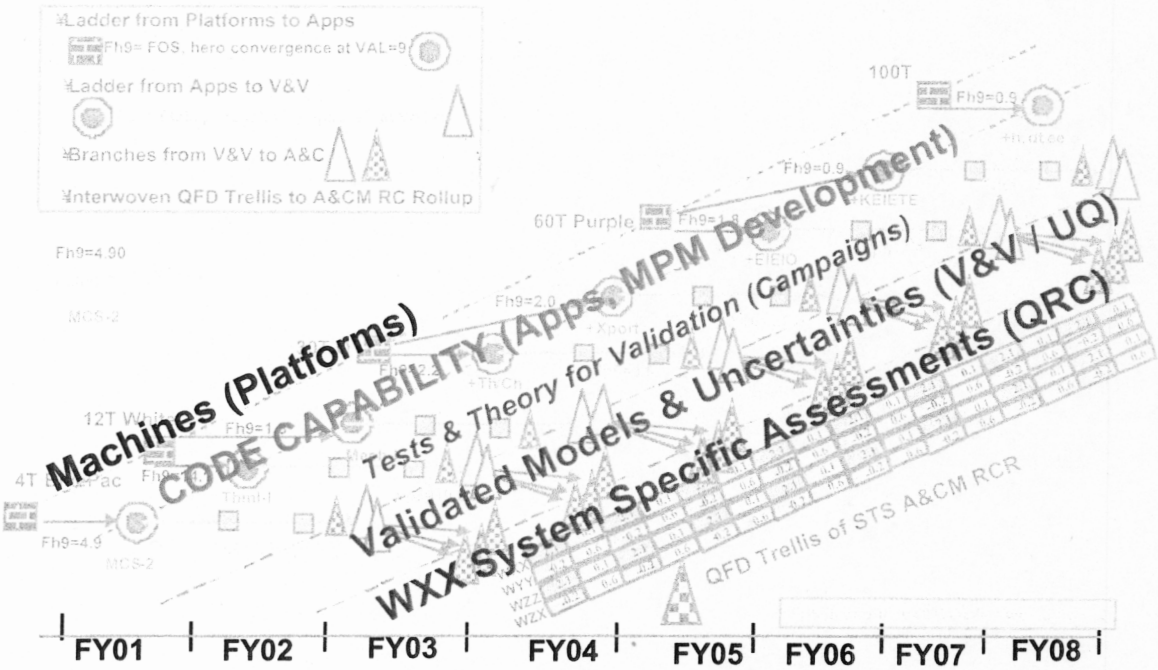


Figure III-7. Simplified flow of the detailed Ladder Chart.

3.1.4 How V&V Level determines need for machine capability, capacity

The 2 Ladders, Branches, and Trellis concept and illustration (Figures III-6 to III-7) was generated and linked to allow us to evaluate our machine capability (peak Teraflops needed during a given run) and capacity (Teraflop rate needed to get all our runs done on schedule) needs. We introduce Factors of Safety (F) as the ratio of available machine capability or capacity (Tflops) to needed capability or capacity. In Figure III-6, we show a notional set of mechanics and physics based milestones (not the actual ASCI Milestone set but a capability-based set similar to them) to help us illustrate the process — and estimate the actual condition — of our machine (platform) capability and capacity plans vs. needs in the coming decade. We used the engineering-oriented set of code capabilities shown in Table III-1 and generated the following factors F for hero h [capability] runs in code development and verification, and production p [capacity] runs for V&V of the level through approximately Figures II-7 to II-12; and on a single system. We will not be able to claim a validated capability across the entire product line (stockpile and environments) without cross-system V&V; so the resulting V&V excess capacity factors Fp will be easily devoured — let alone if we added the Sensitivity and comprehensive Uncertainty Quantification (UQ) of Figure II-13. Code speed and even machine efficiency will indeed matter if our machine capacity is limited even as planned for ASCI. With those caveats, Table III-2 shows the factors of capacity Fh and Fp as we approximated them.

Table III-1. Code feature nomenclature in the capability/capacity ladder:

(The associated V&V applications scale as the level of problem and coupling we can handle grows):
MCS-2: Multiple Code Shootout II (Sam et al, 2001); MPP explicit Lagrangian dynamics
Thml-I: Capability for Thermal Implicit MPP
Mech-I: Capability for Mechanics Implicit MPP
+Th/Ch: Thermal and Chemical kinetics coupled into Mechanics Implicit, MPP
+Xport: Species transport (diffusion, advection) coupled in, MPP
+EIEIO: Explicit / Implicit switching, MPP — with couplings — auto-stepping & restep stability
+KEIETE: Kinetic, Internal, Total Energy conservation, including contact
+h, dt, ee: h-method mesh refinement and coarsening, time step sub-cycling, and error estimation

Table III-2. Estimates of reserve=F in capability (Fh) or capacity (Fp) for feature V&V: (We assume the given Tflops platform and the use of 10% of this machine for our V&V)

		Verification		Validation		
Feature	Relative Model Execution Time	TFlop	hero Reserve, Fh9 @TF	Tflop	production Reserve, Fp6 @TF	V&V+UQ p Reserve, Fp7 @TF
MCS-2:	1	4	4.9	12	7.5	.75
Thml-I	1	12	14.1	30	19.8	1.98
Mech-I:	10	12	1.3	30	2.0	.20
+Th/Ch	10	30	2.2	30	3.8	.38
+Xport	20	30	2.0	60	3.3	.33
+EIEIO	25	60	1.8	60	2.5	.25
+KEIETE	50	60	0.9	100	1.3	.13
+h, dt, ee	100	100	0.9			

The F ratios shown in red mean that we do not have enough machine capability or capacity to do our current estimate of what it will numerically take to verify and / or validate the given capability or application to the VER or VAL level desired. Any of these factors could easily be off by a factor of 2x or 4x; consider for example just one 2x step in global mesh refinement; for an explicit code, we would expect run time to increase by 2^4 or about 16 times.

With that caveat, here is a short explanation of how we got these capacity factor F numbers:

We used these conditions as our approximation to the Validation level VAL=6, or an Fp6 capacity level for Validation. For a system-level Base Case problem of MCS 2 (Sam et al, 2001) we achieved, from our most efficient code as our

Base Case:

- 4000 horsepower (element-steps per millisecond run time) on 2% of Blue Pacific (4Tflop Platform)
- 300,000 time steps
- 1,000,000 elements
- 1 day run time

We used these conditions as our approximation to the [code and *solution*] Verification level VER=9, or an Fh9 capability level for Verification to **Fh9:**

- In our [mainly thermal-mechanical] problems, speed was key versus capability
- 90 day allowable run time for an Fh9 hero mesh [solution] verification run
- MCS-2 took 2% of a 4T machine, ran in 1 day at V=6
- Assume 2^4 twice, and 4x more for improved ASCI-MPM material & physics models
- About 1000x more capacity needed

- Result is 100 day run on 20% of machine vs. 1 day run on 2% of machine for MCS-2
- This appears as Fh9=4.9 , assuming about a year to accomplish this hero run

For the other estimates of Fh9 for the capabilities in Table III-1, we assumed code compute intensity as high as 100x beyond the explicit Lagrangian MCS 2 by the 2008 time frame. For the case of Validation of these capabilities, on two code suites and two systems, per the procedures of Figures II-7 to II-12 (but not addressing all the comprehensive sensitivity and uncertainty quantification of Figure II-13), we estimated Validation capacity needs at Fp6 as follows:

- 400 full scale VAL=6 level runs per milestone
- VAL=6 will produce Figure II-7 through II-12 level quantification but not repeated sets for e.g. Figure II-13.
- Runs need to finish in 1 machine day; in reality about a 3 work-day results cycle
- We assume 10% of the fastest available platform to do this work
- MCS-2 ran in 1 day at 2% of our 4TF machine
- This enables 200*5=1000 runs/year, so Fp6=1000/400=2.5 for MCS-2
- But, with the 2 year delay from development to V&V we have 12T so FP6=7.5 (Figure III-4)
- Subsequent Fp6 based on the Load Factors for added mechanics, etc e.g. a 100x slower code for all the capability in Table III-2.
- We always assume the machine Tflops available 2 years hence

The preceding is of course just one example of the process of resource estimation. Those well versed in finite element development and analyses can see its subjectivity; just as in choices of uncertainty quantification, non-frequentist N* sample size, etc, there is a lot of expert judgment needed. It is clear that we should not plan to the minimum machine capacity / capability that we might calculate; our odds of missing by 2x or 4x in capacity are quite prominent. The result would be an inability to assess uncertainties by the same factor. In translation, the number or QRC cells assessed in the Figure II-14 Matrix would be unacceptably low; or else we would have to assess them all to a much looser uncertainty level.

Optimum schedule and rate of uncertainty quantification may also be determined by the number and experience of the people available to perform (and document) all these analyses. This is true regardless of platform capacity or capability. Either way, the risk is the same: an unacceptable bathtub increase in assessed uncertainty (therefore lower assessed QRC) of our product line. We must be judicious in the balance of estimation of platforms, people, and rate of assessment, and attempt, as we have here, to tie these back to a quantitative process such as that outlined in this work.

We have seen in this example, using this method, that we can project as much as a tenfold shortage in (ASCI Platforms) capacity. We can also see a tenfold shortfall in our designer/analyst teams ability to process and document this tenfold increase in information; solving this is a challenge for perhaps ASCI PSE/VIEWS (efficiency of the designer s working environment). These tenfold shortages are based on a complete Fp7 with Uncertainty Quantification sweep through The Matrix of WXX systems and STS environments once each ten years (to alleviate the bathtub assessed uncertainty increase), *using all the advanced code features available at that time*. What if we do not make such a sweep? Then the values of uncertainty in each cell of the Figure II-14 Matrix will have gone up, or confidence will have gone down, and the Reliability at Confidence (RC) product will go down. We can use QFD and QSV to quantify the risk of not having the ASCI platform capacity etc. to sweep the Matrix , and compare that to the cost and benefit of doing so. However, to do so, we must first estimate how much our uncertainty will be reduced by using all the advanced code features as planned on every analysis, as was assumed here. Fortunately, many of our assessments can be done using today s codes and model sizes. A sizable amount, though hopefully not a majority, require the more advanced (and platform intensive) mechanics and physics. We must balance available platform capacity and timing against the additional [model] uncertainty we will have to accept by running simpler, faster models. This type of example helps clarify and quantify missions and challenges for stewardship, ASCI, and our industry partners in the CTBT era.

3.1.5 Speed, the "Multiple Code Shootout", Timing, and Risk Mitigation

Speed of problem execution (and indirectly machine efficiency) are of secondary importance to high levels of V&V and low assessed error and uncertainty. However, with a severe capacity shortage by any reasonable estimate such as that shown in Table III-1, speed now becomes just as paramount: the *ability* to achieve study results with high V&V levels does us *no good* if we do not have the speed and machine efficiency, balanced against platform capacity, to execute that study with proper prioritization and **timing**.

The environments matrix of Figure II-14 already mentioned has hundreds if not thousands of cells that demand quantitative assessment. We address each cell each year as we assess our product line; but lacking a quantitative validation statement for each cell, this assessment is often intuitive expert or legacy judgment. We should in fact make a statistical validation statement about each cell each year. But we cannot because it takes time first, for the ASCI national labs, and commercial MPP code vendors in some cases, to build the code development capabilities. The reason we have gone for MPP is for size [read: resolution] and then for speed. Since measurement of speed and speed vs. accuracy tradeoffs is so essential to speeding up our ability to address the confidence and plan the needs of the stockpile, we have instituted a process that allows us to measure, report, and encourage enthusiasm for analysis speed at the ASCI level while we simultaneously work the V&V Meter readings. The term Multiple Code Shootout (MCS) is taken in the same spirit as the frequent use of the term Muscle Car Shootout which has an exciting legacy of independence in the drag racing world. That is, color and flare aside, how does performance measure up in a fair, independent setting? A schedule driven setting where V&V is driven independent from the code development effort makes for fair independent and realistic

evaluations of model expediency, code speed, and model accuracy. This should be a balanced setting, driven by program milestones and the time-urgent need to address the cells of the Figure II-14 Matrix, to care about speed *and* quality, and to be driven of necessity by those measures due to the compelling nature of mission and time constraint.

Classes of analyses where we feel the V&V Meters are on the verge of Green , as for the MCS-II analysis (Sam et al, 2001), are where we desire to be for production analyses. In reality the Meters are typically somewhere between 5 and 9 (see Figure I-6) when they are put to use for design and certification use. When the Meters are in the desired 6-8 range, we often select a problem with relevance and good cross-code capability as an MCS candidate.

The Multiple Code Shootout has never had any maximize the number of processors criteria. Our goal has been to run problems with ASCI-era resolution and potential for accuracy, and of a size that is adequate for accuracy and yet manageable in terms of pre and post processing and run time. Given practical, useful problems like this one, a real system level engineering problem that is contact dominated, the speeds and efficiency become more representative of reality than scaling done on a pile of merged brick elements. It is important to see the entire horsepower curve , because it is obviously not going to be optimum in efficiency to run the problems at peak horsepower. Rather, it is faster to run two problems on say half that number of processors each; in the end nearly twice the number of runs will get done. Absolute speed differences of say 20% are in practice not too important. When codes are slower by factors of two and three however, it will become compelling to find an alternative method unless the slower code offers additional features for related analyses not offered by the faster codes. Just as the use of more than one code is a model error risk mitigator, we can say that if run *at equal fidelity*, **speed is a key risk mitigator** for our concern about a lack of platform capacity in the coming decade.

Figure III-8 (Sam et al, 2001) shows, for the four code suites capable of engineering mechanics problems at the MPP level, what we can learn in terms of horsepower as a function of number of processors and element count in the problem.

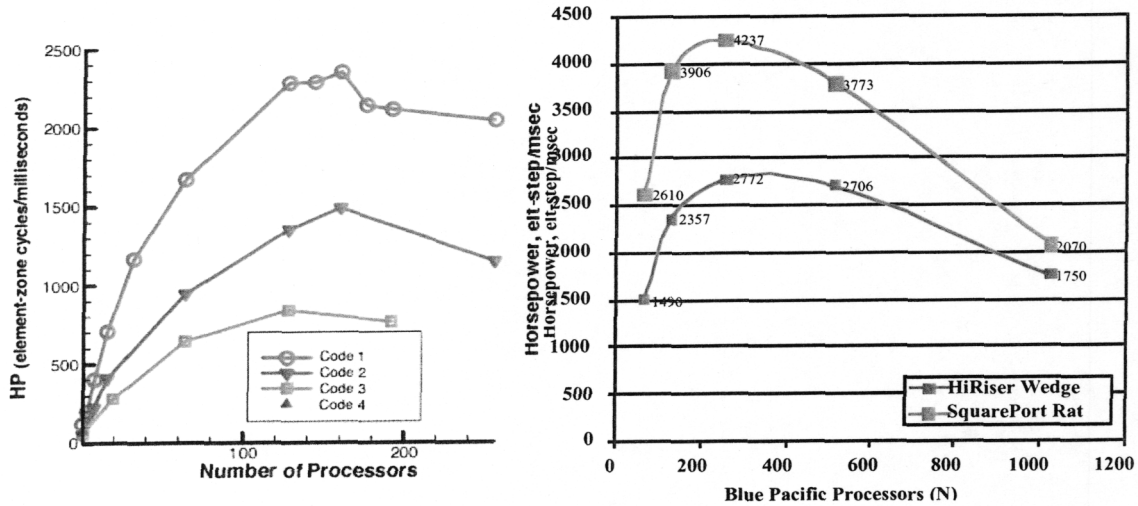


Figure III-8. Multiple Code Shootout II (Sam et al, 2001) modeled a blast environment full reentry system with internal nuclear package components. Left: ~200,000 elements. Right: ~1,000,000 elements.

3.1.6 Measures of "Analysis Horsepower"

There are many cases that involve the conditions, unlike those in MCS-II of Figure III-8, where code and model horsepower as measured by simple element-steps per millisecond is a misleading measure of model performance. A code may carry with its methods the computational overhead needed for higher order formulations and physics, but the extra weight may be worth it. That is, if an element has features such as Eulerian formulation, ALE options, h-method or AMR, p-method polynomial options, sub-cycling, etc, we will not likely see as many element-steps-per millisecond as a simpler Explicit Lagrangian 1-point integrated element formulation. Added capability is often traded off for slower speed, but the end result is that the path to an *equally credible answer* is now faster than before.

There are a few points that should be addressed regarding any analysis that contends a level of accuracy. For example, we define *verification* as the ability to match analytical or known solutions. However, we may for example show that we can get a beam bending solution to any number of significant digits; but only with say 100 or more elements across the beam. In reality, we rarely know our material properties to that level, and even if we do, knowing an elastic deflection to within 1% is usually not worth it; we would rather use our resolution elsewhere. The Verification analysis should tell us when we have converged, and how far we have backed off when the Verified feature and sub-model is put into the larger full system engineering model. A model termed Verified should be within the known solution to a desired level; not just because it has a lot of elements. Indeed, the number of elements to obtain given accuracy may well be a function of the code construction; this is the place where a code with less horsepower on a per-element basis may well redeem itself due to greater accuracy per element.

The known level of accuracy from small scale Verifications should then be incorporated into the larger, perhaps system-level, Validation model. It is known that the sub-models have a given

accuracy capability from the Verification studies. Validation comparisons are similar; in terms of distance from model convergence, comparison to other data, comparison to other codes; the criteria should be *the level of resolution* versus simply the number of elements used.

At a given resolution, we can begin to examine relative problem speed. Each Validation should contain this information, in addition to just speed : What level of accuracy was desired? Was it achieved? How did this trade off against speed, size, or resolution that could have been put elsewhere, where perhaps unknown events might have occurred but were missed due to a coarse mesh?

In addition to code horsepower [and the dilemma of defining it as we go beyond 1-point explicit Lagrangian models], the following are suggested as important:

- Teraflop Machine efficiency
- Horsepower (~1/Grind-Time) or Cell-Cycles / millisecond
- Horsepower at constant accuracy (spatial, temporal, iterative)
- Horsepower at constant fidelity (material and physics models)
- Total User Time:
    - Build Mesh
    - Run, View, Diagnose
    - Calibrate and Validate
    - Display and Conclude

The most logical final note on speed is that the last measure, Total User Time is perhaps most key. Like the common 20th century dilemma of wartime flight, pilots (nuclear designers?) are in perennially shorter supply than airplanes (compute platforms). Of course, this is easy to say, but it requires not only that we recognize that platform shortages are at least a problem we can mitigate with a balanced tradeoff, but that we actually perform such tradeoffs using QSV (Value) and Benefit/Cost Ratio BCR as outlined in the next section, Part IIIb.

3.1.7 Earned Value expressed as Enhanced QRC in The Matrix

Earned Value and Earned Value Management (EVM) are emerging (and in many instances mandated) program management tools in government and industry. The nomenclature is widely used, and a quote from the Earned Value Management Website (EVM, 2002) shows the direct relevance of Earned Value to our V&V-QRC-QSV investment strategy process described in this work: Earned Value is a management technique that relates resource planning to schedules and to technical cost and schedule requirements .to permit the customer to be able to rely on timely data produced by those systems for determining product oriented contract status . We provide here one example of these Earned Value concepts, e.g. resource planning , cost and schedule requirements , and timely data [i.e. enhanced QRC with timing to push back the confidence erosion depicted as the bathtub edge in Figure III-2. That is, if we combine quantitative V&V as described in Part I and Part II of this work, with sufficient machine capacity and code/analysis speed, we can show an example or two of just how all our Quantified V&V can quickly and solidly manifest itself in terms of the equivalent R and C terms across the

stockpile in The Matrix. Figure III-9, the upper Matrix is a notional stoplight of Z_{qrc} terms across systems and environments. The middle Matrix is the same chart — but with U reduced due to Platforms, code capability, V&V, and finally validated simulations leading to smaller U, and a Matrix with sharpened values of Z_{qrc}. The amount of sharpening is shown in the bottom Matrix — we can see just where the benefit of the Ladder Rungs will enter. Sharpening should always be viewed as an uncertainty reduction, but this may lead to a higher — or lower — estimate of Margin M. So, we cannot say the cells will become more green (a higher Z_{qrc}) with better V&V; we can only say that our assessment will be sharper. What we can say is that the risk of having the wrong Z_{qrc} or color for a cell goes down with additional V&V, and therefore our assessments of Reliability at Confidence will improve.

We will need to continually sharpen The Matrix to offset the expertise bathtub encroachment of Figure III-2. We can employ methods like those in (Logan, 2001a), (Booker et al, 2001), and (Nitta and Logan, 2000) to depict our estimation of the tendency of The Matrix to turn yellow/orange over the years. (In other words, margins may decrease with aging or changes of components or environments; uncertainties can increase due to the same factors or simply due to loss of assessment expertise and tools). These same methods can then quantify the value of our success in higher levels of validation and uncertainty quantification.

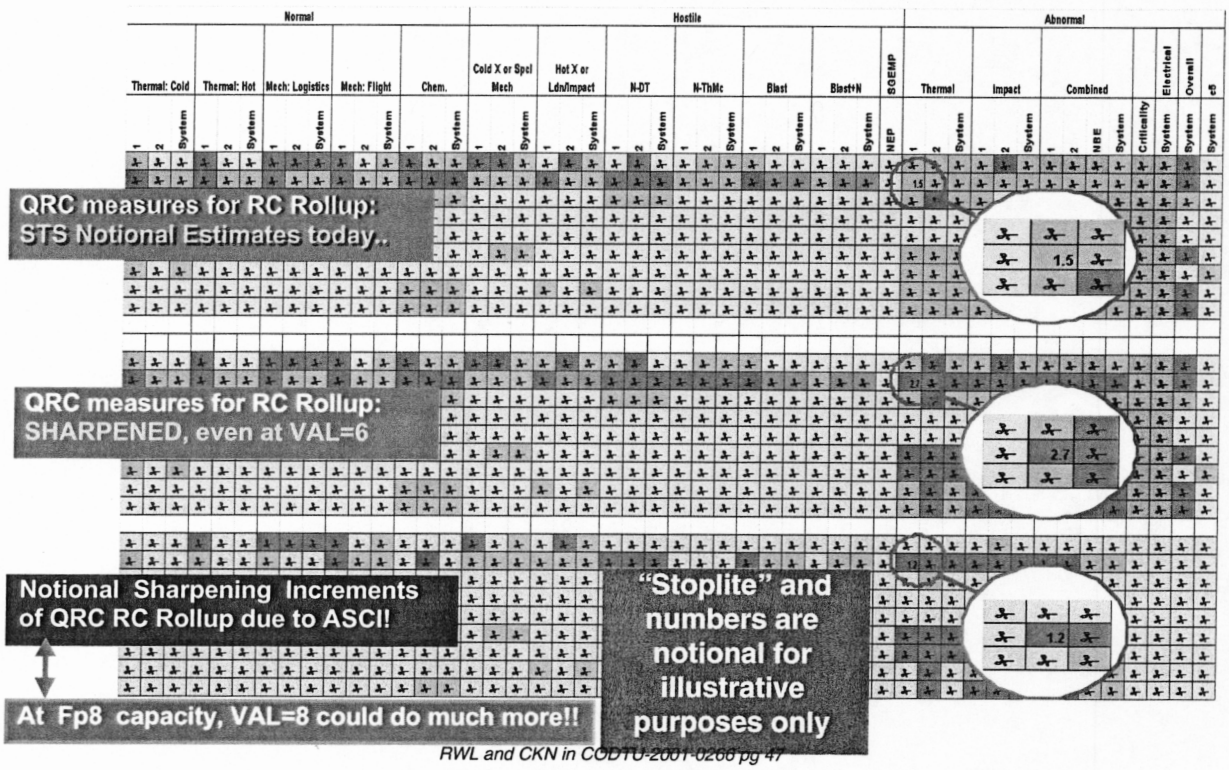


Figure III-9. The RC Rollup Matrix can show improvement in the numerical values (and hence colors) of its cells of uncertainty and confidence; these fold directly into QRC and then QSV as assessed earned value. The amount of sharpening (QSV added) is a direct function of the Validation Level (alias uncertainty quantification). We can thus link platform purchases and timing with QSV over time.

3.2 Part IIIb: Quantitative Systems Value (QSV) and Investment Strategy

How do we decide which cells in the (Figure II-14) matrix to sharpen — (which uncertainties to quantify better — or which margins to raise for example by refurbishing) - and when and how much?
The goal of QSV is to provide, with Quantitative Reliability at Confidence (QRC) input terms, a method to enable quantitative answers to this question, with QRC as a direct and visible term in QSV, and of course V&V and UQ as direct and visible terms in QRC as we outlined in Part II. We can therefore follow the value trace back through the compact QSV equation back through the QRC Matrix , V&V , code development , and Platform capacity Ladder Rungs of Figure III-7, and look at timing and optimum Benefit/Cost Ratio BCR for decisions and plans along the way.

QSV and BCR are really just ways of expressing Value Engineering (VE) concepts. Value Engineering is a fairly general set of concepts, and a general summary (from the VE website www.vetoday.com) is simply the quote that VE helps people creatively generate alternatives that secure essential functions at the greatest worth as opposed to costs . If we make more abstract the concept of worth , as implicit in both VE and QSV, we can think of the Benefit/Cost Ratio BCR as analogous to the more familiar accounting Return On Investment (ROI). Of course, in assuring the safety and reliability of the stockpile, these returns do not come in dollars, we must [subjectively] assign the returns a value based on the *quantitatively assessed* condition of that stockpile.

We assign to our product line, in this case the stockpile, and all subsequent factors and permutations in our analysis a dollar value , with the same focus as would any industry analysis of its business posture. The performance [mission effectiveness], diversity [odds of a common mode failure over time], safety [risk and dollar-estimated consequence], and management [surveillance and refurbishment] are all defined in dollars. It is certainly true that like any such analysis, subjective values abound. However, with all quantities in common units, we can reach numerical, quantitative conclusions and then proceed to *do uncertainty analyses around them.* The objectives are to say, for example, should we keep or retire a given system? With what should we replace it? While working on a system, should we refurbish all its components or just some of them? What is the Benefit-Cost-Ratio (BCR) of doing so? Sometimes we just go with our instincts — but it is preferable, to have a quantitative story that provides a stronger business case. And finally, how do these conclusions change in light of treaties or regulations, and how do they change if we consider Quantified Systems Diversity (QSD) or if we ignore it?

With reduced stockpile levels, quantification of our choices matters all the more; we do not have as many hedges due to numbers of warheads, numbers of systems, or diversity as we once had. To retain economy, we can perhaps avoid refurbishing entire builds of weapon systems — we can assume a block approach where only numbers needed to support the deployed stockpile

would be refurbished. This may actually allow us to live within the tight budget constraints that are inevitable as we deal with the ever present possibilities of recession, war, missile defense, and perhaps boutique weapon or advanced concept development. With these limited block assumptions, it may become advantageous to delete entire systems from our product line (the stockpile), depending on QSD assumptions, in order to maximize the QSV of the force within the budget constraints that we have. For all our systems, such tradeoffs require embedding credible and possible values of assessed QRC into our QSV value equation:

$$\Delta QSV = QSV_0 \int_t PV_F[t]\, \Delta[t]\{\Pi_{i=1,MC/STS}\ (RC^*)_i\}\, dt \qquad [1]$$

The preceding paragraph discussion on the changing nature of the world around us is the reason for inclusion of the Present Value (PV_F) term in Eqn. [1], where $0 \leq PV_F \leq 1$, so PV_F decreases from PVF=1 to PVF=0 as time approaches infinity. The PV_F term does not represent system aging or inflation (or deflation in some instances where purchasing power may increase with time). The PV_F term in Eqn. [1] is essential because for long term planning and investment, we have to account for the fact that we do not know the distant future as well as the near future. Using a PV_D=2% discount per year (2% for example) is a common way to express this fact. This results in a discounted value for future benefits (QSV) and costs as shown in Figure III-10.

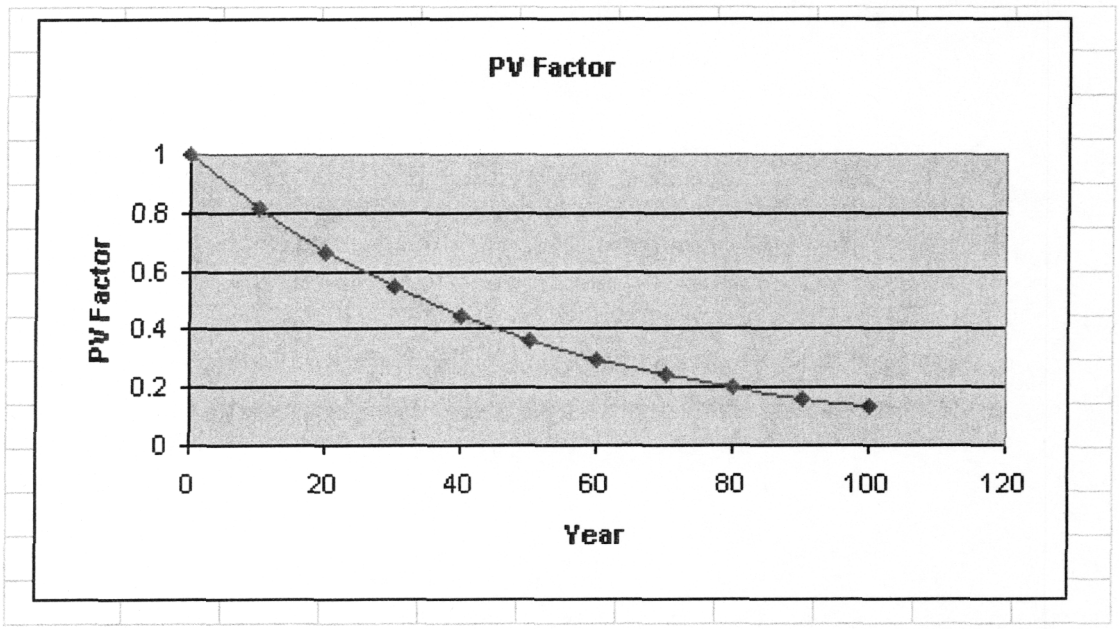


Figure III-10: Present Value Factor at 2% Discount as used in QSV Value Engineering.

Note that the PV chart goes out a full 100 years. This is to avoid discontinuities in the quantification process inherent to long term planning such as QSV. In other words, if the end-point year happens to include a new power system for one system but another system can get by with its old power system for say 5 more years, the latter system *looks* much cheaper to keep; but this is misleading. Using present value (PV_D~2%) as often used in industry for long range economic planning, plus a timeline that goes out 100 years, mitigates these discontinuities

and also represents the fact that we become more uncertain about what is going to happen as the decades go by; this is a standard reason for such discounting. The value of any event then, be it a degradation or a cost to fix the degradation, is PV d in time as shown in Figure III-10. This PV_D=2% plot would seem to be consistent with what our common sense might tell us about the accuracy of our crystal ball into the future decades: It says anything 35 years out is a 50/50 bet, 50 years out is about 1 in 3, 100 years out is about 1 in 9. Any bumps out at the 100 year mark, e.g. did we just miss putting in a new power system, etc, are only counted at about 11 cents on the dollar anyway so they won t affect the decision analysis that can come out of our Value Engineering QSV analysis (using of course V&V d values for the Reliability at Confidence over time.)

So, the use of a low value of PV like PV_D=2% per year in the quantification process to follow is why the estimates have to go out so far. The long, continuous, time frame also lets the longer term payoff investments show the math to justify their existence. But the fact that the long term is PV_D discounted also forces capital intensive endeavors (be they experimental or analysis) to prove their worth in a reasonable time frame, else the PV_D discount factor will fall before they can repay our investment in them.

Better data is needed in many parts of the QRC and QSV models. Many times large numbers are being subtracted from each other with the difference being only a billion or two dollars; uncertainties can easily swing the answer. We illustrate how to use the BCR as a measure of strength of the decision indicator, but still we stress that the method of QSV, while it has great value, must be accompanied by a study of uncertainties and a quest for better information.

3.2.1 Introduction to QSV: Quantitative Stockpile Value

The entire premise of QSV and its related quantities as discussed below is that all the quantities regarding value are expressed in dollars , or some equivalent quantity that can be added, subtracted, and compared in a consistent way — sometimes percent of total product line present-value is a good method also.

We must define a fairly simple definition of the time-integrated present-value of our product line (and indeed each individual product or warhead in our case) as outlined for example in (Logan and Nitta, 2000). To go any further we need some more definitions. Let:

B=Benefit, expressed in value e.g. dollars.
C=Cost, similarly expressed. Then let
BCR=(B-C)/C = Benefit-Cost Ratio. [6a]

The nature of our BCR, dating back to the first 1997 presentations of this work, is not unlike the Signal-to-Noise , Cp Process Capacity Ratio , or statistical Z and M/U concepts we have rediscovered in our business lately (Logan and Nitta, 2001). That is, we can look at the benefit (B-C) of an action; it may be a large or small $$$ value, but if |BCR|<1, the benefit certainty of

our decision is in doubt. High values of |BCR|>10 leave little doubt as to the correct course of action. Of course, some sensitivity studies should be done to assure |BCR| remains high under possible perturbations of the issue under consideration.

Continuing with our definitions, we now have:

QSV=Quantitative Stockpile Value, the portion of our total portfolio value that each product line, or the entire product line, contributes. ΔQSV=(B-C) is a good way to think about any given issue.

QSP: Quantified Systems Performance, this would in fact be identically QSV if not for the following term QSD, which makes QSV a nonlinear function QSV=QSV(QSP, QSD):

QSD: Quantified Systems Diversity: In terms of diversity and a hedge against common mode failure in the CTBT era, each system may not quite be a unique system . The more commonality, the more risk of common mode failure. We may think we have say 10 systems in a product line, but may only have effectively 7 or so unique ones. The resulting assumption in this method where QSV=QSV(QSP, QSD) is that over decades, there are certain [increasing] odds of losing a system due to some unforeseen aging or other unknown unknown that we cannot address in the CTBT era. Less diversity means fewer systems as a hedge to this, and also more chance of losing 2 systems at once due to common mode. QSD attempts to account for this. Just like QSP was offset by QSD, QSD is offset by another value whose acronym is QSM.

QSM: Quantitative Systems Management: (Maintenance). Of course more diversity (QSD) also means more types of components, materials, processes, etc — and systems to keep around. If we keep our diverse stockpile around over many decades, we will face more surveillance, more facility needs, more varied refurbishment needs. We will find QSM to be a significant number over the decades that can offset or even overwhelm QSD. ΔQSM=C=Cost is a good way to think about many issues and tradeoffs.

QSS: Quantitative Stockpile Safety: Described more fully in (Logan et al, 2000) and (Nitta and Logan, 2000), QSS takes the ENDS (Enhanced Nuclear Detonation Safety) thinking from ENDS=yes or ENDS=no to a graded, quantified scale that accounts for more than electrical system safety in prevention of a *nuclear* event — it attempts to include dispersal issues, a risk certainly more likely, as shown over the years, than any potential nuclear detonation in an accident. Suffice it to say, without detail, that our current estimates are that for the sake of this study and its tradeoffs, QSS is in general 1%-2% of QSV. That is like saying that, we do pay a safety penalty for having warheads around, but the BCR of this safety penalty is BCR=50 to 100 which is quite high. We have seen analyses showing BCR=10 or 20 as about the lowest point. Like other entries into the Matrix, (Safety aspects enter into the QRC Cells as well), each Safety QRC and BCR must be diligently quantified.

3.2.2 A Simple Everyday Illustration of QSV and BCR: Stopping at a Traffic Light:

QSV and BCR may seem abstract in concept at first. However, we instinctively use these concepts every day without actually estimating the numbers in our minds. We know the stockpile decisions get so complicated that we need to go back to a process that has numbers [albeit judgment based] so that we can trade off several things at once. But it seems our everyday decisions can get so instinctive, so obvious , that we often bypass this process. To illustrate, consider approaching a red light on your way to a meeting in your typical 4-door, front-wheel drive rental car — an experience most of us have had all too frequently. Let s look at the cost, benefit, and BCR as we approach this red light.

If we stop at the light, we will lose about 2 minutes of time. What if we are late for a SLEP meeting, and since we don t want to jeopardize FPU, we ponder running the light?
It looks bad to stop at the light. So, should we run the light then? Let s run the *numbers* first, before we run the light!

If we run the light, we will see a benefit — we will save about 2 minutes of time! Let s count our time at $1.00 / minute, an average of our $3.00/minute loaded at-work costs and our $0.00/minute cost while we are on travel but it is after the 8 hour work day . So, at $1.00/minute then,

$$B(QSP)=\$2.00$$

Another benefit of running the light is through QSS risk reduction. If sitting at the light, it is possible to have another inattentive driver attempt to defy physics and mechanics and enable his hood and your trunk to occupy the same physical space. Let s estimate the odds of this happening at 1e-4. Further, should this happen, the odds of a fatality resulting are, say, 1e-2. If the value of that life was assigned a value of $4e6, we have a net benefit B value of:

$$B(QSS)=1e\text{-}4*1e\text{-}2*4e6 = \$4.00$$

If we run the light, there is some Cost, C(QSM). It will cost in fuel say 1/8 mile at full throttle, or about:

$$C(QSM)=\$0.02$$

So if we run the light we have:

B=$2.00, QSP Quantitative System Performance [in this case the value of time saved]
B=$4.00, QSS Quantitative System Safety [in this case risk reduction of a rearend accident]
C=$0.02, QSM Quantitative System Maintenance [in this case fuel cost]

(B-C)=$5.58
BCR=279

So far it is looking like running the red light is the way to go. And there s that added QSS risk reduction bonus: If we run the light, we certainly won t get hit in the back!

Of course there is that dump truck approaching us from the left. Maybe we should do the BCR on stopping at the light first [most of our instincts would suggest this if we saw the dump truck].

If we stop at the light, then from a QSP standpoint, there is no benefit — we don t save any time getting to our meeting. So

B(QSP)=$0.00.

And of course there is a cost of stopping at the light — we lose about the same fuel as if we traveled a mile steady state, so

C(QSM)=$0.05.

It still looks mathematically like we should ve run the light! We should consider the dump truck though before coming to closure. The dump truck might T-Bone our car in the driver s door. There might be a fatality. And of course there is the well known relation that the odds of the fatality being us versus the dump truck goes as the ratio of the weights of the two vehicles to the 3.5 power. In other words if there is a fatality it is going to be us.

But, odds are about 9 of 10 we would make it anyway. And, if the dump truck hit us, odds are about 9 of 10 he wouldn t hit us right in the door, and we do have drivers side air bags , maybe. Even so, our QSS analysis would give a benefit from QSS Risk Reduction for not running the light of:

B(QSS)=0.1*0.1*4e6=$40,000.00

Now for the first time we see a large — very large — difference between running the light and stopping. Our BCR for stopping at the light, as our common sense tells us, is now:

B=$40,000.00
C=$0.05
(B-C) = $39,999.95
BCR=+800,000.

To review our lesson: Running the red light looked *GOOD*:

(B-C)=$5.58 net — a positive savings -
BCR=279, and since |BCR|>>10, a strong indicator of this savings.

But stopping at the light looked *slightly better:*

(B-C)=$40,000.00 [rounded]
BCR=800,000

We can see that the tradeoffs need to be done, like the stockpile, in a relative sense; decisions considered in isolation can lead to high consequences.

3.2.3 The process to get QSV and BCR:

So, how does QSV for the *stockpile* work? In essence, we first calculate Quantitative Systems Performance, QSP, for a warhead on its mission, that is, its deployment status with treaty considerations, etc. This gives a number, which is then multiplied by the number or function of the number of such warheads we have. This is repeated for each warhead type and summed to give a total potential. We can do this in more detailed studies — the conclusions depend on many assumptions and on the validity of our assumptions about treaties, world conditions, and on the V&V and QRC level of our assessments of each element of the stockpile.

Next, QSP (Performance) is combined with QSD (Diversity) to form QSV (Value). What is QSD? It is the uniqueness or diversity of features of a given system in relation to the others; a hedge against unknown unknowns that we can no longer address with tests. To do this we made a corporate internal list of CTBT-relevant diversity features across our product line .

These features are catalogued in a product table. That is, we ask, if a feature is good to have, how many systems have it? Say 6 of them. Then, for each of our [say] 10 systems, we have a row; if that system has the feature, we write 5 , because we would lose a system with that feature if we lost the system we are considering. If not, we write 6 ; there would be no loss in diversity of that feature if we lost that system. Each item in our list is a column. The product across all columns is huge. For a diverse system, the product will be small, because we will tend to be down a notch in most columns. So the ratio of [Total Product]/[Individual System Product] will be large; showing a lot of diversity. If the system is non-diverse we won t lose any features by losing it, so the ratio will be small — it is rarely, if ever 1.0, but it can be much smaller than a diverse system and much less than average . As a result, we can compute the QSD Ratio for all our systems, normalized to average, and fold this result into our QSV expression.

This ratio shows that, per the criteria set out, some systems are much more diverse than others. As we noted above, these same systems can often be hard to work on, because of their very diversity. To translate: Higher QSD (Diversity) is good, but higher QSM (Maintenance) is costly. Often, higher QSM accompanies higher QSD. The two trade off automatically in our QSV (Value) expression. Once we have a system s QSD Number , we scale it to a maximum of 1.0 [actually with a cutoff] so a system can represent a whole unique system, but may be only [say] half a system from a QSD standpoint if most of its features look like a partner system.

There is one more component to QSD. Over time, in the CTBT era, we will inevitably lose confidence in the stockpile to some extent — this loss comes from changes in the stockpile, but also changes in the nuclear weapons workforce and our distance from the underground nuclear test era as noted in Figure III-11 from (Joersz et al, 1997).

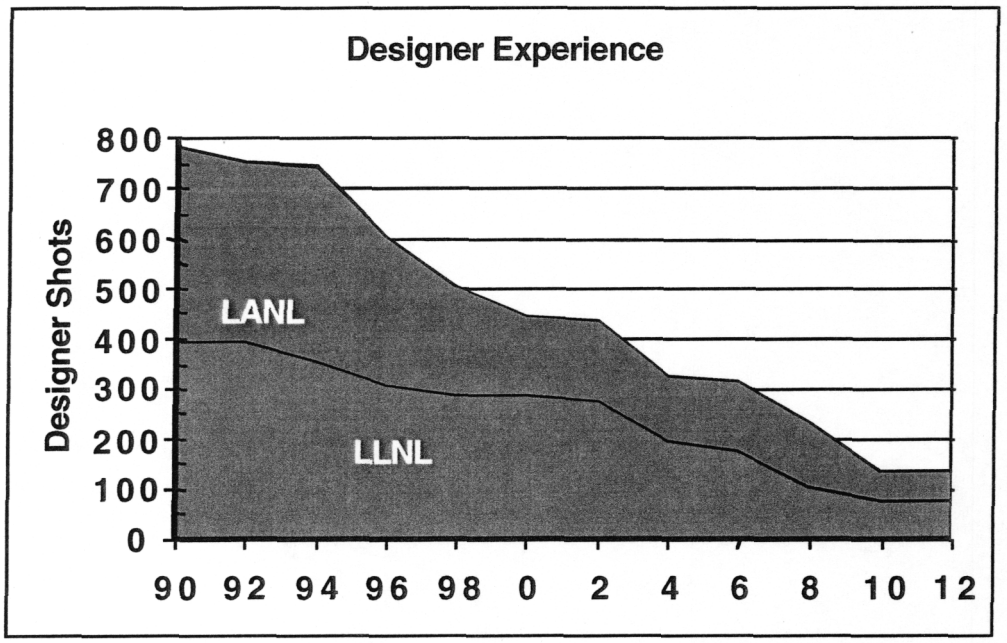


Figure III-11. Declining designer experience in the CTBT era.

A similar plot can be made based on surveys of projected confidence in any given stockpile system, in the CTBT era over a longer term (translate: concern about the edge of the expertise bathtub . The loss of direct experience in Figure III-11 may manifest itself in more assessed uncertainty in our CTBT product line , even though there may not (or may) be a real increase in uncertainty. The best we can do is make sure we take the action to acquire the tools (platforms for codes, models, and V&V) and efficiency (machine efficiency, code speed, and overall user time efficiency) to enable that experience to be put to most time efficient use.

This QSD correction is folded into the QSV(QSP, QSD) process, not by penalizing a particular non-diverse system, but by merely factoring, across the entire stockpile chosen, the fact that there are really fewer nuclear systems than we think due to lower QSD, and so our confidence may fall faster hence leading to lower QSV. This method of handling QSD was first implemented circa 1997, and was the best method we could find at that time. We continually revisit our QSV assessments using more modern methods such as those in (Logan and Nitta, 2001), (Logan and Nitta, 2002), and other works referenced therein.

So we now have a QSV system and unit quantity value , corrected for QSD [or not] as we may assume. There is one more major factor to fold in before we look at some example scenarios. That is the QSM (Quantitative Stockpile Management) term. We consider QSM to be the value [cost actually] of maintenance, surveillance, rebuild, etc. attributed to a system. These costs can be

elusive — again, we need continually better estimates of these costs and interdependencies. We must also attempt to account quantitatively for the fact that certain components on certain systems take more effort to maintain or refurbish and are more expensive to deal with; the balance of QSD and QSM offsetting each other.

The QSM life cycle cost number, over the life of the QSV process, is significant. This number, along with some major bumps in R&D cost as refurbishment options are planned and evaluated, forms the QSM number to retain [and therefore surveill, assess, refurbish] a given system. QSM must extend over the many decades of the extended life of the systems under consideration.

We discuss QSV for the CTBT nuclear stockpile in (Logan and Nitta, 2000), but a generic example of the combination of QRC and QSV needed to do this will illustrate the simplicity but fragility and interdependency of the method.

Suppose we could update a certain component, a widget , on our systems, to assure more confidence and a higher assessed QRC over the decades of the CTBT era. Of course, such an update would carry a cost, QSM, along with it. As an example of the use of QRC, QSV and BCR to quantify our decision making process, we took a quick look at this issue and postulated that all of our stockpile systems could be given an appropriate widget upgrade . We included complete QSM costs as best we could, based on best available information and estimates. We further assumed the same development and unit cost for each type of widget; this needs refinement in some areas that range from obvious to not so obvious.

How did we quantify the Benefit, B? To do so, we used a combination of methods for Quantitative Reliability [R] and Confidence [C] for the nuclear package, as outlined in Part II of this work. We used the most up to date design analysis data available to us for each system, along with standard estimates of uncertainties, or specific values where we could reference them for particular systems. Now, we could quantify the changes in equivalent [R] and [C], and evaluate the DQSV and then BCR:

BCR=(B(DQSV)-C(QSM) / C(QSM)
[6b]

We looked at the QSV, BCR of widget upgrade for 'all' systems. Here are the results in order of highest BCR [ie highest priority] in this first-cut analysis:

1. BCR=18.0 WII, with special caveat
2. BCR=18.0 WCC, with special caveat
3. BCR=11.0 WHH
4. BCR=10.0 WDD
5. BCR= 6.0 WEE
6. BCR=02.2 WFF, with special caveat
7. BCR=02.0 WBB
8. BCR=01.4 WAA
9. BCR=00.2 WGG

The WAA through WII represent particular systems in our product line. Now, given a fixed QSM budget, we can, after consideration of plant capacity, timing, present value, and other factors, choose a priority order.

By refining these BCR s, and looking at changes in BCR and (B-C) as we trade off possible implementation dates for our widget upgrade hypothesis, we can now quantitatively fold these widget options in with our other certification decisions to see what we might do first, and what we might delay, with some quantified input for justification and closure. A BCR$\geq$10 seems at first glance to be a compelling thing to do. A closer look is needed here.

For example, since B=Benefit=DQSV, and DQSV is proportional to QRC, we see that the validity of our QSV prioritization depends directly (in fact nearly linearly) on the level of our V&V and UQ. If our assessed uncertainty is large (or even guessed versus a rigorous assessment), we still have a simple BCR ordered widget priority list ; but is it right? Only a credible, prioritized, balanced assessment of V&V-UQ-QRC-RC Rollup will tell us if our priority-ordering is sensitive to things we are not so sure of.

The quantified Benefit / Cost Ratio (i.e. Value Engineering) closure of the RC Rollup method using QSV allows quantitative, traceable decision criteria that are *one* tool we use in our complex, nonlinear, data and statistics-starved world of nuclear package stockpile stewardship. An example of this closure is depicted in (Logan, 2001a); a suggestion regarding the value of testing or preserving certain stockpile assets. The same method, with help and credibility that V&V gives to The Matrix, could be used to trade off refurbishment dates and content, and facilities and quantities of validation testing needs, etc. Of paramount importance are the uncertainties U — the value of the assets and the size of the error bars are major factors in the effort we spend reducing U to an acceptable level — and hence major factors in our determination of validation metrics for this and other scenarios.

Continuing our Risk=Likelihood*Consequence analogy that has closed Part I and Part II, Figure III-12 shows the new terms introduced in that analogy as they come out of Part III. Others (Kilkauskas, 2002) have also presented the concept of quantifying assessments and model uncertainties into a Risk dollar quantity.

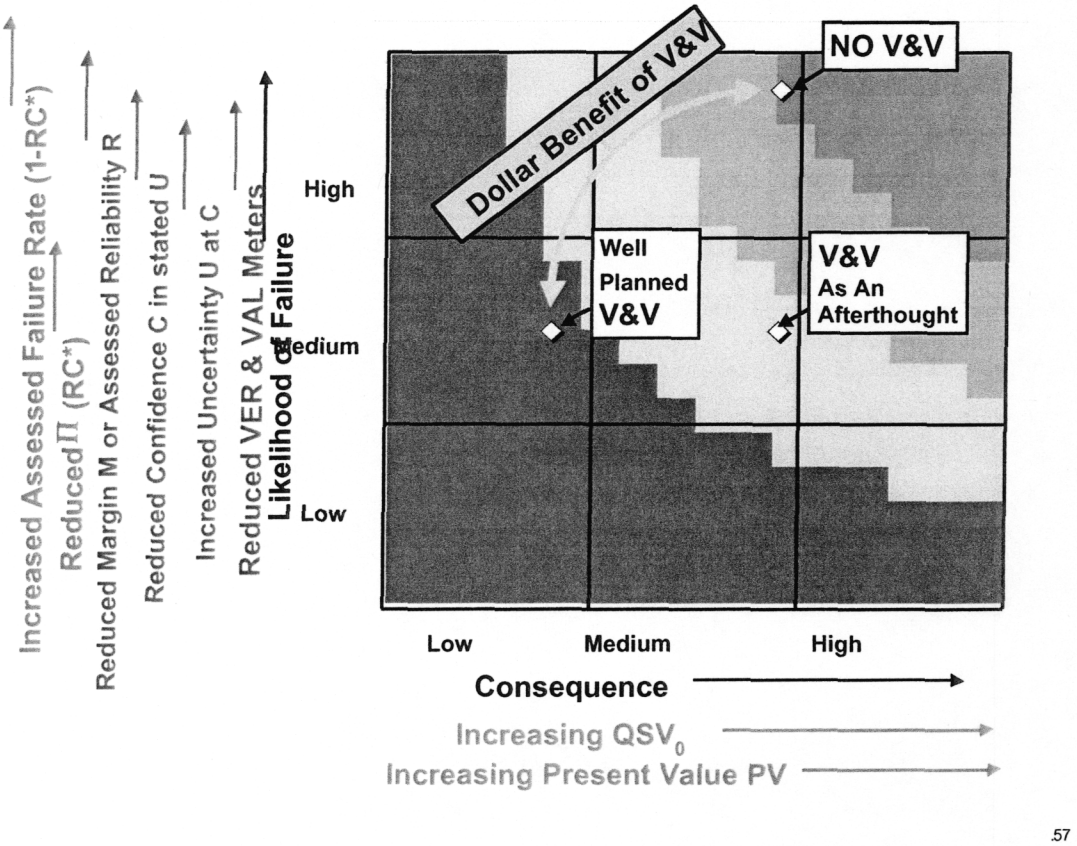


Figure III-12. Risk=Likelihood*Consequence and analogous terms from Part I (green), Part II (blue), and Part III (red). Likelihood (of failure due to model error) becomes analogous to assessed (1-RC*), using the rolled-up RC* product term; Consequence is expressed in Value Engineering / Earned Value [ie dollars] terms, and Risk=Likelihood*Consequence is the multiplicative product. In other words, *(Model Error) Risk is equivalent to the ΔQSV assessed due to model error*. In this way, the [dollar] value of model error risk reduction due to V&V can be quantified. We can illustrate with situations where Risk [as dollar value] is lowered by doing V&V to reduce likelihood of model failure, and where Risk is lowered by doing V&V early on to reduce Consequence as well.

4 Summary and Conclusions:

4.1 Part I: V&V and UQ

Our V&V methodology goals include a document leading to quantitative V&V statements as phrased in Part I and detailed in Part II. This should be done for each V&V study addressing part or all of the 84-point criteria of Part I of this work. It will rarely be cost effective to address all 84 points for a given analysis. However, as risk and consequence increase, it will be more important to address more of these criteria.

In this era of increasing audit pressure for quality, a quantified, documented methodology is essential. Documentation should not be a waste of time or a nuisance, it should be, as it evolves with the help of those affected by it, a source of pride. We must encourage the philosophy that even a documented analysis with known flaws has far more credibility than an undocumented analysis with unknown flaws. There is never enough time to document everything on a checklist that one should do — but this policy is an effort to encourage making the time. We must provide a policy of rewarding documentation of *good* work with *supportable* conclusions. We hope these evolving guidelines lead us to an era where movie of the week analyses will be considered in their true context, and in contrast, analysis that have a high [reviewed] Quality Meter reading will be those used to retain our credibility for stewardship and the nuclear deterrent.

The authors welcome improvements, suggestions, and refinements. The concept of a couple of simple 0-10 Meters is gaining nationwide popularity. Maybe round automotive gauge styles aren t for everyone, but they have attracted attention and focus as a way to quantify V&V; we have found the illustrative meters to be a good starting point. Numerous agencies have their own numbered levels for the pedigree of a code and quantitative levels for the pedigree of validation of a simulation.

This version of our V&V Methodology and the 84 Point criteria [10+20+19+35] are always a work in progress. It is true that a checklist can be helpful if it is not overly constraining as to become a hindrance without adding value. The 84 points listed should be considered — and revised year by year — in that spirit. Others in the national community are engaged in similar activities; quantified V&V is just coming of age in many areas. However, we are obligated to start the process given the maturity of ASCI. Quantitative measures such as the V&V Meter readings and how many of the 84 points did we address will allow us to track and revise our process as we evaluate its cost / benefit to the future of our mission. Quantitative statistically based V&V statements will be the next evolutionary step.

4.2 Conclusions: Part II (QRC) into Part III (QSV)

The advantage of Z_{qrc}, leading to Quantified Reliability R at Confidence C (QRC), then assessed Log Failure Rate as a "magnitude" indicator - with QRC rolled up into QSV at the system level - is that it lets us:

* Relate Margin "M and Uncertainty U" to R .
* Demand that we associate "R" with a "C" - and quantifies that "C" based on eg "Fuzzy N*".
* Show how better assessments of M and U - and increasing the effective number of coin flips N* - quantitatively tightens U allowing higher quantified C.
* Express this situation as value *(dollars)* via QSV - and so we can ***protect*** **the investments in stewardship assessments, tests,** etc by *Quantifying* their *Stockpile Value*. This is a continuing need in the current budget environment, especially where purely technical analyses remain under utilized by some. Perhaps these audiences have lacked a clear link between science and V&V and Value and Investment Strategy — *it is our hope that QRC-into-QSV will provide this link.*
* Provide a numerical [albeit judgment folded] estimate of "how much confidence do we need and how do we get it".
* Provide a quantitative V&V statement as a lower statistical bound. This has the advantage of being *quantitative and assessable*, but also the advantage in that we recognize that the upper bound on a product may still be quite high (even unity); this avoids conflict with assertions, whatever the source, that high reliability is expected or promised from a given product. The V&V-based lower bound *assessment* and the upper bound *assertion* can both be right; and both have appropriate uses.
* Prioritize, since the BCR {Benefit/Cost Ratio] in QSV shows us how urgently we should address a given issue, especially when integrated over time as most clearly in the Engineering Index (EI) works such as (Booker et al, 2001).

We recognized early on that methods like FOM/QMU or its 1-tailed convolution into CF cannot by themselves bring this quantification and closure around an investment strategy. Therefore, while the concepts of QMU and CF can certainly be *components* of a system to closure, we prefer methods like QRC/QSV that show the path to closure instead of "open-ended conservatism" like the FOM/QMU=2 and CF=.33 examples detailed above. A great value could be added if methods like QMU and CF are used at the local level [eg as in the current QMU/CF working groups] to find the areas where FOM/QMU and CF are considered low — and plan specific technology efforts around those. We can then do the more detailed closure via methods like QRC [and equivalents at Los Alamos and Sandia] and fold into an Investment Strategy method like QSV or its equivalent. For the near future, we would suggest that the traditional Margin M continue to be shown, as well as other chosen measures of expressing uncertainty.

The use of $Z_{qrc}= M/U|_{N^*,\sigma}$ as used in QRC provides for a path to closure using either:

1. A standard U , so that the closure through QRC to QSV and Benefit/Cost Ratio BCR would continue to be proportional to Margin , as our intuition yearns for, but also
2. A specific U , so closure through QRC to QSV and BCR tradeoffs could quantify, the benefit of investments to reduce that specific U.

4.3 Conclusions and the Future: Improvements and Refinements

Quantified stockpile assessment and certification methodologies:

Part II of this narrative has described the relation between several specific methods, represented by acronyms used for various terms in quantified assessments, notably FOM/QMU, CF, and QRC/QSV. These and other acronyms — and the many methods developing in the community not yet tagged with an acronym — need to be pursued. We note that QRC/QSV (and the combination of TRB and EI discussed elsewhere) have the potential to provide:

- Quantification
- Closure
- Prioritization
- Planning
- Tradeoffs

We have also expressed our method in a compact form — although as we have noted, each term in the summations of the equation can represent multiple careers of work and knowledge — and partnerships of many agencies involved:

$$\Delta QSV = QSV_0 \int_t PV_F[t]\, \Delta[t]\{\Pi_{i=1,MC/STS}\ (RC^*)_i\}\, dt \qquad [1]$$

Our major caveats are that the methods are new, they are complicated, and they require technical experience and discretion to develop, implement, and use for decisions — because the outputs are quantitative and the implications are far reaching. It is equally important to think through a **methodology with closure**. We realize that even this closure with QSV is essentially a Likelihood method. But, it is an essential step. Otherwise, we will end up fulfilling the prophecy of Yogi Berra, often echoed by Stan Trost, DP Emeritus and namesake of the infamous Trost Chart :

If you don't know where you're going, you'll end up somewhere else.
- *Yogi Berra, as so frequently quoted by Stan Trost while at DOE-HQ.*

5 Acknowledgements:

The authors have had many helpful discussions with many colleagues in this rapidly evolving area. We would especially like to acknowledge helpful discussions with our partner lab colleagues. Discussions and insight from Bill Oberkampf, Tim Trucano, and Marty Pilch of Sandia have motivated our work. We also had many helpful discussions and concepts contributed by Cheryl Faust, Jane Booker, Mike Hamada, Ron Dolin, David Sharp, Karen Pao, Merri Wood, Peter Adams, Dawn Flicker and Scott Carman, of Los Alamos and Livermore respectively. These colleagues at our partner labs are pushing the barriers of the quantitative linkages between validation, uncertainty, and confidence in the stockpile and it is our privilege to interact with them. We also would like to thank Carolyn Mangeng, Bill Reed, and Rich Wagner for encouragement to pursue these concepts and this work.

6 Appendices

6.1 Appendix A: Verification & Validation Definitions given by others

For context, we provide here several definitions of V&V at this point, and highlight those that are our favorites.

VERIFICATION DEFINITIONS:

Verification: (Roache, 1998): Verification ~ solving the equations right.

Verification: (Schlesinger, 1979): Substantiation that a computerized model represents a conceptual model within specified limits of accuracy *Quoted from (Oberkampf and Trucano, 2002) pg 12.*

Verification: (IEEE, 1991): The process of evaluating the products of a software development phase to provide assurance that they meet the requirements defined for them by the previous phase *Quoted from (Oberkampf and Trucano, 2002) pg 13.*

Verification: (DoD, 1996): Verification is the process of determining that a model or simulation implementation accurately represents the developer s conceptual description and specifications. *Quoted from (DON, 1999) enclosure 2 pg 2.*

Verification: (AIAA, 1998): The process of determining that a model implementation accurately represents the developer s conceptual description of the model and solution to the model. *Quoted from (Oberkampf and Trucano, 2002) pg 14.*

Verification: (Aubrey et al, 2001): The process of determining that a simulation code accurately represents the code developer s description of the model [e.g., equations, boundary conditions, etc.)

Verification: (Cafeo and Roache, 2002): Verification of a CODE: The process that determines that the computer code accurately represents the mathematical equations.

Verification: (Cafeo and Roache, 2002): Verification of a CALCULATION: The process that determines that the computer calculation for a particular problem of interest accurately represents the solutions of the mathematical model equations .

VALIDATION DEFINITIONS:

Validation: (Roache, 1998): Validation ~ solving the right equations.

Validation: (Schlesinger, 1979): Substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model. *Quoted from (Roache, 1998) pg 26.*

Validation: (IEEE, 1991): The process of testing a computer program and evaluating the results to ensure compliance with specific requirements. *Quoted from (Oberkampf and Trucano, 2002) pg 13.*

Validation: (Mehta, 1995): Validation is defined as the process of assessing the credibility of the simulation model, within its domain of applicability, by determining whether the right simulation model is developed and by estimating the degree to which this model is an accurate representation of reality from the perspective of the intended user. *Quoted from (Roache, 1998) pg 26.*

Validation: (DoD, 1996): Validation is the process of determining the degree to which a model or simulation is an accurate representation of the real world from the perspective of the intended uses. *Quoted from (DON, 1999) enclosure 2 pg 2.*

Validation: (AIAA, 1998): The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model. *Quoted from (Oberkampf and Trucano, 2002) pg 14.*

Validation: (Aubrey et al, 2001): Determining the degree to which a code provides an accurate representation of chosen physical phenomena.

Validation: (Cafeo and Roache, 2002): Validated Model: A model that has confidence bounds on the output. A validated model output has the following characteristics:

1. **The quantity of interest**
2. **An estimate of the bias**
3. **A set of confidence bounds.**

A validated model is one where we can make a formal statement after running the model similar to:

I am 90% confident that if I build and measure the quantity of interest, that it will fall within the confidence bands (of uncertainty) shown around the model output

7 List of Symbols, Abbreviations, and Acronyms:

A&C	Assessment & Certification
ASCI	Accelerated Strategic Computing Initiative
B	Benefit, usually in $$$, Millions ($M), or Billions ($B)
BCR	Benefit / Cost Ratio
C (in BCR)	Cost, usually in $$$, Millions ($M), or Billions ($B)
C_i (in QRC)	Confidence, in the ith environment (i omitted if only one)
CF	Confidence Factor: More properly Capacity Factor
CTBT	Comprehensive Test Ban Treaty
DoD	Department of Defense
DOE	Department of Energy, Specifically NNSA
DOF	Degrees Of Freedom (model)
DP	Defense Programs
DSW	Directed Stockpile Work
EI	Engineering Index, due to LANL-ESA
ENDS	Enhanced Nuclear Detonation Safety
ESA	What WX is now called
EVM	Earned Value Management
FMVSS	Federal Motor Vehicle Safety Standards
FOM	Figure of Merit, "M/U" in QMU
FR	Failure Rate
HIC	Head Injury Criterion, from Motor Vehicle Safety
LANL	Los Alamos National Laboratory
LLC	Limited Life Component
LLNL	Lawrence Livermore National Laboratory
M	Margin, where Factor of Safety = M+1
M_i	Margin, in the ith environment
MC	Military Characteristics
μ	population mean
N	Frequentist N, Number of trials as in coin-flipping
N*	Inferred, Weighted, Bayesian or Fuzzy N
NeLoFaRA	Negative Log Fail Rate, Assessed
NEP	Nuclear Explosives Package
NPR	Nuclear Posture Review
PDF	Probability Distribution Function
QMU	Quantified Margins and Uncertainties
QRC	Quantitative Reliability at Confidence
QSD	Quantitative Stockpile Diversity
QSM	Quantitative Stockpile Management (Maintenance, Refurbishment, Surveillance)
QSP	Quantitative Stockpile Performance
QSS	Quantitative Stockpile Safety
QSV	Quantitative Stockpile Value
PV_D	Present Value Discount Rate, e.g. 2%
PV_F	Present Value Factor, $0<PV_F<1$
R_i	Reliability in the ith environment (i omitted if only one)
RC*	Reliability at Confidence product equivalent
RCR	RC Rollup, across all system environments in TMX
ROI	Return On Investment
S_{cij}	Sensitivity in Capability in the ith environment to the jth parameter
SLEP	Stockpile Life Extension Program [now known as Refurbishment]
START	STrategic Arms Reduction Talks
STS	Stockpile [to] Target Sequence
σ	population standard deviation
TMX	The Matrix of MC/STS System Environments
TRB	The Reliability, Bayesian
U	Uncertainty, General or "System" [in V&V always at a confidence C]
U_{cij}	Uncertainty in Capability for the ith environment due to the jth sensitivity and parameter
U_{ci}	Uncertainty in Capability for the ith environment
U_{pj}	Uncertainty in Parameter for the jth parameter [material, tolerance, etc.]
UQ	Uncertainty Quantification
V&V	Verification & Validation
VE	Value Engineering
Z	Standard Normal Distribution Variable for variable X, $Z=(X-\mu)/\sigma$
Z_{qrc}	Standard Normal Distribution Variable, $Z=M/U|_{N^*,\sigma}$ in QRC

8 References:

(Adams et al, 1999): P. Adams, C. Nakleh, D. Sharp, M. Wood, and J. Glimm, Quantification of Uncertainty and Stockpile Confidence (U), in LA-CP-00-293, SRD, October, 1999.

(AIAA, 1994): AIAA, Editorial Policy Statement on Numerical Accuracy and Experimental Uncertainty, *AIAA Journal*, Vol. 32, No. 1, Jan. 1994, p. 3.

(AIAA, 1998): AIAA, Guide for the Verification and Validation of Computational Fluid Dynamics Simulations, AIAA-G-077-1998, Reston, VA, 1998.

(ASCI, 2000): DOE-Defense Programs, Accelerated Strategic Computing Initiative (ASCI) Program Plan, DOE/DP-99-000010592, January, 2000.

(ASME, 1994): ASME, Journal of Heat Transfer Editorial Policy Statement on Numerical Accuracy, *ASME Journal of Heat Transfer*, Vol. 116, Nov. 1994, p. 797.

(Aubrey et al, 2001): J. Aubrey, F. Brown, S. Clancy, A. Dana, D. Harris, A. Heath, T. Seed, M. Kirkland, J. Kamm, A. Mathews, T. McKee, J. Painter, D. Post, W. Rider, D. Shirk, and S. White, Report of the Activities for Completion of the V&V Level-1 Milestone for Initial Validation Methodology for Early-Time Primary Behavior by the Los Alamos National Laboratory (U), SRD Report, LA-CP-01-259, 30 March 2001.

(Booker et al, 2001): J.M. Booker, R. Dolin, C.L. Faust, and M.S. Hamada, Engineering Index: A Qualification Metric, 4th Biennial TriLab Engineering Conference on Modeling and Simulation, October, 2001.

(Booker et al, 2002): J.M. Booker, T.R. Bement, and K.F. Sellers, Linking Probability Theory and Fuzzy Sets — A Study in Uncertainty Assessment, AIAA-2002-1570, April, 2002.

(Cafeo and Roache, 2002): J.A. Cafeo and P.J. Roache, private communication of draft V&V definitions, April, 2002.

(Clough, 1997): R.E. Clough [approval], DTED Quality Assurance Policy and Plan, M-078-37/ENW-84-001, Revision 3, LLNL, February, 1997.

(DON, 1999): Department of the Navy, Verification, Validation, and Accreditation (VV&A) of Models and Simulations, SECNAVINST 5200.40, DoD, April, 1999.

(EVM, 2002): http://www.acq.osd.mil/pm/evbasics.html, Illustrative Explanation of Earned Value, Earned Value Management Website, 2002.

(Flicker et al, 1999): D.G. Flicker, M.G. Sheppard, K.D. McLenithan, H. Martz, and J.M. Booker, Performance and Uncertainty Analysis (U), in LA-CP-00-293, SRD, October, 1999.

(Goodwin, 2001): B.T. Goodwin, private communication to certification workshop, Los Alamos, December, 2001.

(Hamada, 2001): M.S. Hamada, in LANL/LLNL Alternative Reliability Assessment Methodologies Peer Review (U), LANL Meeting Material, ESA-WE-01-04125, February, 2001.

(Heath et al, 2002): A.R. Heath and Board, ASC V&V Strategy and Plan, LA-UR-02-4314 Draft Version 2.2, 10 July, 2002.

(Hecker, 1974): S.S. Hecker, A Cup Test for Assessing Stretchability , Metals Eng. Qtrly, p. 30-36 (Nov).

(Hodges et al, 2001): A. Hodges, G. Froehlich, D. Peercy, M. Pilch, J. Meza, M. Peterson, J. LaGrange, L. Cox, K. Koch, N. Storch, C. Nitta, and E. Dube, ASCI Software Quality Engineering, Goals, Principles, and Guidelines , DOE/DP/ASC-SQE-2000-FDRFR-VERS2, February, 2001.

(Horrillo, 2001): T.A. Horrillo, 2001 Annual Assessment , CODT-2001-0374, LLNL, 6 June 01, 27 pgs.

(IJNMF, 1994): P.M. Gresho and C. Taylor, Editorial , *International Journal for Numerical Methods in Fluids*, Vol. 19, 1994, p. iii.

(JFE, 1993): C.J. Freitas, Editorial Policy Statement on the Control of Numerical Accuracy , *ASME Journal of Fluids Engineering*, Vol. 115, No. 2, June 1993, p. 339.

(Joersz et al, 1997): E. Joersz, R.L. Hymer, R.B. Barker, K.C. Malley, J.H. Birely, R.L. Wagner, W.J. Howard, and L.W. Woodruff, A view of DoD s Requirements for DOE s Denfense Programs in 2010 , August, 1997.

(Juzaitis and Sharp, 2001): R.J. Juzaitis and D.H. Sharp, Certification in the 21st Century , Los Alamos draft SRD report, June, 2001.

(Kane, 1986): V.E. Kane, Process Capability Indices , J. Qual. Tech. 18 No. 1, Jan 1986, p. 41.

(Keane, 2001): R. Keane, HKS Inc. private communication to audience re ABAQUS and ISO 9000, at Los Alamos, August, 2001.

(Keeler, 1968): S.P. Keeler, Understanding Sheet Metal Formability , Machinery **74**, No. 6-11.

(Kilkauskas, 2002): M.L. Kilkauskas, Estimating V&V Resource Requirements and Schedule Impact , Proc. Foundations 02: Foundations for V&V in the 21st Century, NTSA Event 31V0, Johns Hopkins Univ., October, 2002.

(Kogure and Akao, 1983): M. Kogure and Y. Akao, Quality Function Deployment and CWQC in Japan , Quality Progress, October, 1983, p. 25.

(Krauser, 2001): W.J. Krauser, private communication to certification workshop, Los Alamos, December, 2001.

(Levine and Twining, 1995): R. Levine and B. Twining, DOE/AL Quality Criteria (QC-1) , DOE/AL, Revision 9, July, 1995.

(LLNL-ENG, 2001): LLNL, *Engineering Policies & Procedures Manual,* UCRL-AR-145342-7.

(Logan and McMichael, 1995): R.W. Logan and L.D. McMichael, "Modeling the Dynamic Crush of Impact Mitigating Materials", *PVP-307, Transport and Storage of Radioactive Materials, Ed. R.W. Carlson, R.S. Hafner, and W.H. Lake, ASME Book H00970-1995.*

(Logan et al, 1995): R.W. Logan, S.A. Perfect, and R.D. Parkinson, "Energy Absorption in Aluminum Extrusions for a Spaceframe Chassis", *Proc. IX Intl. Conf. on Vehicle Structural Mechanics*, SAE (1995).

(Logan, 2001a): R.W. Logan, DSW A&C Integration of LANL s TRB and LLNL s CRUMBS Reliability and Decision Methodology: A Preliminary Example: 2nd Edition (U), LLNL SRD CODT-2001-0265, April, 2001.

(Logan, 2001b): R.W. Logan, C-6 Integration with DSW A&C Margins & Uncertainties FY01 L2 Milestone, LLNL CODTU-2001-0572, July, 2001.

(Logan and Maker, 1997): R.W. Logan and B.N. Maker, The Relative Sensitivity of Formability to Anisotropy, SAE Paper 970440, in Sheet Metal Stamping: Development and Applications, (SP-1221), SAE, Warrendale, PA (1997) p. 77.

(Logan and Nitta, 2000): R.W. Logan and C.K. Nitta, Quantitative Stockpile Value (QSV), Rides for Nine (R49), and relevance to Nuclear Posture Review (NPR), LLNL CODT-2000-0247, March 2002.

(Logan and Nitta, 2001): R.W. Logan and C.K. Nitta, V&V Methodology: A 69-point checklist, Metrics, and Quantitative Certification Methods with STS examples, LLNL CODTU-2001-0266, December 2001.

(Logan and Nitta, 2002): R.W. Logan and C.K. Nitta, Certification Acronyms: FOM/QMU, CF, and QRC/QSV, LLNL CODTU-2002-0520, June 2002.

(Miller, 1995): G.H. Miller [approval], DNT Directorate Quality Assurance Plan, DNT-QA.0, LLNL, February, 1995.

(Montgomery, 1985): D.C. Montgomery, Introduction to Statistical Quality Control, 2nd Edition, John Wiley & Sons, 1985, 1991.

(Moore and McCabe, 1989): D.S. Moore and G.P. McCabe, Introduction to the Practice of Statistics, W.H. Freeman & Co., 1989.

(Nitta and Logan, 2000): C.K. Nitta and R.W. Logan, W76 Dual Revalidation Independent Review Team Final Report (U), SRD LLNL COPJ-00-0043, April 1, 2000.

(Oberkampf and Trucano, 2000): W.L. Oberkampf and T.G. Trucano, Validation Methodology in Computational Fluid Dynamics, AIAA 2000-2549, Denver, CO, June 2000.

(Oberkampf and Trucano, 2002): W.L. Oberkampf and T.G. Trucano, Verification and Validation in Computational Fluid Dynamics, SAND2002-0529, March 2002.

(Oden, 2001): J.T. Oden, private communication to the ASCI Non-Nuclear panel, June, 2001.

(Pao et al, 2000): K.I. Pao, B. DeVolder, J. Glimm, J. Grove, M. McKay, and D.H. Sharp, A Statistical Framework for Error Modeling for Hyperbolic Conservation Laws, presentation at 1st SIAM Conference on Computational Science and Engineering, Washington, DC, Sep 2000.

(Pao, 2001): K.I. Pao, private communication to NEDPC-01 audience, Nuclear Explosives Design Physics Conference, Livermore, CA, October 2001.

(Pilch et al, 2000): M. Pilch, T.G. Trucano, J.L. Moya, G. Froehlich, A. Hodges, and D. Peercy, Guidelines for Sandia ASCI Verification and Validation Plans — Content and Format: Version 2.0, SAND-2000-3101, January 2001.

(Roache, 1998): P.J. Roache, *Verification and Validation in Computational Science and Engineering,* Hermosa Publishers, Albuquerque, NM.

(Sam et al, 2001): D.D. Sam, S.S. Murty, and S.A. Perfect, Muscle Code Shootout II, A Scaling and Performance Study of the W-Program ASCI Finite Element Codes , LLNL CODTU-2001-0415, May, 2001.

(SAND, 1993): Sandia National Laboratories, Reliability Assessment Dept, Nuclear Weapon Reliability Methodology , SAND93-0704, June, 1993.

(Schwer, 2002): L. Schwer, private communication, August, 2002.

(Storch et al, 2001): N. Storch, C. Nitta, R. Klein, T. Quinn, S. Louis, D. Sam, E. Dube, L. Cook, D. Miller, and P. Miller, LLNL Site-Specific ASCI Software Quality Engineering Recommended Practices , Version 1.0, LLNL UCRL-ID-143698, May, 2001.

(Trucano et al, 2001): T.G. Trucano, R.G. Easterling, K.J. Dowding, T.L. Paez, A. Urbina, V.J. Romero, B.M. Rutherford, and R.G. Hills, Description of the Sandia Validation Metrics Project , SAND-2001-1339, July 2001.

(Trucano et al, 2002): T.G. Trucano, M. Pilch, and W.L. Oberkampf, General Concepts for Experimental Validation of ASCI Code Applications , SAND-2002-0341, March 2002.

(Trujillo, 2001): D.A. Trujillo, private communication, Black Mesa Example, October, 2001.

(Wood, 2000): M.M. Wood, private communication, 16 November, 2000.

9 Distribution

LLNL (64):

T.F. Adams	L-095
J.D. Alvarez	L-096
T.S. Carman	L-031
E.A. Chandler	L-170
T.F. Chen	L-125
S.K. Chidester	L-125
R.B. Christensen	L-160
D.L. Crawford	L-066
A.H. DePiero	L-125
E.I. Dube	L-095
J.P. Dow	L-125
P.O. Egan	L-097
M.F. Foltz	L-125
D.H. Fujino	L-096
B.T. Goodwin	L-160
R.M. Greenman	L-183
P.J. Harwood	L-125
H.S. Hopkins	L-125
T.A. Horrillo	L-125
J.J. Hsu	L-096
M.M. Hurd	L-005
K.C. Johnson	L-160
B.R. Jorgensen	L-125
J.M. Kenneally	L-231
R.I. Klein	L-023
B.E. Knapp	L-125
A.S. Lee	L-125
T.Y. Lo	L-125
R.W. Logan (6 copies)	L-125
G.L. Mara	L-466
T.L. McAbee	L-095
C.F. McMillan	L-170
G.H. Miller	L-466
B. Moran	L-096
S.S. Murty	L-125
T.A. Nelson	L-125
R.L. Newmark	L-208
C.K. Nitta (6 copies)	L-096
D.A. Nowak	L-160
M.L. Olsen	L-125
S.A. Perfect	L-126
D. Pilkington	L-095
J.A. Rathkopf	L-030
D.D. Sam	L-125
D.E. Stanfel	L-031
S.M. Taylor	L-125
S.R. Trost	L-119
J.V. Tyler	L-160
C.P. Verdon	L-030
R.A. Ward	L-030
P.D. Wapman	L-125
L.G. Wiley	L-097
K.A. Winer	L-096
B.J. Wu	L-097

Los Alamos National Laboratory (51)
Mail Station 5000
P.O. Box 1663
Los Alamos, NM 87545

P.J. Adams, MS B220, X-2
E.D. Aragon, MS F633, ADWEM/WS
J.B. Aubrey, MS F663, X-5
J.M. Booker, MS P946, ESA-EA
S.W. Doebling, MS P946, ESA-WR
C.L. Faust, MS C936, ESA-WE
D.G. Flicker, MS F664, X-4
M.D. Garcia, MS C936, ESA-WE
S.P. Girrens, MS P946, ESA-EA
F.M. Guerra, MS P946, ESA-WR
J.A. Guzik, MS B220, X-2
M.J. Haertling, MS F633, ADWEM/WS
M.S. Hamada, MS F600, D-1
E.M. Hanson, MS P946, ESA-DO
D.B. Harris, MS B220, X-2
A.R. Heath, MS F663, X-5
S.S. Hecker, MS G754, MST-DO
R.D. Idzorek, MS F633, ADWEM/WS
J.D. Immele, MS A148, DIR
D.L. Jaeger, MS F629, ADWEM
R.J. Juzaitis, AD-WPC
J.R. Kamm, MS D413, X-3
K.M. Keeler, MS C936, ESA-WE
M.W. Kirkland, MS F664, X-4
K.R. Koch, MS F661 X-DO
R.A. Kracjik, MS B218, X-DO
W.J. Krauser, MS B218, X-DO
R.W. Macek, MS P946, ESA-WR
M.R. MacInnes, MS B220, X-2
R. Mah, MS A107, AD-WEM
C.A. Mangeng, MS A148, DIR
D.R. McCoy, MS F652, AD-WPC
J.E. Morel, MS D409, X-6
W.R. Oakes, MS F645, X-8
K.I. Pao, MS F663, X-5
J. Peery, MS F652, ADWP/ASCI
W.J. Rider, MS D413, CCS-2
D.E. Post, MS B218, X-DO
J.J. Ruminer, MS P946, ESA-DO
L. Salazar, MS A107, ADWEM
S.L. Shalles, MS F633, ADWEM/WS
D.H. Sharp, MS B213, T-13
M. Sheppard, MS F664, X-4
P.A. Stevens, MS F633, ADWEM/WS
B.C. Trent, MS F664, X-4
D.A. Trujillo, MS F633, ADWEM/WS
D.L. Tubbs, MS B220, X-2
R.L. Wagner, MS A153, DIR
L.E. Witt, MS F633, ESA-DO
M.M. Wood-Schultz, MS B220, X-2
R.B. Wood-Schultz, MS B218, X-DO

Distribution (continued)

Sandia National Laboratories (25)
Albuquerque, NM 87185
Livermore, CA 94550
Attn:

H.J. Abeyta MS 0453 2101
T. C. Bickel MS 0841 9100
R.L. Bierbaum MS 9202 8205
E.P. Chen MS 9161 8726
J.L. Handrock MS 9042 8727
D. R. Henson MS 9007 8200
C. M. Hinckley MS 9013 8231
T.O. Hunter MS 0151 9000
D.M. Kwon MS 9014 8242
S.E. Lott MS 0139 9905
W. L. Oberkampf MS 0828 9133
J. M. McGlaun MS 0835 9140
R. G. Miller MS 9007 8201
J.L. Moya MS 0824 9130
T.L. Paez MS 0557 9133
M. Pilch MS 0828 9133
J. Polito, Jr. MS 0130 2800
J. S. Rottler MS 0429 2100
P.A. Sena MS 0427 2104
J.H. Stichman MS 0457 2000
W.J. Tedeschi MS 0479 2113
T. G. Trucano MS 0819 9211
M.O. Vahle MS 0139 9900
P. J. Wilson MS 0139 9902
J.B. Woodard MS 0102 2

EXTERNAL DISTRIBUTION

Marvin L. Adams
Texas A&M University
Nuclear Engineering Dept.
129 Zachry Engineering Center
Mail Stop 3133
College Station, TX 77843-3133

Ted Belytschko
Department of Mechanical Engineering
Northwestern University
2145 Sheridan Road
Evanston, IL 60208-3109

John H. Birely
1341 Monterrey Rd NE
Rio Rancho, NM 87124

Thomas D. Burton
Dept. Chair
Texas Tech University
Dept. Mechanical Engineering
Box 41021
Lubbock, TX 79409-1021

John A. Cafeo
General Motors R&D Center
Mail Code 480-106-256
30500 Mound Road
Box 9055
Warren, MI 48090-9055

Steve Castillo, NMSU
Klipsch School of ECE
Dept. 3-0, Box 3001
Las Cruces, NM 88003

James C. Cavendish
General Motors R&D Center
Mail Code 480-106-359
30500 Mound Road
Box 9055
Warren, MI 48090-9055

Fred S. Celec
Deputy ATSD-NCB
Nuclear Matters
3050 Defense Pentagon
Rm 3C125
Washington, DC 20301-3050

EXTERNAL DISTRIBUTION (cont'd)

Department of Energy (11)
Attn:
Dave Beck, NA-12
Everet Beckner, NA-1
Glenn Bell, NA-115
Diane Bird, NA-115
David Crandall, NA-11
Joe Gazda, NA-122
B. Ed Pate, NA-114
Maurice Pitt, NA-115
William Reed, NA-114
Jamileh Soudah, NA-114
Jim Van Fleet, NA-113
1000 Independence Ave., SW
Washington, DC 20585

Yvonne DesLauriers, NWCA
AF NWCA/WM
1551 Wyoming Blvd. SE
Kirtland AFB
Albuquerque, NM 87117-5617

Virginia T. Dobey
Acton Burnell, Inc.
DMSO
1901 N. Beauregard St.
Suite 500
Alexandria, VA 22030

Robert G. Easterling
7800 Northridge NE
Albuquerque, NM 87109

Ashley Emery
Dept. of Mechanical Engineering
Box 352600
University of Washington
Seattle, WA 98195-2600

James M. Ferguson
1856 Grandview Dr.
Oakland, CA 94618

John Fortna
ANSYS, Inc.
275 Technology Drive
Canonsburg, PA 15317

Mike Giltrud
Defense Threat Reduction Agency
DTRA/CPWS
6801 Telegraph Rd.
Alexandria, VA 22310-3398

Stan Gooch
USSTRATCOM / J53
901 SAC Blvd. STE 2E9
Offutt AFB, NE 68113-6500

Alejandro Graf
Advanced Technical Vehicle Operations
Ford Motor Co.
MD116-PDC
20901 Oakwood Blvd.
P.O. Box 2053
Dearborn, MI 48121

James Gran
SRI International
Poulter Laboratory AH253
333 Ravenswood Ave.
Menlo Park, CA 94025

Barry W. Hannah
Director, SSP
Nebraska Ave. Complex
287 Sommers Court NW
Suite 10041
Washington, DC 20393-5446

Ted Hardebeck
AD, Strategic Planning
USSTRATCOM / J5
901 SAC Blvd.
Offutt AFB, NE 68113-6500

Tim Hasselman
ACTA
2790 Skypark Dr., Suite 310
Torrance, CA 90505-5345

Richard Hills
College of Engineering, MSC 3449
New Mexico State University
P. O. Box 30001
Las Cruces, NM 88003

William F. Hosford
Dept. Materials Science and Engineering
University of Michigan
Dow Building, North Campus
Ann Arbor, MI 48109

David Keyes
Old Dominion University
Dept. Mathematics & Statistics
Norfolk, VA 23529

Noboru Kikuchi
Dept. Mechanical Engineering
University of Michigan
G.G. Brown Building, North Campus
Ann Arbor, MI 48109

Michelle L. Kilkauskas
JASA Director
NAWCWD
1 Administration Circle
Code 418000D
Ridgecrest, CA 93555-6001

Hyoung-Man Kim
Boeing Company
M/S: ZC-01
502 Gemini Ave.
Houston, TX 77058

Edward W. Larsen
Dept. of Nuclear Engineering and Radiological Sciences
Cooley Bldg. North Campus
University of Michigan
Ann Arbor, MI 48109

Kaye Lathrop
190 Cedar Lane East
Ridgeway, CO 81432

Robert O. Lewis
The Boeing Company
499 Boeing Blvd.
Huntville, AL 35824-6402

Robert F. Lucas
Chairman
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Suite 1001
Marina Del Ray, CA 90292-6695

Robert Lust
General Motors, R&D and Planning
MC 480-106-256
30500 Mound Road
Warren, MI 48090-9055

Lynda K. Liptak
USAF
3550 Aberdeen Ave., SE
Kirtland AFB, NM 87117

B.N. Maker
LSTC
7374 Las Positas Road
Livermore, CA 94550

Hans Mair
Institute for Defense Analysis
Operational Evaluation Division
1801 North Beauregard Street
Alexandria, VA 22311-1772

William R. Martin
Dept. of Nuclear Engineering and Radiological Sciences
Cooley Bldg. North Campus
University of Michigan
Ann Arbor, MI 48109

Dan Meiron
California Institute of Technology
MC 217-50
Pasadena, CA 91125

Paul C. Messina
Director
Center for Advanced Computing Research
California Institute of Technology
1200 E. California Blvd.
Pasadena, CA 91125

Kim Molvig
Dept. Nuclear Engineering (24-206)
Massachusetts Institute of Technology
77 Massachusetts Ave.
Cambridge, MA 02139

B. Mullins
HQ USAF/XON/NWCA
1480 Air Force Pentagon
Washington, DC 20330-1480

Ahmed Noor
University of Virginia
NASA/Langley Research Center
MS 369
Hampton, VA 23681-0001

J. Tinsley Oden
Texas Institute for Computational and Applied Mathematics
Mail Code C0200
University of Texas at Austin
Austin, TX 78712-1085

Dale Pace
Applied Physics Laboratory
Johns Hopkins University
11100 Johns Hopkins Road
Laurel, MD 20723-6099

Alex Pang
Computer Science Department
University of California
Santa Cruz, CA 95064

Allan Pifko
2 George Court
Melville, NY 11747

J.N. Reddy
Dept. of Mechanical Engineering
Texas A&M University
ENPH Bldg., Rm 210
College Station, TX 77843-3123

Patrick J. Roache
1108 Mesa Loop NW
Los Lunas, NM 87031

D. Robbins
HQ USAF/XON/NWCA
1480 Air Force Pentagon
Washington, DC 20330-1480

Sunil Saigal
Carnegie Mellon University
Dept. of Civil and Environmental Engr.
Pittsburgh, PA 15213

Len Schwer
Schwer Engineering & Consulting
6122 Aaron Court
Windsor, CA 95492

Paul Senseny
Factory Mutual Research Corporation
1151 Boston-Providence Turnpike
P.O. Box 9102
Norwood, MA 02062

Mark Shephard
Rensselaer Polytechnic Institute
Scientific Computation Research Center
Troy, NY 12180-3950

Don Simons
Logicon
222 W. 6th St.
P.O. Box 471
San Pedro, CA 90733-0471

T.P. Slavik
LSTC
7374 Las Positas Road
Livermore, CA 94550

S.C. Tang
Scientific Research
Ford Motor Co.
MD3135-SRL
P.O. Box 2053
Dearborn, MI 48121-2053

Ben Thacker
Southwest Research Institute
6220 Culebra Rd.
P. O. Drawer 28510
San Antonio, TX 78284-0510

Susan B. Turner
BWXT Y-12
P.O. Box 2009
Oak Ridge, TN 37831-8238

Reid Worlton
2031 Roosevelt Court
Grand Junction, CO 81503

Ren-Jye Yang
Ford Research Laboratory
MD2115-SRL
P.O. Box 2053
Dearborn, MI 48121

Sidney Yip
Dept. Nuclear Engineering (24-208)
Massachusetts Institute of Technology
77 Massachusetts Ave.
Cambridge, MA 02139

Simone Youngblood
DOD/DMSO
Technical Director for VV&A
1901 N. Beauregard St., Suite 504
Alexandria, VA 22311