QMU and Nuclear Weapons Certification

What's under the hood?

David H. Sharp and Merri M. Wood-Schultz

In the article "Science-Based Stockpile Stewardship," Ray Juzaitis has described the main elements of the nuclear weapons program as it is taking shape today. The cornerstone of the program is nuclear weapons certification. Our purpose in this article is to explain our approach to certifying nuclear weapons in the posttest era.

Full-system nuclear tests and conservative designs have provided a high degree of confidence that stockpiled nuclear weapons will perform safely, reliably, and to specifications when their condition and use are within the tested envelope. Confidence was established through the certification procedure, whose outcome was a guarantee that the stockpiled weapon will achieve specific performance levels (military characteristics) under stipulated operating conditions (stockpile-to-target sequences, STS). Scientific judgment plays a critical role in determining the *sufficiency* of the criteria on which a certification is based and, to some extent, in determining whether the criteria have been met.

The methodology for certifying nuclear weapons has always included aboveground experiments, nuclear tests, and simulations of weapons operation. These elements were tightly interwoven, and no single element was sufficient by itself. Full-system tests were particularly important, however, in that such tests swept away many, although not all, uncertainties about the performance of nuclear weapons.

The need to answer questions about the stockpile has not gone away with the cessation of nuclear testing. For example, aging can alter the state of a weapon, and although not all observed aging defects are serious, some may be. Evaluating the effects of aging becomes increasingly important as weapons are kept in the stockpile well beyond their designed lifetimes. Similar questions arise concerning the effects of manufacturing or design flaws that may come to light, as well as the effects of planned refurbishments and modifications. Questions arising from the possible need for weapons of new design are looming.

Thus, there is a compelling need for assessments of how weapons will perform in an untested configuration. That is the problem. Plainly, there is no complete substitute for nuclear tests as a source of confidence in such assessments. Developing predictive capabilities that can support certification in the posttest era is therefore a tremendous challenge. Can this challenge be met with an improved scientific understanding of the behavior of nuclear weapons derived from a new generation of large- and small-scale nonnuclear experiments, better physics modeling, and more powerful computing? In this article, we will look at what needs to be done to answer this question, starting in the next section with a discussion of quantification of margins and uncertainties (QMU), a methodology created to facilitate analysis and communication of confidence in an assessment or certification.

Confidence is so central to certification that the use of predictive simulations in this context needs to be discussed first. Confidence in predictions of nuclear weapons performance, as with all scientific predictions, will be based on the track record, that is, on the scope and success of past predictions. But in matters concerning health, safety, or security, the cost of incorrect predictions can be very high, and one will often have just one chance to get the right answer. In such cases, the issue of confidence in prediction comes up with particular force, as compared with cases in which predictions are used mainly to guide the development of science. In both cases, one wants correct predictions; it is the consequences of incorrect predictions that are different.

An analogy can be drawn between nuclear weapons certification and Food and Drug Administration (FDA) approvals of new drugs. Just as the FDA requires demonstration of actual efficacy before approving a new drug, we require *positive evidence* that a nuclear weapon will work; absence of evidence that it will not work is not sufficient. Likewise, just as the FDA requires documentation of contraindi-

cations and side effects, leading to a lot of fine print in drug advertisements, our certification and validation studies come with some fine print. All this is not mere fussiness; in both cases, the driving force is the need for high confidence in predictions about the behavior of very complex systems.

The Role of QMU in Maintaining Stockpile Confidence

QMU, currently under development at Los Alamos and Livermore National Laboratories, is a framework that captures what we do and do not know about the performance of a nuclear weapon in a way that can be used to address risk and risk mitigation. The QMU framework is explained here in its simplest form, for example in a deterministic rather than a probabilistic form. Like any other part of science, QMU will evolve on the basis of experience gained through its use in actual applications.

The basic idea of QMU is to evaluate confidence in terms of the degree to which the operation of a weapon is judged to lie within "safe" bounds on judiciously chosen system or operating characteristics. A useful operating characteristic might pertain to the system configuration at a critical juncture in its operation, or it could relate to a time-dependent or time-integrated characteristic of the system.

QMU does not determine predictions or their uncertainties *per se*; these are inputs to a QMU analysis. It is designed to analyze and communicate the confidence in a conclusion or decision based on those predictions. Confidence—not the rigorous statistical determination of confidence intervals but the intuitive concept—is intrinsically hard to quantify. It is typically determined through a mental "calculation," weighing factors that may well differ from person to person. The person producing an assessment may not even realize all the factors that were considered. This degree of fuzziness can confound any attempt at analysis of confidence, its quantification, or its communication. It is clearly necessary to identify well-defined characteristics of a system on which discussions of confidence can be based.

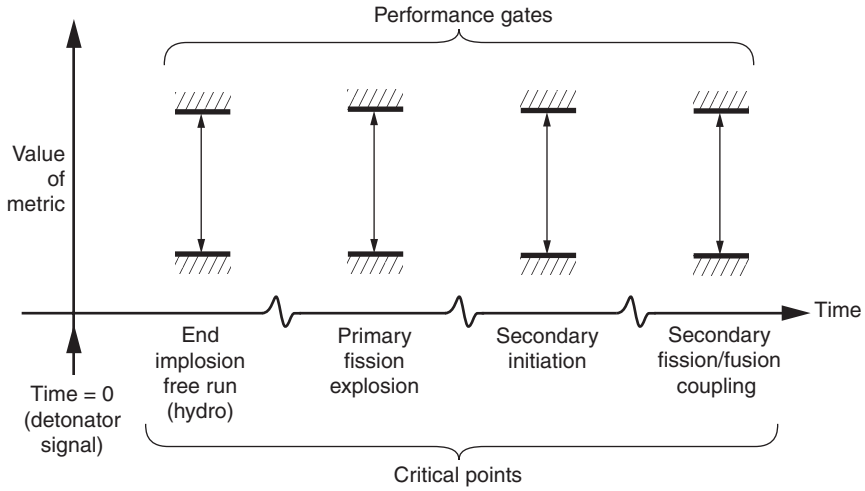

Figure 1. An Illustrative Timeline for the Operation of a Thermonuclear Weapon

"Characteristic" is a broad term. It can mean any function of the physical variables determining the performance of the system. Such characteristics can be static or dynamic, measured or calculated, intuitively clear or obscure. To give just one example, the amount of fissile material in a primary and its peak compression during operation are characteristics of a nuclear weapon.

In QMU, the characteristics of a system that are used to evaluate confidence are termed "metrics." Metrics and the other basic concepts necessary for QMU will be discussed with the help of Figures 1 and 2. Figure 1 shows a schematic timeline for the operation of a thermonuclear weapon, along with a schematic application of QMU. Four metrics are shown. The first one—the pit energy—is an operating characteristic of the system at the time when the kinetic energy of the imploding pit is at its maximum, so this metric is based on a snapshot of device behavior. In contrast, the system yield clearly is a characteristic that depends on the entire history of the device operation. What all metrics have in common is that they are high-level indicators of some aspect of the system's operation.

Defining useful metrics requires an understanding of the strengths and weaknesses of the tools used to evaluate the metrics. We define a complete set of metrics as one that, taken as a whole, is sensitive to *all* the important and potentially inadequate aspects of the simulations and measurements used in the evaluation. Data must be available to validate a useful gate for each of these metrics. Together, these requirements will affect the scope of weapons issues that can be addressed.

The process of evaluating a metric is conceptually straightforward for both measured and calculated metrics. Determining the uncertainty in the metric being evaluated is also relatively straightforward for a measured metric, but not for a calculated metric. In the latter case, each aspect of the calculation—databases, physics models, and numerical methods—may have errors. Error and uncertainty in predictive simulations are thorny problems, discussed in "Estimating Uncertainties" below.

The vertical bounds associated with each metric in Figure 1 represent the range of values for that metric that are judged to be acceptable. This range is termed a "gate" (see Figure 2). Metrics and gates are intended to delineate safe parameter regimes. Each metric used in QMU must be assigned an appropriate gate. It is clear that setting an appropriate gate is crucial to using QMU successfully.

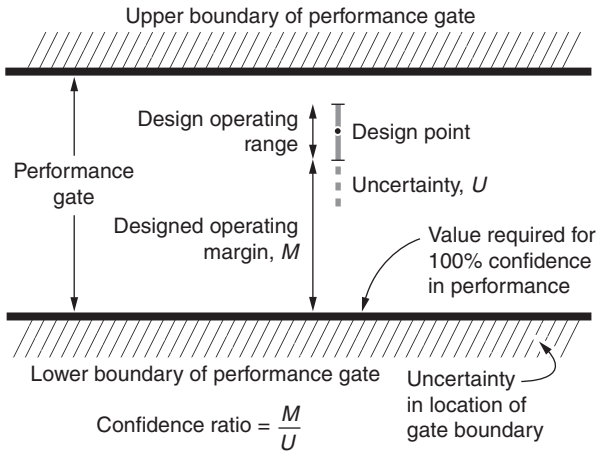

Figure 2. Key Elements of a Performance Gate: Boundaries, Margins, Operating Range, and Uncertainties

The procedure for setting a gate is to evaluate the metric for a set of successfully tested configurations. It is important that this set include variations in whatever parameter is being considered to ensure that the effects of variations in that parameter are represented in setting the boundaries of the gate. The set of metric values used to set the gate boundaries is thus known to correspond to successful performance, and the conservative presumption is made that systems having metrics outside this range—systems for which this metric does not fall within the gate—will not work. Nuclear test data are absolutely essential in defining valid gates. There is no substitute within the constraints of existing predictive capabilities. As much nuclear test data as possible are used to maximize confidence in the location of a gate. The criterion for successful device operation supplied by QMU is that system performance must lie within *all* the defined gates. Confidence that this criterion has been met derives from the "safety'" margin at each gate.

Margin is simply a measure of how much "room" is left between a metric at the limit of its operating range and the boundary of its gate. The details of the uncertainty in a metric, its gate, and the resulting margins are shown conceptually in Figure 2. We note that a range of metric values, called the designed operating range, results from the effects of different STS environments and manufacturing tolerances on device operation and from the effects of any intrinsic variability on device performance.

The confidence ratio (CR), defined at the bottom of Figure 2, summarizes the situation at each gate. The CR is simply the ratio of the margin in a metric to its uncertainty. The gate where the CR is smallest is the aspect of performance most likely to be or to become problematic—the weakest link. If we were very sure that we had not underestimated the uncertainty U, then 1.0 would be an acceptable CR. However, U is generally known imprecisely. An acceptable CR, therefore, will depend on scientific judgment as to the accuracy of U. The use of scientific judgment is common throughout science, and its role in weapons certification will be discussed in more detail below.

The CR is an example of a figure of merit, and other figures of merit could easily be defined. Confidence in overall device operation is represented by the CRs for the entire set of gates, and any CR approaching 1—a weak link in the chain—is a warning flag.

Three QMU functions—defining metrics, setting gates, and evaluating uncertainty—have been discussed, and acceptable confidence was quantified in terms of the CR. We stress that this confidence derives not from the QMU formalism itself but from the quality of the science used in applying it.

Even if QMU gives an acceptable CR, a fundamental question remains: Is this procedure *sufficient* to guarantee acceptable performance? This issue was touched upon above as the requirement that *all* the important vulnerabilities be adequately constrained by the QMU metrics. Whether or not this has been done can only be based on expert, or scientific, judgment. Scientific judgment is always essential in reaching a decision on the basis of incomplete or inconclusive evidence and therefore has always played a significant role in certifying nuclear weapons. The foundation of science is that experiment is the sole judge of truth, but the use of expert judgment is legitimate when it is *provisional*, in the sense that it is subject to challenge and correction through the scientific process itself. In a posttest era, we must continue to rely on expert judgment. Expert judgment can still be challenged on the basis of nonnuclear tests, predictive simulations, and peer review, although the standards are softer than the ones set by full-system tests.

QMU is designed to facilitate such challenges to expert judgment. It is also flexible enough to incorporate all the criteria on which a certification might be based, and that is why it can be used as the methodology for certification. It is important to keep in mind that QMU, like other tools, does not determine the adequacy—or in this case, the sufficiency—of the product it is used to create. That aspect still depends on the craftsman (the designer) and the raw materials (the data).

Determining the Behavior of a Complex System

There are basically two ways in which to learn about the integrated behavior of complex, real-world systems: One is full-system testing (observation), and the other is full-system simulation. Full-system tests and simulations are complementary but are not interchangeable. Full-system tests always provide more confidence than simulations because they give a definitive answer as to *whether* a particular device worked under the specific conditions realized in the experiment. Also, they typically provide some detailed data on the internal conditions during the experiment. However, full-system tests are not equally definitive in telling *how* the device works, and it can be prohibitively expensive to explore device behavior over a broad range of operating conditions in this way. In contrast, full-system simulations are usually cheap (compared with hardware), have zero risk, are controllable, and allow access to the details of the physical processes. The price of these advantages is the need for complete and precise knowledge of the operation of the system. Because such knowledge is often not available for complex systems, simulations come with myriad opportunities for errors.

Increasing the scope of stockpile-related questions that can be answered with confidence and without nuclear testing requires that the boundary between what can be reliably established by full-system simulation and what must be proved by a full-system test be shifted. Correct and reliable prediction using a simulation presents two core issues: One concerns data, and the other concerns the integration of information pertaining to subsystems into full-system simulations.

Experimental Data. Data are needed to define initial conditions and parameter values for specific problems and to validate or constrain models. For complex problems, a lot of detailed data are needed to validate models for predictive purposes, and the data requirements go well beyond what is needed for interpolation. Even in a laboratory setting, detailed quantitative data about fluid motions, for example, are often hard to come by. The problem of getting well-diagnosed, accurate data is very much more difficult for nuclear weapons because they operate in a regime that is far from laboratory conditions.

For complex systems, data are usually sparse, relative to the need, so it is often necessary to combine data from multiple, diverse sources when testing a model. Some of the data requirements are being met through integral experiments at facilities such as the Dual-Axis Radiographic Hydrodynamic Test (DARHT) Facility at Los Alamos, the Z-machine at Sandia National Laboratories, the Omega Laser at the University of Rochester, and eventually the National Ignition Facility (NIF) at Lawrence Livermore National Laboratory and through subcritical underground experiments conducted at the Nevada Test Site (NTS). Laboratory-scale experiments can play a vital role in building predictive models, although their usefulness is occasionally underestimated. Sometimes, model parameters can be determined from a more basic theory. For example, the viscosity coefficient appearing in the Navier-Stokes equation could be calculated from kinetic theory. However, archival nuclear test data remains the crucial core of the data used for certifying weapons because it provides the only faithful integration of the interactions among the various parts of a functioning device.

From Subsystems to Full Systems. The task of building an understanding of full-system behavior from a knowledge of component subsystems is one of the most difficult aspects of modeling complex systems. Multiscale science, that is, consistent representations of physical processes that extend over more than one length (or time) scale, is often a problem. A completely adequate "microscopic" model, which can include properties that profoundly influence large-scale behavior, is often not feasible for use in full-scale studies, so a method is needed to incorporate the essential fine-scale information into macroscopic simulations. Areas of weapons physics where this issue arises include modeling the initiation of high explosives, materials damage modeling, and the fluid-mixing problem.

The next part of the integration problem is to model full-system performance of a complex device starting from models of the individual components or processes. We refer again to Figure 1, this time to illustrate how one might decompose a complex system into pieces that can be studied independently or, at least, conditionally. The initiating event in a nuclear weapon (seen at the far left of the timeline in the figure) is the detonation of a high explosive (HE). The physics and chemistry of detonations are extremely difficult subjects, which have been studied at Los Alamos since World War II. Fortunately, the HE detonation is unaffected by the physics occurring in the nuclear regime of device operation, so HE can be studied and modeled using information from aboveground (nonnuclear) experiments. Doing so allows an HE detonation model to be developed and tested independently of the downstream physics.

The next stop in Figure 1 is the pit implosion, during which the flow of dense materials occurs. Because the material flow is driven by the HE, the pit implosion is conditional on the HE simulation, yet it is, at this point, still independent of nuclear-phase processes. Like the HE model development, laboratory experiments and large-scale

experiments (for example, those at DARHT) can supply extremely useful guidance in modeling the flow of dense materials.

Two observations are appropriate at this point. The first (and rather obvious) observation is that an accurate full-system simulation must be built from accurate, reliable models of the individual processes that are occurring: detonation, material flow, neutronics, and so on. There is no reason to think that coupling together poorly modeled processes or subsystems will produce anything but a poor model of the whole system. The second observation is that accurate modeling of the coupling of different physical processes, for example, of neutron and radiation transport to material flow, can itself be very difficult to achieve.

The timeline in Figure 1 continues to the time when criticality is achieved. A calculation must now couple material flow to nuclear and thermonuclear processes, for which the data is not as detailed or systematic as that available for the earlier processes. We are left without a guarantee that predictive models can be validated for this late-stage operation, but steps can be taken to improve our understanding.

The first step has already been stated: Begin with a good model. A good model, say for fluid mixing, will be internally consistent and will agree with a broad range of results from large- and small-scale nonnuclear experiments, with few if any adjustable parameters. The validation experiments must include all the important aspects of the weapons process—for example, change of flow from laminar to turbulent—and at a variety of parameter values demonstrate predictive capability.

A model that works well in the laboratory regime is not necessarily correct in the weapons regime. But one can still test the model postdictively in the weapons regime, using comparisons with a portion of the NTS database to constrain any free parameters and the results of applying the model, with no additional parameter adjustment, to the remaining NTS data as evidence of the model's predictive power. If sufficient data exist, this procedure will provide a fairly good means for establishing confidence in models for the explosion phase of operation. However, if a bottom-line result reflects a sensitive dependence on initial conditions or other problem parameters, then its reliability may be subject to question.

How far will all this take us toward meeting the goal of a predictive capability for assessment and certification? This will certainly depend on the question one is trying to answer; we will be able to deal with some questions using predictive science but not with others. The boundary will be set by the scope and power of the predictive models that we are able to develop—an explanation that requires an explanation.

Scope refers to the number and variety of cases in which the theory has been tested. Knowing the scope is important in building confidence that one has identified the factors that limit the applicability of the theory. Power is judged by comparing what is put into the model with what comes out. Theories that correctly predict a wide range of phenomena with just a few input parameters are powerful; phenomenological models—those that are calibrated to data and hence closely tied to specific problems in their formulation and predictions—are less powerful. Nevertheless, they are extremely useful, and are in fact the default solution to the problem of producing assessments when adequate fundamental models are not available. Monitoring progress toward predictive capability is the job of validation and uncertainty quantification, the topics of the next section.

Estimating Uncertainties

Nuclear weapons performance is calculated using complex computer programs, or codes. These codes combine databases for various physical quantities (equations of state, opacities, and so on), multiple physics models, and algorithms for solving the physics equations to calculate the operation of the weapon, given its initial state. Like all codes, weapons codes are approximate representations of reality. As we call on them for actual predictions, as opposed to interpolations or small extrapolations, to help answer questions about weapons that deviate from their tested condition, knowing how accurately the codes describe the real world, that is, knowing the error in code predictions, becomes of paramount importance.

Validating a code is not like proving a mathematical theorem. Nuclear weapons simulation codes must simulate coupled, nonlinear, multiscale physical processes, and the most important and difficult-to-model aspects of weapons behavior (which occur during the explosive nuclear-energy production phase) are not accessible to laboratory experiments. This leads to reliance on integral data from nuclear tests and to the additional complication of having only indirect inferences about weapons behavior from this data.

Nevertheless, determining uncertainties in simulation-based predictions revolves around the answers to a few basic questions: What do you need to predict? What factors can lead to errors in the predictions? How can you get a handle on these errors?

Errors in predictions can come from poor-quality input data, incomplete or insufficiently accurate physics models, and inaccurate solutions of the governing equations. Some of these, for example, equation-of-state errors, error models for material damage or fluid mixing, instrumental

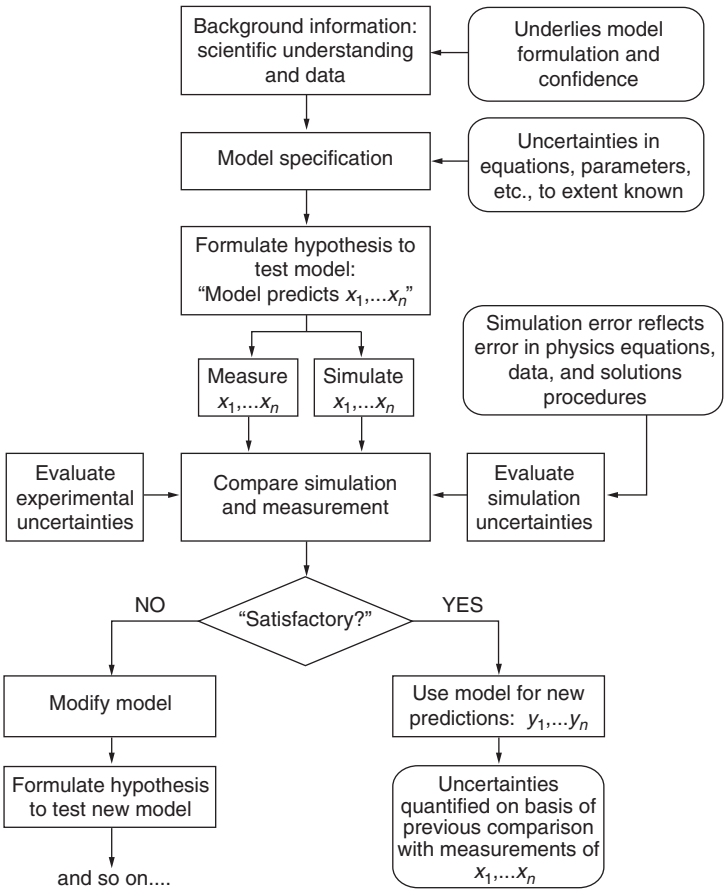

Figure 3. The Scientific Method—the Basis for Quantifying Uncertainty in Predictions

errors (as in DARHT or NIF), and errors in numerical solutions can be determined through standard experimental methods with sufficient resources. Solution errors are a distinctive feature of predictions made using large-scale simulations. They contribute to the total error in a prediction and must also be considered when drawing conclusions from comparisons of data with predictions. As discussed in "The Role of QMU in Maintaining Stockpile Confidence," the explosion-phase physics is problematic. In addition to the calculations being exceptionally difficult and the physical regime being inaccessible in the laboratory, there are significant uncertainties about some aspects of the physics.

Determining the error in a simulation is directly analogous to determining the error in an experiment. A direct determination of error is sometimes possible for simple experiments and for simple simulations by comparing a measurement or prediction with a standard or an analytical solution. In contrast, errors in complex experiments and simulations must be calculated by breaking down the end-to-end operation into components (subsystems) amenable to separate error analysis. The subsystem results must then be painstakingly combined to produce the overall uncertainty in the specific quantity of interest. Although this procedure is obligatory in complex situations, it has the virtue of showing which sources of error are most influential and of providing guidance for reducing the errors one by one. Incremental progress will be made at mitigating the effects of errors, but significant uncertainties in predictions based on simulations will remain for the foreseeable future.

The analysis of uncertainties has many aspects, but they can be combined into a simple, coherent framework as shown schematically in Figure 3. This figure simply displays the main steps in the scientific method but in a probabilistic setting in order to include uncertainties. It shows a "forward step," which goes from a hypothesized model to predictions that are compared with experiment, a "backward step" that consists of model improvement as a result of the comparison with data, and then new predictions. We note that Figure 3 has an alternative, and equivalent, interpretation in terms of the steps in Bayesian statistical inference.

The approaches to error analysis and uncertainty quantification discussed above pertain to prediction of events within or very near the regime of the data set used for validating the model. Outside this regime, uncertainties cannot be assessed, and predictions may be wrong. "Known unknowns"—that is, recognized phenomena for which adequate models are lacking—are a common source of error in simulations. Then there are "unknown unknowns." By definition, they cannot be dealt with directly, but an attempt is made to address them through "What if?" exercises and, most important, through conservative design. In the QMU framework, that means ample margin. Certainty is still not in the cards, so our highest priority is to avoid catastrophic systemic failures, rather than failures resulting from isolated low-probability events.

The Future of Certification

The demands on certification procedures derive from our responsibility to identify and remediate factors, including obsolescence, that could place the nation's nuclear deterrent at risk. Managing these risks will present a broad range of challenges to our certification and assessment capabilities. Some of these challenges can be dealt with confidently, by using improved predictive capabilities, but others will stress these capabilities to the point at which further nuclear testing may be needed to maintain confidence. Questions in the latter category may include certification of weapons of new design and assessment of severe weapons defects because they deal with weapons behavior well outside the tested range. How a particular question is dealt with is a matter of judgment, and QMU should be helpful in explaining the basis for confidence in such judgment.

Risk mitigation for the nuclear stockpile problems will be accomplished through (1) surveillance to monitor the actual condition of the weapons, (2) predictive assessment of the impact of changes observed in the surveillance program, especially the identification of possible failure modes (flagged in QMU by loss of margin at a gate), and (3) the ability to refurbish, remanufacture, or modify a weapon system to remedy defects (diagnosis is hollow unless followed by treatment). The last option requires a functional manufacturing capability, which is an extremely complex and expensive undertaking. Other options restrict the weapons' potential use, such as changes in the STS.

One sometimes hears that the problem of recertification can be obviated simply by "making them [warheads] the way we used to." This is appealing, but it cannot address design flaws, new designs, *or* the simple fact that, for all practical purposes, we cannot make them the way we used to. Thus, the need for a more predictive scientific understanding of weapons operation cannot be sidestepped so easily.

Nor would the need for better predictive capabilities be completely obviated by a return to testing. The stockpile questions that need to be answered would inevitably outstrip the number of tests authorized or conducted to answer them, as it has occurred in the past. It is a fact of life that larger political considerations affect, and sometimes override, technical needs. Moreover, human and institutional factors will continue to profoundly influence the stockpile stewardship program. An example of such a factor is the need for an integrating goal that can focus both questions and efforts within the weapons program. Ironically, one of the critical roles of nuclear testing was to provide exactly this focus. The challenges posed by weapon assessment and certification will be met through a combination of the currently recognized steps of science, each held to higher standards of control and error analysis than is customary, and through integration of the various parts of the weapons program so that they effectively support the development of comprehensive predictive capabilities. Successful stockpile stewardship will produce tight estimates for the outcomes of critical events and will identify corrective actions where necessary. Failure, in terms of inadequacy, will be recognized as estimates that are too loose—that is, too uncertain or too unreliable—to be useful.

Predictive science applies to phenomena resulting from understood or acknowledged causes. In time, an increasing number of such causes can be studied and brought within the predictive framework with a corresponding increase in confidence in our ability to identify the factors that limit the use of our models to assess the behavior of untested weapons. True failure could still occur in those cases in which the unrecognized cause and unanticipated effect are significant. The fundamental question of the sufficiency of our certification procedure—"How will we know if we have made a mistake?"—will never go away. Nevertheless, we believe that, with effort and determination, the nuclear weapons community can go a long way toward meeting the challenge of certification as it is presented today. ■

For further information, contact David Sharp (505) 667-5266 (dcso@lanl.gov).