

Risk Reduction as the Product of Model Assessed Reliability, Confidence, and Consequence

Roger W. Logan
rwlogan@llnl.gov

Cynthia K. Nitta
nitta1@llnl.gov

Steven K. Chidester
chidester1@llnl.gov

University of California
Lawrence Livermore National Laboratory
Livermore, CA 94551

This paper presents our methodology for verification and validation (V&V) of models, coupled to quantification of risk and risk reduction, and the use of risk, quantified as a dollar value, in the value-engineering and decision trade-off process. We begin by defining a simple measure of quantified risk, as the numerical product of three basic terms: the lower-bound model assessment of a product's reliability, multiplied by the confidence level used in the assessment, multiplied by the actuarial consequence value of the risk or risk reduction. We discuss our process for obtaining each of these three terms, and of other terms needed to complete the quantification process. An explosives impact test example is used to show the effects of model choice, data, and uncertainty methods on the resulting quantified risk reduction value. We then show how we use the quantified risk reduction value in a benefit/cost ranking method of value engineering. In this way, we can follow the simulation results from our explosive impact model all the way from V&V to risk to the investment trade-off decision process. Naturally, the nature of most such methodologies is still evolving, and this work represents the views of the authors and not necessarily the views of Lawrence Livermore National Laboratory.

Keywords: Verification, validation, V&V, confidence, reliability, risk, benefit/cost ratio, BCR

1. Introduction: The NNSA-TriLab V&V View: Credibility in Stockpile Modeling

The unprecedented growth in computing power available today is leading to greatly increased use of computer simulations. In order to establish and retain trust in these simulations, it is important to be able to assess the accuracy of the simulation. This assessment is often formalized through processes known as model verification and validation, or model V&V. But even if we have assessed the accuracy of our simulations and gained trust in them, there remains the rapidly growing

frontier concerning what to do next. The potential to use more numerous and complex simulations to provide model-based assessments of reliability and risk is tempting, with today's computational tools and relatively inexpensive computing capacity. However, we must establish trust in any linkage we make from our simulation to a risk analysis, just as we establish trust in the simulation in its own right. In this work, we provide one example of a path or framework to make this linkage from computer model, to V&V, to reliability, and then to risk analysis. Naturally, the path we illustrate is not unique, but represents one path forward.

The National Nuclear Security Agency (NNSA) and TriLabs (Los Alamos, Sandia, and Lawrence

Livermore) recently published a consensus summary of V&V within the DOE-NNSA “Advanced Strategic Computing” (ASC) Program Element. We repeat the language verbatim in this introduction so as to retain the terminology as agreed [1]:

Why:

The ASC Verification & Validation (V&V) program exists to establish a technically rigorous foundation of credibility for the computational science and engineering calculations required by the NNSA Stockpile Stewardship Program. This program emphasizes the development and implementation of science-based verification and validation methods for the support of high-consequence decisions regarding the management of the U.S. nuclear stockpile. The V&V process reduces the risk of incorrect stockpile decisions by establishing that the calculations provide the right answers for the right reasons.

What:

V&V is the multi-disciplinary process of demonstrating credibility in simulation results. Credibility is built by collecting evidence that (1) the numerical model is being solved correctly and (2) the simulation model adequately represents the appropriate physics. The former activity is called Verification and requires intimate knowledge of the mathematical model representing the physics, the numerical approximation derived from that model, software quality engineering (SQE) practices, and numerical error estimation methods. The latter, termed Validation is accomplished by comparison of simulation output with experimental data and quantifying the uncertainties in both. Broad knowledge of modeling and experimentation, augmented with a deep understanding of statistical methods, are necessary for Validation.

Impact:

Computer simulations are used for analysis of all aspects of weapon systems, as well as the analysis and interpretation of experiments. The credibility of our simulation capability is central to the credibility of the certification of the nuclear stockpile and is established through rigorous and quantitative V&V analyses. Regardless of whether or not we return to nuclear testing, V&V *establishes* credibility by providing evidence to support questions such as, “Why should we trust the simulation’s results?” Insufficient confidence or credibility in our simulations, will lead us to an incorrect decision pertaining to the reliability, performance and safety of the nuclear weapons.

Delivery:

The V&V process delivers traceable, reproducible, and formally reviewed conclusions supporting confidence in stockpile simulations, to include, but not limited to:

- Documented analysis and conclusion of the confidence level of the models as a result of the V&V activities
- Repository of test results associated with unit/regression/system tests, verification and validation tests. Additionally, list of test data used

- Documented code (feature) and solution (model) verification
- Documented V&V environment (constraints, assumptions, and tools)
- Repository of code versions, fixes and other patches used during the V&V phase.

Integration:

The ASC V&V process provides the most quantified evidence available to the programs laid out in the Nuclear Posture

Review:

- Validated stockpile simulations as certification evidence for the Annual Assessment Review.
- Validated simulations to assess and help quantify the benefit or impact of changes in Stockpile Life Extension Plan (SLEP) content and timing.
- Quantitative Validation of pit performance simulations, as inputs to the high risk assessment of the need and timing for a Modern Pit Facility.
- Methods and assessment of predictive adequacy, as we face challenges of certification of Advanced Concept Initiative weapon system certification under the Comprehensive Test Ban Treaty (CTBT) era.
- Quantify the model assessment benefit of a return to nuclear testing using V&V methods and quantitative model validation.

V&V Enables Credibility in Simulation Capabilities via These Steps

1) Planning

- *Identify* and document the programmatic and weapons systems needs that drive the detail and accuracy requirements of the physics and engineering calculations.
- *Ensure* that planned activities are sufficient to meet requirements; document the gaps.
- *Review* archived data, and plan small-scale experiments and tests needed to meet the V&V requirements.
- *Ensure integration* with other ASC program elements, experimental campaigns, and directed stockpile work.

2) Verification

- *Confirm the implementation* of sound software quality practices to ensure that the codes are sufficiently free of defects and errors.
- *Conduct Code/Algorithm Verification* to assure the code is correctly solving the mathematical equations.
- *Perform solution verification* to provide evidence that the time and space discretization of the mathematical model is adequate for the intended application.

3) Validation

- *Align* planned validation experiments with the needs of the mathematical model and the application.

- Assess model accuracy with respect to experimental data to provide evidence that:
 - The *correct* physical equations are being solved; and
 - The *important* time-dependent geometric, material, and thermodynamic features of the weapon and its components are adequately represented.
- Quantify the predictive capabilities and limitations of the models within their database and in the application parameter space.

4) Predictive Adequacy Estimation

- Present evidence of an adequate understanding (through expert judgment and quantitative rigor) to make defensible risk-informed stockpile decisions.
- Provide adequate understanding of variabilities and/or uncertainties and quantify their integral impact on design margins.
- Validate sensitivities to identify which physical phenomena most influence weapon safety and performance.

The NNSA-TriLab four-step procedure (1, 2, 3, 4) corresponds identically to the “ABCD” V&V procedure we will introduce in section 3 below.

2. From V&V to Reliability to Risk

One way to take the V&V process from requirements to V&V and eventually to the investment strategy is shown in Figure 1. The V&V process consists of modeling and experimental data activities sufficient to establish

models with confidence (C) bounds on uncertainty (U). We then obtain margins (M) and reliability (R) equivalents, and quantified reliability at confidence (QRC) [2]. Margin (M) is simply factor-of-safety (FOS) minus one, where FOS equals capability divided by requirement. In this simplest implementation linking QRC to risk and benefit/cost, QRC is defined as *model assessed reliability* (R) times the assessed confidence (C) in the model, times 1000, a multiplier of convenience. This leads to a maximum value for QRC = 1000, and a minimum of QRC = 0. Lacking a perfect world with infinite data and infinite knowledge, we prefer to think in terms of $0 < QRC < 1000$ so that it is not confused with a true system reliability R_s , where $0 < R_s < 1$. This *deterministic* interpretation of QRC forms one axis of our risk diagram (Figure 2) in the simple case described here. Multiplication of reliability times confidence is one simplified way to avoid spurious results that may be obtained by attempting risk analyses with a deterministic value of model assessed reliability *without* accounting for confidence. The simple multiplicative implementation of QRC used here is one way to avoid this trap and opens the door to more complex risk and decision analyses where confidence is treated more thoroughly [3].

A value engineering–related quantified systems value (QSV_0) per unit time, t , is defined and then adjusted as a function of QRC, with the economic function of present value factor (PV_F) in the time (t) domain. PV_F falls from $PV_F = 1.0$ to $PV_F = 0.0$ over time, as a function of the assumed discount rate. The discount

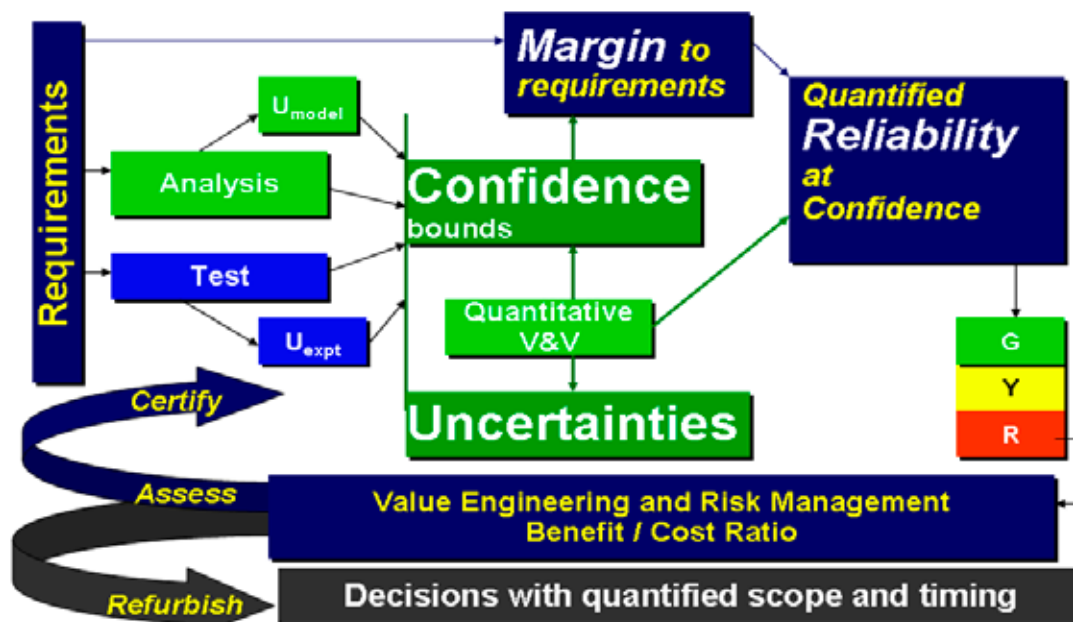


Figure 1. Flow diagram from system requirements through V&V, through uncertainty quantification and margins; onward through QRC, then value engineering and risk management via benefit/cost ratio. More detail is described by Logan and Nitta [2].

rate may be chosen to represent the traditional time value of money, coupled with a partial representation of uncertainty about the value of a product or credibility of a scenario far out into the future:

$$\Delta QSV = QSV_0 / 1000 \int_t PV_F[t] \Delta QRC[t] dt \dots \quad (1)$$

In terms of linkage to risk, the QSV_0 term can be thought of as the *risk consequence*; the value of the system working, not working, or the negative consequence of an accident or a system failure. The QRC term is used here as a simple form of *risk likelihood*. In this way, ΔQRC compares the difference between two assessments of likelihood; perhaps one with a better model, or one with more data, etc. Then ΔQSV becomes the assessed change in value or reduction in risk due to our investment in a better model, in more data, or to a mitigation of risk likelihood in general.

Key to the investment strategy process, and its linkage back to V&V, is the benefit/cost ratio (BCR): benefit (\$B) is proportional to risk reduction, expressed as ΔQSV . With the above relation, $\Delta Risk$ is proportional to ΔQSV , and hence to ΔQRC ; and finally ΔQRC links to the fidelity of our quantitative V&V statements (confidence bounded uncertainties). Then BCR is (Benefit - Cost)/Cost. The BCR is well accepted as a key to the investment strategy process [3]; its linkage back to V&V is provided by the QRC analysis. Quantified V&V shows us that there is not a unique BCR—we must explore its bounds for any given decision. Due to the non-uniqueness of any given BCR, we suggest that a decision input consider three 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

These “V&V bounds” may be the result of a non-deterministic treatment of the BCR as in Logan et al. [3] and other works, or due to generation of V&V and reliability-related quantities using several different models and model forms. The terms “high BCR” and “low BCR” are relative, and the absolute values of a given BCR depend on details that are often as local as internal accounting systems and consensus values for product (or accident) value. The goal of linking the model to the BCR is to enable informed decisions by comparing the relative BCR of multiple options assessed with a consistent modeling, V&V, and risk analysis process. The risk management and BCR activities depicted in Figure 1 may either lead to product certification, or to refurbishment, or to a BCR-based justification for additional experimental

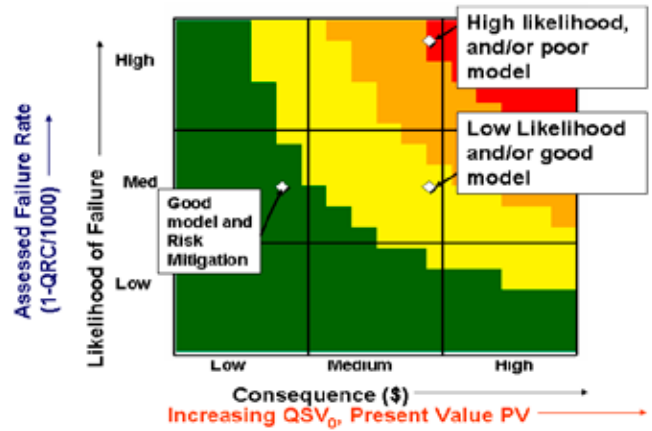


Figure 2. Dollar benefit of V&V and quantitative certification, expressed as a standard Risk = Likelihood × Consequence matrix. For the sake of simplicity of analysis demonstration we use a symmetric risk matrix. Likelihood becomes analogous to assessed (1 - QRC/1000); consequence is typically expressed in dollar terms.

and modeling activities during another product assessment cycle.

The end product of the roadmap shown in Figure 1 is the certification of a product. Implicit in this product certification is the accreditation of the simulation including the computer code, model, V&V, application, and the expert judgment employed throughout. The end product and dollar benefit can be explained using a Risk = Likelihood × Consequence matrix as shown in Figure 2. A better model assessment can reduce assessed risk, either by reducing assessed likelihood or by showing opportunities to mitigate risk consequence. V&V plays a quantified role, one that is now directly proportional to risk reduction and value engineering quantities. We will relate each aspect of Figure 1 and Figure 2 to an example of impact testing of explosives. In this example, our goal is to quantify and bound the reaction thresholds, so that we can quantify our model assessed likelihood of avoiding an accidental explosive reaction of any level. Our goal will be to assess the likelihood of avoiding any explosive reaction; this is a far tighter requirement than simply avoiding an accidental explosive detonation.

Before we proceed with the V&V process, we have to know the requirements our product or system will have to meet, and which of these our model is to address. After that, depending on both the fiscal and scientific ability to perform a certain level of V&V assessment, we proceed with various degrees of qualitative and (ideally) quantitative validation.

2.1 Qualitative Validation Ratings

We emphasize our preference for *quantitative* V&V with numerical confidence bounds, but we also recognize the need for *qualitative* assignment of the V&V level for particular simulation capabilities. For qualitative V&V, we suggest the use of 0–10 rating scales such as those discussed by Logan and Nitta [2] or those for validation adequacy status given on a 0–5 scale from Trucano et al. [4]. We discuss these scales, which may be normalized to a unit scale to give a qualitative V&V rating of $0.0 < \text{VER} < 1.0$ for verification (VER), and $0.0 < \text{VAL} < 1.0$ for validation (VAL) [5].

2.1.1 Qualitative V&V Ratings

For qualitative V&V, we suggest the use of 0–10 rating scales such as those discussed by Logan and Nitta [2], for example. 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 qualitative 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 time-constrained 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 detailed results 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 *at a minimum* some kind of graded scale V&V measure is used. Such a graded-scale, single-valued *qualitative* V&V indicator may in fact represent the pedigree of a very thorough and *quantitative* V&V process. This qualitative process “0–10” rating is not related to or indicative of any of the actual numbers resulting from quantification of uncertainty, confidence, or reliability we discuss later on. The 0–10 rating is only a number representing the level of V&V *process* that was used to generate any numbers in the quantitative assessments of uncertainty and confidence.

There is nothing wrong with using a lower quality or conceptual analysis to make a point or point out an area of risk. However, to avoid having the audience take such examples with verbatim precision, a verification (VER) and validation (VAL) meter or equivalent *as a minimum* should be used.

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

- Reflect the inclusion of “expert judgment-based” V&V aspects;
- Motivate a more scientific V&V process; because, hopefully, audiences and decision makers will support the quantitative activities needed to raise a low “V&V meter level” to a higher one;
- Motivate a more scientific decision process;
- Provide fundamentals for rational discourse on quantitative V&V details;
- Provide a rational basis for common understanding and expression of V&V level; and
- Provide an expression of *relative* information and level regarding V&V.

A VER meter with a scale that reads 0–10 is simple, and it should be fairly clear in intent. We suggest that the following factors help in setting this somewhat subjective but informative 0–10 verification scale *for each code feature*.

Reading of 0–1.5:

- *Code or feature has a theoretical document and user’s manual*
- *Version control*
- Software quality engineering (SQE) guidelines
- Extensive code coverage regression

Reading of 1.5–3.5:

- Basic verification suite

Reading of 3.5–7.5:

- Most of element types verified
- Most of materials verified
- Most of contact (modeling collisions and friction) verified
- Order of convergence demonstrated

Reading of 7.5–9.9:

- Most of couplings (e.g., thermal-mechanical or thermal-chemical couplings) verified
- More complex verification studies such as *method of manufactured solutions* (MMS) [6]

Naturally, any such scale reading has to take into consideration the features as used for the application regime and fidelity of interest. For example, we use the term “most” regarding some of the features and verification activities; whether “most” means 60% of the features, 90% of them, etc. are criteria that depend on a given application and accreditation process. Obviously such a meter is still subjective, and still qualitative in how we specify the value of the meter. We should of course remember that this simple 0–10 scale is only an attempt to summarize (and compel) a more rigorous verification (and then validation)

process. More desirable are quantitative *verification statements* that describe the order of convergence of a given code feature, the ultimate accuracy of the answer that may be achieved, and the ability to obtain a given accuracy compared to analytical results when this code feature is used alone or in combination with other code features.

Like the VER meter, we introduce the concept of a VAL meter, with a scale that reads 0–10. The following factors are helpful in setting this somewhat subjective but informative 0–10 validation scale *for each model and application*.

Reading of 0–3.0:

- Runs the first time step
- Runs desired model to completion
- Obtains an answer: “blind trust” fidelity
- Calibrated model (the model can be *calibrated* to give an acceptable answer, but with perhaps non-physical “tuning” of model parameters in a way that precludes quantification of uncertainties or confidence)

Reading of 3.0–5.5:

- Solution verification (space, time, and iterative domain)
- Sensitivities qualitatively correct

Reading of 5.5–7.5:

- Integral validation to more than one system level test
- Hierarchical validation to numerous subsystem tests
- Integral or hierarchical validation across different systems
- Quantitative validation statement (confidence bounded uncertainties over the range of the experimental data)

Reading of 7.5–9.9:

- Predictive validation bound assessed (confidence and prediction intervals where we have data are quantitatively extended to application regions where we may lack data)
- All uncertainty terms quantified
- *Validated* sensitivity slopes
- “Fully validated” (an oxymoron; we can only hope to quantify what is *sufficient* validation)

The V&V process is much deeper and more quantitative than any single summary number can depict. But, the 0–10 graded scale meters go beyond a “yes/no” V&V statement for communicating fidelity quickly.

Another qualitative scale for validation adequacy status has been given on a 0–5 scale from Trucano et al.

[4]. A slightly modified and annotated version of this 0–5 scale is presented here:

- 0/5 = 0.0: (Inadequate): No significant comparisons with experimental data—and therefore no measure of correspondence with any such data. These are sometimes very preliminary “what if” analyses. They can be useful as a guide for the next set of analyses, but it is exceedingly dangerous to base any design or certification decisions on them.
- 1/5 = 0.2: (Inadequate): Ad hoc comparison of experiment “pictures” with prediction “pictures.”
- 2/5 = 0.4: (Incomplete): Ad hoc (nonstatistical) comparisons of experimental data (that may or may not be statistically significant).
- 3/5 = 0.6: (Incomplete): Statistical comparison of data and calculations that does not quantify predictive capability of the model or correlation over the parameter space of the database. The degree of extrapolation (if any) may not be quantified. The database may not be statistically significant or fully relevant to the application. *For example, in Figure 3, the “database” is reflected in the square boxes on the graph. The experimental data falls between temperatures of 20°C and 170°C. There are enough data points (about 16) to be statistically significant, but we note that all these data were not measured at “standard conditions,” e.g., the composition, impact geometry, etc. under consideration for reliability and risk assessment. If we take temperature outside the range of 20°C to 170°C, we are “outside” the parameter space of the database, and we are contending that our validation has gone beyond Level 3/5 = 0.6.*
- 4/5 = 0.8: (Adequate): Predictive capability of the model or correlation is quantified over the parameter space of the database. The degree of extrapolation is quantified. *For example, in Figure 3, the solid lines of model prediction clearly go outside the experimental database. We can obtain the degree of extrapolation impact temperature directly from Figure 3, but the extrapolation of other impact condition quantities from the data is not obvious in Figure 3. There is a statistically significant database that is fully relevant to the application.*
- 5/5 = 1.0: (Adequate): Predictive capability of the model or correlation is quantified over the full parameter space of the application. *As implied in Figure 3, the application (temperature range conceivably from 0°C to 200°C) may extend well beyond the parameter space of the database (impact temperature of 20°C to 165°C). There is*

a statistically significant database that is fully relevant to the application. We can contend that the analysis output shown below in Figure 3 meets this criterion, but that is not obvious from the limited information presented in that plot. We have not assessed any caveats or dangers of extrapolation, such as an autocatalytic reaction temperature at about 180°C, which will cause reaction in the absence of any impact at all! We cannot overstate the dangers of statistical extrapolation unless one is certain of a continuum of physics. Even if we had a continuum for extrapolation, quantitative validation of the model is not a statement that the model is adequate; it is only an attempt to supply the information needed for an adequacy assessment.

A similar “V = 0 to V = 5” level scale is proposed by Youngblood and co-authors [7], with a scale comparable to our 0–10 meters and the 0–5 scale of Trucano et al. [4].

We suggest that, as a first step in the V&V documentation process, such a rating system (e.g., 0–5, or 0–10, or even 2/5 = 0.4) be used. Within the documentation of the model and the verification and validation activities performed on it, an estimate of the V&V level of the work should be stated in the text by the authors and/or reviewers of the report.

We feel that some such overall rating is an important contribution to quantitative V&V for two major reasons:

- 1) Some type of overall rating is better than having an audience wonder if the work was at Level 0.10 (1 out of 10 on the V&V meter scale) or Level 0.95 (9.5 out of 10 on the V&V meter scale).
- 2) The diligence of most analysts, having rated their own work at, say, Level 0.4, will make them ask themselves, “How could I make this work reach Level 0.6, and would that be worth the effort?” This will motivate analysis and documentation to move from *qualitative* validation into *quantitative* validation.

In this work, we will focus on a *quantitative* validation process, with an “ABCD” process, leading to a *qualitative* rating of VAL > 0.4, and a quantitative validation statement resulting from validation analysis depicted in Figure 3.

2.2 Quantitative Validation Statement

For validation to achieve a *qualitative* rating goal greater than Level 0.4, the work should include the simulations and analysis necessary to generate a *quantitative* validation statement (annotated referring to Figure 3 and Figures 5–8 for this example) supporting the definition of validation given by Cafeo and Roache [8]:

Validated Model: A model that has confidence bounds on the output (the middle solid line in the plot). A validated model output has the following characteristics:

- 1) The quantity of interest (output);
- 2) An estimate of the bias (i.e., confidence bounds are not centered around the model output); and
- 3) A set of confidence bounds (the outer solid lines in the plot, drawn at an assessed confidence level). The confidence bounds may be statistically based or have some other basis, as long as the assumptions and caveats are stated.

A validated model is one where evidence supports a quantitative validation statement such as [8]: “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.”

We stress that for *quantitative* validation, improvement lies not in what we can *sketch*, but in what we can *assess* quantitatively. The quantitative validation process need not replace the V&V level expressed by the “0–10 meter reading.” Rather, the quantitative validation process such as the “ABCD” V&V process described next is what allows us to credibly claim a high “V&V meter reading.”

2.2.1 Example Quantitative Validation for VAL ~ 0.5

In Figure 3, we show the results of a validation study example from a design safety study of the threshold velocity for the onset of *any* explosive reaction for heated explosives impacted by steel projectiles [9]. We will describe some of the details of the experiments, and then proceed to the validation of the model used to characterize them, including an assessment of its predictive capability.

In the experiments, impact tests were performed on heated energetic material samples to help us develop and validate accident scenario models involving combined thermal and impact conditions. To determine heated thresholds, Steven test targets (see Figure 4) containing explosives PBX 9404 or LX-04 were heated to the range of 150–170°C and impacted at velocities up to 150 m/s by two different projectile heads fired from a gas gun. The projectiles had masses of 1.2 kg and 1.6 kg, and spherical radius heads of 30 mm and 6.4 mm, respectively.

For the impact tests using the apparatus in Figure 4, the measured and ambient threshold velocities for explosive reaction are shown in Table 1 and Table 2. The following trends were observed:

- 1) Threshold velocity increases with temperature.
- 2) Threshold velocity increase with temperature is more pronounced for LX-04 (higher binder and lesser explosive content) than for PBX 9404.

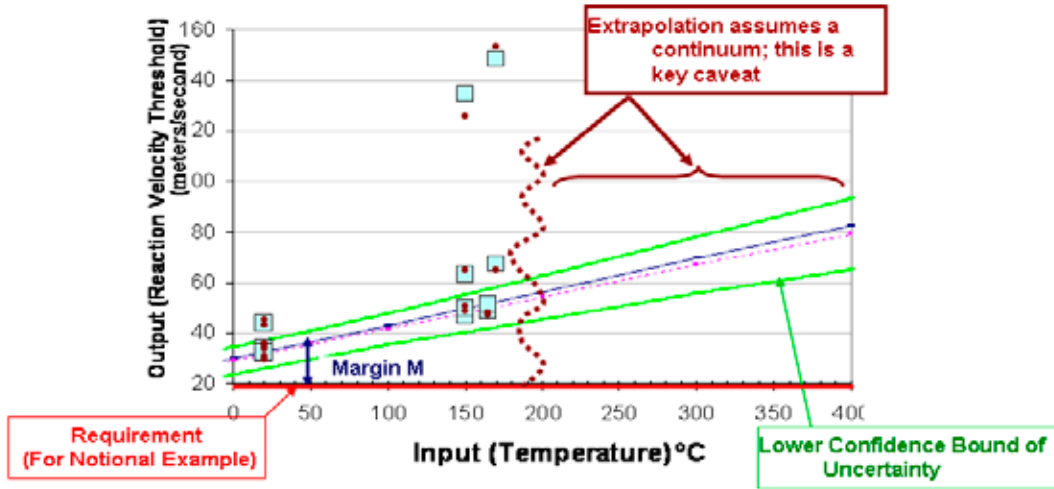


Figure 3. Linking requirements and V&V to risk management: V&V quantifies the agreement between test data (squares depict approximate test uncertainty) and model output (dots), then corrects to a standard operating condition (middle solid line), performs a bias correction (dotted line), and provides confidence bounded uncertainties (outer solid lines). These are tested against a requirement (e.g., the starting ordinate velocity), leading to model + system assessed reliability at confidence, the ordinate of the risk diagram in Figure 2. The wavy vertical line is an autocatalytic reaction temperature that violates our continuum physics assumption, one danger of extrapolating a model with limited physics.

- 3) Threshold velocity increases with projectile head radius.
- 4) The threshold velocity increase with radius is more pronounced for LX-04 and barely noticeable for PBX 9404.

In order to be useful in assessing potential accident scenarios, a predictive model for this thermal + impact experiment must, as a minimum, capture the four phenomena observed in the data within adequate confidence bounds (as depicted in Figure 3). In order to avoid undue physics “surprises,” the model should also be physically and geometrically based so as to capture the physical and geometric phenomena during the impact. We are developing finite element models to suit this purpose; for example, to capture the autocatalytic reaction “cliff” at about 180°C depicted by the wavy line in Figure 3. However, for this simple example we will use an empirical model, based on regression correlations, to capture the phenomena observed in the data using a few empirical “adjustables,” but not so many of these “tuning dials” as to preclude demonstration of the process leading to a quantitative validation statement.

Of course, any realistic model validation statement contains caveats and assumptions. Such is obviously the case with a regression-based model. As stressed in any standard statistics text, extrapolation of empirically based models can be particularly dangerous. Furthermore, several physical and chemical variables discussed in Switzer et al. [9] are ignored for this model validation; we plan eventually to account for these in our mechanical-thermal-chemical finite element model.

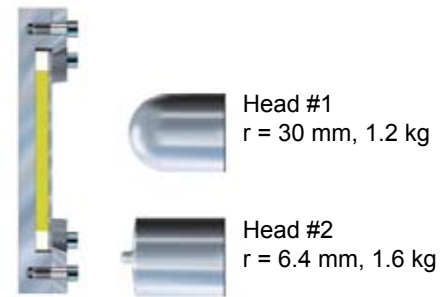


Figure 4. Steven impact test: 30 mm (Steven, Head #1) and 6 mm (Duff R/T, Head #2) projectiles

Nevertheless, the simple empirical model used here, comparing to nominal and even truncated threshold values, will serve our purpose in demonstrating the method and results of our quantitative validation process and will allow us to carry those results into a risk management construct.

We will use these experimental results, and a statistical validation method that does not preclude “tuning dials” but accounts for them statistically, to generate validated confidence bounds on a model for impacts over the regime of these experimental data.

The plot in Figure 3 is actually the end result of completing the ABCD validation process described next. The plot shows numerous small dots of experimental data (9+7=16 data points) for this Steven impact scenario. The scatter depicts the variation in head radius, explosive type, and test temperature. If our intended application of the impact model involves thermal conditions, we wish to quantify reliability and confidence during an impact scenario where impacts

Table 1. Data for Steven/Duff tests: Head #1, $r = 30 \text{ mm}^*$

Test #	High Explosive (HE) Type	Percent Explosive	Test Temp (°C)	Nominal Threshold Velocity (m/s)
1-1	PBX 9404	94	20	34.0
1-2	PBX 9404	94	20	36.0
1-3	PBX 9404	94	20	35.7
1-4	PBX 9404	94	150	50.5
1-5	PBX 9404	94	165	47.2
1-6	LX-04	85	20	45.0
1-7	LX-04	85	20	43.0
1-8	LX-04	85	150	*125.7
1-9	LX-04	85	170	*153.2

*caveats in Switzer et al. [9] (These tests showed no reaction at the test velocity shown, which was the highest achievable.)

Table 2. Data for Steven/Duff tests: Head #2, $r = 6.4 \text{ mm}^*$

Test #	High Explosive (HE) Type	Percent Explosive	Test Temp (°C)	Nominal Threshold Velocity (m/s)
2-1	PBX 9404	94	20	29.1
2-2	PBX 9404	94	150	48.8
2-3	PBX 9404	94	165	48.2
2-4	LX-04	85	20	30.7
2-5	LX-04	85	20	30.5
2-6	LX-04	85	150	64.8
2-7	LX-04	85	170	64.7

*caveats in Switzer et al. [9]

of a known velocity must be tolerated without reaction. We develop a model, providing the middle solid line, to answer this question. Quantitative validation requires a quantitative assessment of the confidence bounds on this model. This is determined by comparison to a known number of data points (N) as shown in the plot. The model error between the measured output and model output is generally too small to be shown in the graph above but is used, along with experimental error, variability, and assumed probability distribution functions (PDFs), to construct a confidence bound (the outer solid lines in Figure 3) on our analysis. Any adjustable (calibration) parameters used in the model must be counted as model degrees of freedom (K), so our effective number of data points becomes ($N - K$). Fewer data points (N) or more model adjustables (K) will result in wider confidence bound lines in the plot. Model adjustables are a fact of life; there is neither the time nor funds to avoid them all. It is simply important to quantitatively account for them in the validation assessment. If the model adjustables result in confidence bounds and uncertainties that are excessively wide or large, then an adequacy assessment as discussed below will compel further work to eliminate the adjustable parameters or quantify their range and distribution.

3. Summary of ABCD Validation

To begin to fulfill the formidable promises in the *quantitative validation statement* above, and allow us to progress through the map in Figure 1 leading from requirements through V&V through risk and benefit/cost, we need a quantitative process for V&V. One procedure we have used, and have seen used with success and clarity by others, can be described as an ABCD process for quantitative validation. We use the terminology “ABCD” to represent the steps in model validation, following the nomenclature first conveyed to us by Pilch [10]. The ABCD process can be summarized as follows.

3.1 “A” of ABCD Validation: Plan and Select Test and Model Matrix

During “A” of ABCD, we assess and plan tests and validation analyses. We do some preliminary scoping analyses to obtain a feel for whether there is sufficient independent data to enable validation in light of the fact that some amount of calibration (model tuning dials) is usually unavoidable. We try to obtain a “reasonable” fit of data to analysis—with minimal (and documented!)

model tuning; see Figure 5. If we cannot quantify the number of calibration parameters, then a quantitative validation will indeed be impossible. If there are too many “calibration tuning dials” or if their impact is too large (e.g., $K < N$, but K is significantly large) a quantitative validation will still be possible, but the confidence bounds will be so wide that they will only quantify the need for a better model and/or more data. (Strictly speaking, it is only *mathematically* necessary that $K < N$. However, general practice recommends that $K < N/3$ when considering the amount of data versus free parameters, and that $K \ll N/3$ if possible.) Notice that the raw data shown in Figure 5 are the same raw data overlaid on the V&V’d model of Figure 3. We can see from the overlay that the raw data may often appear to be “all over the map” in any given plotting plane, even with a good model fit as shown in Figure 5. This is because the model has a validation pedigree sufficient to account for numerous “cones of extrapolation” since the data set will typically vary more than one input parameter at a time.

3.2 “B” of ABCD Validation: Solution Verification

Solution verification involves an assessment of mesh convergence in spatial, temporal (and iterative convergence) domains (see Figure 6). The order of convergence should be established, and the proximity to convergence. This is used to quantify an estimate of the error and uncertainty of the discretization used, compared to infinitely fine discretization, with an explanation of how much error will be accepted due to the unfortunate necessity of not running all the runs fully converged. (With our simple regression model, solution verification is not an issue in this study).

3.3 “C” of ABCD Validation: Model Validation

The model validation process must account for experimental variability, measurement uncertainty, epistemic uncertainty (lack of knowledge) due to sparsity of data, model uncertainty, meshing uncertainty, and model discretization (weak form of finite elements) uncertainty. To do so, part “C” of ABCD involves performing the finite element analyses and then generating information such as that shown in Figure 7. That is, error and bias in the analysis versus experiment comparison must be quantified in each dimension of the application domain (i.e., input quantity) and for each decision quantity of interest (i.e., output quantity). The uncertainty bounds at specified, numerical estimates of confidence should be established (with the many caveats listed above including small sample studies and distribution forms assumed). Part “C” then enables a quantitative validation statement

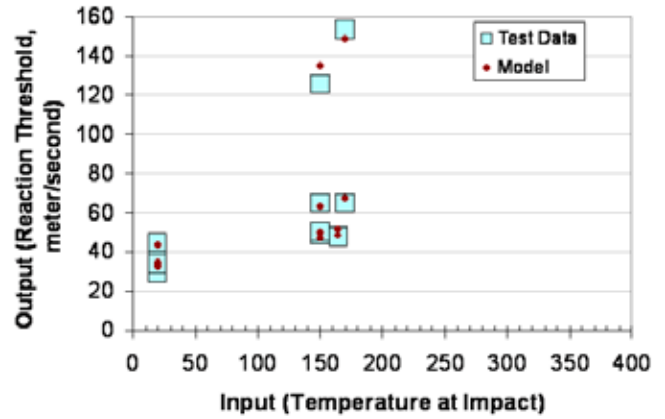


Figure 5. “A” of ABCD: Plan the validation with the data available; cursory model fit, may be *calibration* or *validation* as ABCD is completed

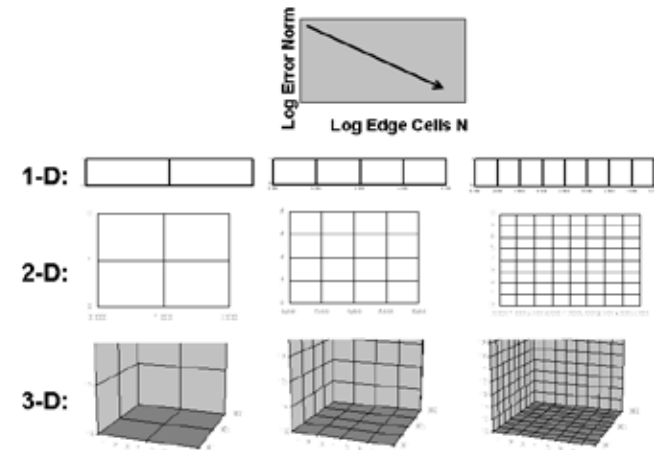


Figure 6. “B” of ABCD: Solution verification quantifies the proximity to convergence, and the model uncertainties we may be forced to accept

(uncertainty at a specified confidence) as defined above, but only over the domain of the data. Part “C,” the model validation portion of our ABCD validation, can be performed *top down* from the *integral* system level, or built in a *hierarchical* fashion (*bottom up* from component tests and models), or, ideally, both top down and bottom up. Since our actual application space is often outside the domain of the data, we must proceed one step further, to “D” of ABCD.

3.4 “D” of ABCD Validation: Predictive Adequacy

To assess *predictive adequacy*, it is necessary to go from quantified confidence over the *tested* domain (the data referent) to quantified confidence over the *application*

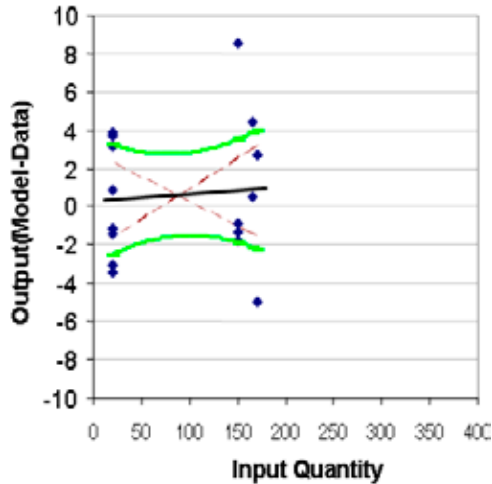


Figure 7. “C” of ABCD: Assess model against experiment, account for calibration “tuning,” and generate a quantitative validation statement with confidence bounded uncertainties—over the domain of the data (Steven test and model example)

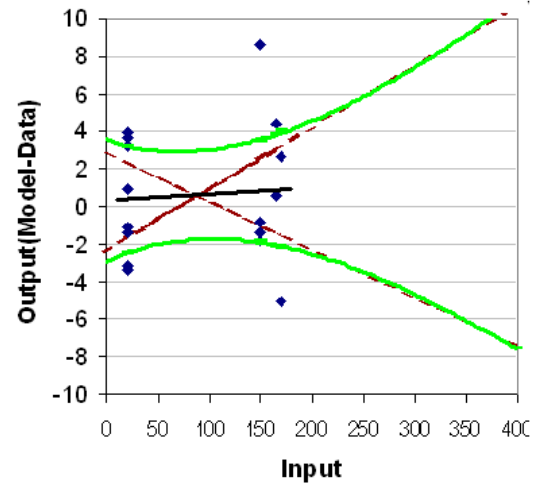


Figure 8. “D” of ABCD: Extend, using statistical methods and expert judgment caveats, the confidence (and prediction) intervals beyond the validation (data referent) domain, out to the application domain (Steven test and model example)

domainspace; Figure 8 shows an example where we only have validation data in an input quantity range of ~20 to ~160, but our application may be as high as ~400 for input quantity. To make quantitative statements about predictive adequacy, we must choose some statistical method to extrapolate beyond the data. Obviously our confidence beyond the data will be lower than within the data range. It should also be obvious that we must assure, to our best expert judgment, that there are no hidden “cliffs” or physics changes beyond the data, so our assumption of a smooth statistical extrapolation is at least reasonable. If this last condition cannot be met, we cannot make a credible quantitative validation statement outside the data range, and therefore certification or underwriting of a high-risk product in this range would be irresponsible. An example is the wavy vertical line shown in Figure 3. This represents the temperature of an autocatalytic reaction in the explosive that will occur even in the absence of any impact. This cliff or physics change is just beyond the referent but not captured by the simple empirical model used here. The statistical extrapolation hence loses meaning. Topics such as evidence theory [11] and judgment theory can make appropriate contributions to guide the future of this area.

3.5 Quantitative Validation Statement

To summarize, our goals on achieving “D” of ABCD are to obtain a quantitative validation statement. A validated model with a quantitative validation statement for performance analyses has the following

features, as depicted graphically in Figure 3 and discussed by Logan and Nitta [12]:

- The validation has a statistical nature, with small sample corrections often needed due to sparsity of experimental data especially at the full system test level.
- The validation must quantify uncertainty at a stated statistical confidence.
- The validation applies for a specific quantity of interest and application domain.
- The validation must show the origins of its data for comparison and, perhaps, the role of expert judgment to select the data and assess the comparison.
- Confidence-bounded “cones of assessed uncertainty” extend out to the application domain of interest as shown in Figure 3 and Figure 8; bounds are quantified with caveats.
- Software quality assurance (SQA), code verification, expert judgment, and planning (“A” of ABCD) help mitigate physics “surprises” during this extrapolation.
- The validation must allow us to assess a reliability measure from margin and assumed or known probability distribution function, normal distribution, or other. That is, the validation should provide confidence and prediction intervals with uncertainties at specified confidence levels. These quantities may be obtained by numerous methods (statistical or

other), but they are the quantities inherent in the “quantitative validation statement” and the quantities needed for subsequent reliability and risk analysis inputs.

- The validation must *provide the information* to address adequacy before stating whether a given model is “validated as adequate for its application” or not.

4. Achieving Predictive Adequacy

Achieving the ABCD validation sequence provides the quantitative information needed to make a decision about predictive adequacy. In the following section, we suggest one numerical metric that is used during the adequacy decision process. We call this metric *quantified reliability at confidence*, or QRC. With a specific requirement, and a quantitative validation statement (the output of ABCD validation using the process just described), we can use the metric QRC to help make our adequacy decision; that is, do we need to spend more time and money refining our model, data, and validation, or have we reached the point of diminishing (validation) returns for our application?

4.1 V&V and Quantified Reliability at Confidence (QRC)

Given a model expressed with this quantitative validation statement and sufficiently quantified information about the system requirements in its environments, we can then assess measures of QRC. Figure 3 above shows the result of the ABCD validation process mapped to the application, for a PBX 9404 explosive and a 30 mm head impact. In this case, the quantities plotted are the *input quantity*, the temperature of the target being impacted, and the threshold velocity for reaction (the *output quantity*). An overly simplified way to show this using the example in Figure 3 is as follows. If the required impact velocity to be survived for our application is 20 m/s, and we deploy our design over a temperature range of 50–200°C, our model assessment is that the minimum nominal velocity threshold will be about 35 m/s. Our nominal (50% confidence) margin (M) in velocity (m/s) can be expressed as

$$M = 35 - 20 = 15 \text{ (m/s)} . \quad (2)$$

Assuming a Gaussian distribution for this simple example, we assess the standard deviation of the fit to the data, and apply a small sample correction or coverage factor based on the number of data points, N , compared to the number of adjustables, K , as discussed

above. We then have uncertainty (U) expressed as a standard uncertainty (at 1 standard deviation):

$$U = \sigma = 5.7 \text{ (m/s)} . \quad (3)$$

Our quantitative validation method requires that any such U be evaluated at an assessed level of confidence. If we meet this requirement and for the simple example here assume a normal Gaussian distribution for error of model fit to the data in the plot of Figure 3, we can then use the statistical quantity Z as the reliability index, β , or

$$Z_{\text{qrc}} = \beta = (M / \sigma) / \sigma = (15) / 5.7 = 2.632 . \quad (4)$$

In equation (4), $X\sigma$ is the number X of Gaussian standard deviations, leading to percent confidence, C . In this case, for the numerator we have done the computation at the 50% one-tailed confidence level. That is to say, we assess that half the time the measured velocity threshold will exceed 35 m/s, and half the time it will not. Use of increased confidence level ($C > 50\%$) will make for a smaller numerator, especially with small numbers of experimental tests. Use of increased confidence level will also usually compel small sample corrections to the estimate of σ , the denominator. These corrections lead to a lower value of reliability index $\beta = Z_{\text{qrc}}$, meaning more area of the PDF is below the requirement level (e.g., 20 m/s velocity). This in turn gives a lower reliability, R , since R in this case is simply the inverse of the standard normal distribution ϕ , or $R = \phi^{-1}(\beta)$, for increased confidence, C . At some value of confidence, C , and the associated reliability, R , the QRC product reaches a unique maximum. This is the unique measure we call the QRC number (scaled by 1000 \times , so maximum QRC = 1000). In our example, with $C = 50\%$, $R = 99.6\%$, or QRC = 500, if we iterate with various values of confidence, C , expressed here as the number of sigmas (σ), we find that:

$$\begin{aligned} \text{QRC} &= 500 \quad \text{at } 0.0 \sigma, \text{ or } 50\% \text{ 1-tailed confidence,} \\ \text{QRC} &= 841 \quad \text{at } 1.0 \sigma, \text{ or } 84\% \text{ 1-tailed confidence,} \\ \text{QRC} &= 912 \quad \text{at } 1.5 \sigma \text{ (maximum),} \\ \text{QRC} &= 298 \quad \text{at } 2.0 \sigma, \text{ or } 97.7\% \text{ 1-tailed confidence.} \end{aligned}$$

The use of a pre-defined expression for our chosen PDF form, such as Z_{qrc} , taken directly from the statistical standard normal variable, Z , has several advantages. Most important is that it leads to *quantified reliability (R) at confidence (C)*, QRC. The value of R in this simple example is the one-tailed area under the statistical Z -curve up to the computed value of Z_{qrc} from equation (4). The value of C determines the breadth of the bell curve, or PDF. Expressing confidence, C , as a percent is

an attempt to quantify the fact that we can only validate to a sample distribution (the data we actually have) instead of the population distribution (the data that would exist over the lifetime of product application). Simple expressions enabling a QRC analysis can also be developed for other forms of PDF such as the uniform distribution.

4.2 QRC as Lower Bound “Model + System”

It is vital that we remember that QRC is not the “system reliability.” Were our model perfect and the data used in the validation plentiful with complete relevance, such would be the case. *Rather, we view the QRC number as a lower bound assessment of the (model + system) reliability at confidence.* An improvement (or decrement) in QRC may represent a change in the physical system, or simply a change in our model’s assessment credibility in that assessment. When we assess QRC at the system level, we can combine this with measures of quantitative system value (QSV, see equation (1) and discussion) and place these values on a risk diagram for investment and decision inputs. The whole process therefore lets us:

- Relate margin (M) and uncertainty (U) to R.
- Associate reliability (R) with a confidence (C)—and quantifies that C based on the specific number of model and test data quantities used for the validation.
- Show how better assessments of M and U—and increasing the “effective number of data points” (N) or reducing the number of “tuning dials” (K) quantitatively tightens U, allowing higher quantified C.
- Measure quantitatively our progress in charts like Figure 3.
- Use the quantities in Figure 3 (QRC) as decision inputs via the risk diagram in Figure 2.

4.3 QRC and Benefit/Cost Ratio (BCR)

After the QRC portion of the analysis, we use these quantities in a model for quantifying business decisions based on inputs from V&V and QRC, with 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). For a simple example, consider a product or event with an assessed dollar value QSV_0 (see the horizontal axis on the risk matrix in Figure 2). If the assessed value of the product is proportional to its *lower-bound model assessed reliability* (expressed as the unitless QRC), and if we fix present value term

$PV_F = 1.0$ in equation (1), then the value assessment of the product from our validated model is simply, for a QRC that is constant over a chosen lifetime Δt ,

$$QSV = QSV_0 / 1000 \times \Delta t \times QRC . \quad (5)$$

This simple example contains a number of assumptions, for example that there is no undue penalty for the instances (QRC < 1000) where the model assessment indicates the product does not perform as required. The benefit, \$B, of having the product is then this QSV value (our product is assessed to work). We can improve benefit, \$B, in at least two ways: either by improving the physical product (and hence the next assessed QRC), or by improving the model—refining and lowering uncertainty and hence also raising the *assessed* QRC. The latter is important because it allows us to attach a direct dollar benefit, $\Delta \$B$, to the V&V process. Of course, either improving the physical product (tighter manufacturing tolerances, etc.) or improving the model (V&V, model or code improvements, etc.) will cost a dollar amount, $\Delta \$C$. We can use a benefit/cost ratio:

$$BCR = (\Delta \$B - \Delta \$C) / \Delta \$C . \quad (6)$$

The BCR, computed here for product (or model) improvement, gives us a quantity to help answer the questions, “Was our product or model improvement process worth the cost?” or “Is the BCR higher to improve the *product* or to improve the *model*?” We link the 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 simplified concept of closure was introduced in the form of a simple equation (1) to integrate V&V, QRC, and QSV quantities with the economic function of present value factor (PV_F) in the time domain to represent the discounted value of the product or the assumptions in the out years. This equation enables quantification of benefit/cost trade-offs and timing decisions.

4.4 Raising QRC and the BCR

Following the V&V process, with ABCD leading to a quantitative validation statement, and then using a metric such as QRC as an input to benefit/cost and risk analyses, we can enter into the decision process with a quantitative analysis that is objective (at least on a relative basis) and is then balanced with expert judgment. However, what if the decision process, due to lack of validation data leading to a wide confidence bound (and a low QRC metric), or due to a high-consequence/high-risk application, leads us to conclude

that we have *not* achieved predictive adequacy?

If QRC is “too low” for predictive adequacy, we are implicitly saying there would be a large benefit, ΔB , from raising QRC (i.e., a tighter ABCD validation bound). Since we use $1 - QRC$ (or more accurately, $1 - QRC/1000$) as our *model + system* likelihood of failure, the increased benefit, ΔB , of raising QRC can be quantified, for example, in terms of product liability avoidance or increased value to the customer.

Once we know the ΔB we could achieve for a given ΔQRC , we can look for the increased cost, ΔC , needed to achieve this. We have to examine what we need to do to achieve the ΔQRC . This might include:

- More V&V analysis;
- More data (higher N for more confidence);
- Better physics (less tuning dials, K); and/or
- Bigger computing platform (reduced convergence uncertainty from “B” of ABCD).

As an example of this process, let us reconsider the model assessed QRC numbers of from the example of Figure 3 and equations (2)–(4). One of our future goals is to develop a more physics-based finite-element model (FEM) of these impact events and to predict threshold velocity. We will be able to use the same ABCD validation process and extend our validation to compute the QRC number as done in equations (2)–(4) and discussion. However, hopefully we would have fewer tuning dials. That way, even if our model fit to the data were the same as we achieved in Figure 5 (“A” of ABCD), the QRC numbers would increase due to a higher value of $(N - K)$. Will this model development be worthwhile? Of course, the FEM is more physics based, and it seems subjectively the right thing to do. But, we can offer a numerical assessment of how much we might gain in assessed QRC using such a model.

In Table 3, we list in the second column the QRC numbers from our best empirical model ($K=8$ dials). We then calculate new QRC numbers in the third column

of Table 3; assuming that our physics-based FEM of the impact could achieve the same fit of model to data but with, say, only ($K=4$) dials. Naturally, there is an increase in QRC for the model + system assessment, shown in the fourth column. Was the increase a good investment, considering just the four impact scenarios shown? If we assume a notional risk-consequence value of \$2 billion for an unplanned explosive reaction accident, we can calculate the benefit of our model investment as reduction in the risk due from the model + system assessed QRC increase as (in our simplified example) $\Delta B = \Delta QRC \times \2 billion . This value is shown in the fifth column of Table 3. Now, let us further use a notional value of cost, ΔC , of \$8 million per explosive type to cover the code development, verification, test data, and model validation to achieve the QRC increases in Table 3. We can then calculate the $BCR = (\Delta B - \Delta C) / \Delta C$ of developing such a model for each explosive type. These BCR values, one for each explosive type, are shown in the final column of Table 3.

If the numbers in this analysis were firm rather than notional, we would say that development of *both* new physics-based FEMs is a good thing to do, even with a cost of \$16 million, because the BCR ranges from $8.0 < BCR < 14.5$. We might start with the model for PBX 9404, with its higher BCR of 14.5.

Naturally, such a series of numbers as in Table 3 cannot *fully* replace expert judgment in the decision process. This should be more evident than ever after following all the assumptions and uncertainties in this example validation with calculation of risk reduction and BCR. We further stress that in our QRC effort to construct single measure combinations of reliability and confidence, we have only shown one such measure in this analysis. The “lower bound” QRC we discuss in this work is essentially a quantification of a reliability that includes the intertwined reliability of the *model and the system*, within the stated (and inferred) statistical confidence obtained from the V&V process. Outside

Table 3. Notional process of calculating the assessed risk reduction and BCR of the development of an improved impact model for each explosive type

Test Condition	Peak QRC at $K = 8$	Peak QRC, Same Fit but $K = 4$	ΔQRC with FEM	ΔB , Risk Reduction Benefit	BCR for FEM Development
LX-04 H = 30 mm	987	995	8	\$16M	8.0
LX-04 H = 6 mm	928	956	28	\$56M	
PBX 9404 H = 30 mm	912	944	32	\$64M	14.5
PBX 9404 H = 6 mm	891	921	30	\$60M	

the confidence interval, we quantify reliability with what amounts to a “coin flip” or $R = 50\%$. There are of course more complex ways to map both reliability and confidence onto a symmetric (or asymmetric) risk diagram. However, the analysis process, from V&V to QRC to risk to BCR, can provide a numerical defense of our decision, showing a process with quantitative diligence backed up by expert judgment for a balanced and defensible decision process.

4.5 Integral and Hierarchical Validation

To address the potential danger of insufficient predictive adequacy for a product application after completing V&V, QRC, and risk assessment we suggest a two-pronged approach to validation: a path of both *integral* validation and *hierarchical* validation.

Integral validation uses a system level model and system level data (with a high-level decision quantity, e.g., threshold impact velocity in the Figure 3 example discussed above) to perform the ABCD process and provide information for an adequacy statement. The integral validation process is relatively fast and provides a quantitative “first cut” to the bottom line. However, much of the information about the sources of the uncertainties in our validation is implicit and cannot be fully extracted. This is an example of why integral validation may not provide an adequate model for decision purposes. If it does not, we have only limited ability to improve the model because of its top level nature and reliance on top (system) level data. Therefore, it is usually good from a benefit/cost perspective to proceed in parallel with the longer, more expensive process of hierarchical validation.

Hierarchical validation attempts to build up to a system level response by performing, e.g., the ABCD validation process at the component and subsystem level. These responses are then combined by a system roll-up or Monte Carlo-like analysis leading to a hierarchical validation statement. The hierarchical bound may be as wide as or wider than the bound achieved with integral validation. However, with the hierarchical process, we can see which components and terms in the model and data have contributed to the total uncertainty. We can then go back and repeat portions or all of the ABCD validation process with more component data in the areas needed (a targeted experimental program) and obtain a tighter bound without having to do large numbers of system level tests.

The integral and hierarchical methods both have appropriate uses; integral gives a fast “bottom line” answer, but with limits on our course of action if that answer is not good enough. Hierarchical validation gives us a much more complete answer

and understanding, but is slower and more expensive to achieve. Comparing final system level validation statements from both methods is a good crosscheck on inputs to the decision process [13]. If the confidence bound from integral validation is different from the confidence bound from hierarchical validation, then either the model form (conceptual model or mathematical model) or assumptions about parametric uncertainty distributions should be re-examined.

4.6 Risk Management with QRC and BCR

Of late, there has been much interest in quantitative processes for model accreditation and system certification, a quest for “confidence” that is more than just “low,” “medium,” or “high.” 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. There are several methods developed at our partner national laboratories as described and referenced by Logan and Nitta [2]. These works have helped to motivate our own methodology.

As illustrated with the example of Table 3, the QRC number enters into the risk diagram of Figure 2. As the level of V&V goes up, our likelihood of risk (due to poor model V&V) goes down. Multiplying this V&V-QRC-based likelihood by the separately determined consequence in Figure 2 provides a direct (dollar benefit) measure of V&V. This is shown in Figure 2, where *adequate* V&V reduces the *likelihood* of failure, and *timely* V&V (i.e., concurrent with design versus post-mortem) can further reduce risk if we design our business and products to minimize the negative *consequence* of failure as well.

5. Summary and Conclusions

Our V&V methodology includes a process leading to quantitative V&V statements; that is, confidence bounded uncertainty on a quantity of interest over a specified regime of interest. The concept of a Level 0.0 to 1.0 scale is suggested as a *first step* toward quantitative V&V. Quantitative statistically-based V&V statements are the next evolutionary step.

The advantage of *quantified reliability (R) at confidence (C)*, QRC, is that it lets us do the following:

- Relate margin (M) and uncertainty (U) to R.
- Demand that we associate R with a C—quantifying C based on probabilistic, possibilistic, or evidence-based methods.
- Show how better *models and assessments* of M and U—and increasing the “effective number of coin

flips, or data points" N —quantitatively tightens U , allowing higher quantified C .

- Provide a numerical (albeit judgment-folded) estimate of “how much confidence do we need and how do we get it.”
- Link to a quantitative V&V statement to provide a lower statistical bound for reliability, risk, and benefit/cost analyses. This has the advantage of being *quantitative and assessable* but also the advantage of allowing us to 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.
- Express this situation as value (dollars) by linking QRC to risk likelihood and dollars to risk consequence, leading to benefit/cost ratio (BCR)—and allowing us to *protect* investments in computing, assessments, tests, etc. by *quantifying* their BCR. This method can provide a clear link between “science,” V&V, reliability, risk, “value,” and “investment strategy” —*it is our hope that V&V-to-QRC-to-Risk-to-BCR is one way to provide this link.*

The end product methodology and dollar benefit can be explained using a risk matrix with Risk = Likelihood × Consequence. *Risk* can also be quantified and viewed, as we illustrate, as the *risk* due to inadequate or mistimed V&V. The use of the BCR enables us to balance the benefits of qualitative and quantitative V&V and timing in a demonstrable way. Too little V&V is insufficient, while too much V&V is inefficient. The use of a quantified risk matrix and the BCR method lets us show how we can determine the level of V&V we feel is appropriate. The evolution from V&V to reliability at confidence to risk is suggested as a tangible way to justify the benefits of investment in V&V.

7. Nomenclature

Δ\$B	Benefit, usually in dollars
BCR	Benefit/cost ratio Δ
Δ\$C	Cost, usually in dollars
C	Confidence, a numerical value
E	Explosive content
FEM	Finite-element model
FOS	Factor of safety; for example, load capacity divided by load requirement
H	Head radius

K	Tuning dials
M	Margin, where <i>factor of safety</i> = $M + 1$ (M may be expressed with units or be dimensionless)
N	Number of trials as in coin-flipping
PDF	Probability distribution function
QRC	Quantitative reliability at confidence; unitless
QSV	Quantitative system value, integrated over time; units typically dollars
QSV ₀	Quantitative system value, per unit of time; units typically dollars per time t
PV _F	Present value factor, $0 < PV_F < 1$
R	Reliability
T	Temperature
t	Time
U	Uncertainty, general or “system” (in V&V always at a confidence, C)
V&V	Verification and validation
Z	Standard normal distribution variable for variable X, $Z = (X - \mu)/\sigma$
Z _{qrc}	QRC analog to Z, where function Z _{qrc} provides R for a chosen PDF, U, C, and M
β	Reliability index; $\beta = Z_{qrc}$
φ ⁻¹ (β)	Inverse of the standard normal cumulative distribution φ
μ	Population mean
σ	Population standard deviation (estimate)

8. References

- [1] Soudah, J., M. Pilch, S. W. Doebbling, and C. K. Nitta. *V&V – Capability in Stockpile Modeling and Simulation*. SAND-2004-08678. April 2004.
- [2] Logan, R. W., and C. K. Nitta. *Verification & Validation (V&V) Methodology and Quantitative Reliability at Confidence (QRC): Basis for an Investment Strategy*. LLNL UCRL-ID-150874. November 2002.
- [3] Logan, R.W., C. K. Nitta, and S. K. Chidester. *Design for Six Sigma with Critical-to-Quality Metrics for Research Investments*. SAE-2006-01-0995, in Special Publication SAE-SP-2032, April 2006.
- [4] Trucano, T. G., M. Pilch, and W. L. Oberkampf. *General Concepts for Experimental Validation of ASCI Code Applications*. SAND-2002-0341. March 2002.
- [5] Logan, R. W., and C. K. Nitta. “Verification & Validation: Goals, Methods, Levels and Metrics.” In *Proceedings of 2003 Summer Simulation MultiConference*, Society for Modeling and Simulation International, 2003.
- [6] Roache, P. J. *Verification and Validation in Computational Science and Engineering*. Albuquerque, NM: Hermosa Publishers, 1998.
- [7] Youngblood, S. “A VV&A Roadmap.” In *Proceedings of 2003 Summer Simulation MultiConference*, Society for Modeling and Simulation International, 2003.
- [8] Cafeo, J.A., and P. J. Roache. Private communication of draft V&V definitions to authors, April 2002.

- [9] Switzer, L. L., K. S. Vandersall, S. K. Chidester, D. W. Greenwood, and C. M. Tarver. "Threshold Studies of Heated HMX-Based Energetic Material Targets Using the Steven Impact Test." Paper presented at 13th APS Topical Conference on Shock Compression of Condensed Matter, Portland, OR, July 2003.
- [10] Pilch, M. Private communication with authors, October 28, 2002.
- [11] Oberkampf, W. O., and J. Helton. *Investigation of Evidence Theory in Engineering Applications*. AIAA-2002-1569, October 2002.
- [12] Logan, R. W., and C. K. Nitta. *Validation, Uncertainty, and Quantitative Reliability at Confidence (QRC)*. AIAA-2003-1337, January 2003.
- [13] Paez, T. L., T. D. Hinnerichs, T. K. Hasselman, and G. W. Wathugala. "Comparing Bottom-Up and Top-Down Approaches for Total Uncertainty Quantification." Paper presented at 7th U.S. National Congress for Computational Mechanics (7th USNCCM), Albuquerque, NM, 2003.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract W-7405-Eng-48. The authors wish to thank Lori Switzer, Kevin Vandersall, Daniel Greenwood, and Craig Tarver for their help in providing experimental results and theoretical assistance enabling the V&V and risk analyses.

Author Biographies

Roger W. Logan has led numerous program elements and technical efforts in weapon system assessment, certification, simulation, and experimentation. He is currently Senior Staff in Engineering Management. He has 25 years experience in code development, simulation, and more recently, in formal mathematical methods for V&V from a user standpoint and from a benefit/cost investment perspective. He has written extensively on the use of simulation for weapon system certification during the current era in the absence of nuclear testing. He holds B.S., M.S., and Ph.D. degrees from University of Michigan, University of California, and University of Michigan, respectively.

Cynthia K. Nitta has held numerous positions in the technical and leadership aspects of nuclear design and certification. She is currently Primary Certification Campaign Leader and has extensive experience ranging from code development to simulation to integration of V&V into the weapon assessment and certification process. She is a 2002 LLNL Edward Teller fellow and a design physics group leader in the Defense and Nuclear Technologies Directorate. She holds a B.S. from Princeton University, with M.S. and Sc.D. degrees from the Massachusetts Institute of Technology.

Steven K. Chidester has extensive experience in design and systems issues involving high consequence events, where either safety or performance or both are highly visible and crucial to success. He has pioneered the use of internationally known tests used for consistent and reliable evaluation of explosives safety, and has directed and performed extensive analysis and experimentation in this field. His work in explosives safety testing and simulation is widely recognized in the literature. He holds a B.S. from Oregon State University.