VALIDATING SAFEGUARDS EFFECTIVENESS GIVEN INHERENTLY LIMITED TEST DATA

ALAN SICHERMAN

Lawrence Livermore National Laboratory*

Livermore, CA

UCRL--96045

DE87 013224

Abstract

A key issue in designing and evaluating nuclear safeguards systems is how to validate safeguards effectiveness against a spectrum of potential threats. Safeguards effectiveness is measured by a performance indicator such as the probability of defeating an adversary attempting a malevolent act. Effectiveness validation means a testing program that provides sufficient evidence that the performance indicator is at an acceptable level. Traditional statistical techniques are useful in designing a testing program when numerous independent system trials are possible. However, within the safeguards environment, many situations arise for which traditional statistical approaches may be neither feasible nor appropriate. Such situations can occur, for example, when there are obvious constraints on the number of possible tests due to operational impacts and testing costs. Furthermore, these tests are usually simulations (e.g., staged force-on-force exercises) rather than actual tests, and the system is often modified after each test. Under such circumstances, it is difficult to make and justify inferences about system performance by using traditional statistical techniques.

In this paper, we discuss several alternative quantitative techniques for validating system effectiveness. The techniques include: (1) minimizing the number of required tests using sequential testing; (2) combining data from models inspections and exercises using Bayesian statistics to improve inferences about system performance; and (3) using reliability growth and scenario modeling to help specify which safeguards elements and scenarios to test.

Introduction

The Safeguards Validation Problem

In designing nuclear safeguards systems, a key issue is how to validate safeguards effectiveness against a spectrum of potential threats specified by regulatory guidance. Potential threats include malevolent acts such as theft or diversion of special nuclear materials or sabotage by a variety of adversaries including terrorists and insider employees. Safeguards effectiveness is often expressed by a performance indicator such as the probability of defeating an adversary attempting a malevolent act. Effectiveness validation means a testing program providing sufficient evidence that the performance indicator is at an acceptable level. Analytical models and inspections help in identifying potential vulnerabilities. However, confidence in safeguards system performance comes from demonstrating safeguards effectiveness by subjecting the system to a sequence of tests or trials.

Staged force-on-force exercises are often used to test certain aspects of physical security systems. Tests of material control and accountability systems may be performed to check for adherence to procedures and to determine the system's capability in detecting discrepancies introduced by the testers. Typically, each system trial is based on a particular "scenario" that includes target location, a specific threat, operating conditions, and time of day. Obviously, there are many possible scenarios. To quote an anonymous source, "It is conceivable we may have to run a force-on-force exercise once a day in order to validate safeguards against all the possible combinations." Unfortunately, these tests are costly and have significant operational impacts on facilities. Therefore, judgment of safeguards adequacy is often based on a few tests or exercises. In addition, the safeguards system is modified after each test which adds further complications to these judgments.

Traditional Statistical Approaches

Statistical approaches to validation depend on the meaning we assign to "acceptable validation." Conservatively, "acceptable" implies a high confidence that the system performs at or exceeds a desired level. In order to achieve such confidence using traditional statistical approaches, it is usually necessary to conduct a

*Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

This document is
PUBLICLY RELEASABLE
Barry Steele
Authorizing Official
Date: 2-15-07

DISCLAIMER

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

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

large number of tests. Often, a system which performs even better than the desired level is required.

A more pragmatic approach is to define acceptable as "within a reasonable range of the desired (good) level, rather than an undesired (bad) level." Alternatively stated, test results are explained in terms that imply the system is "good" rather than "bad." More specifically, classical statistics focuses on two types of errors we might make by accepting or rejecting a system based on a set of tests. One kind of error (called Type I) is rejecting a good system and the other called (Type II) is accepting a bad system. A traditional statistical approach is to devise a testing plan with a fixed number of tests such that the probability of making each type of error is less than a prescribed level. For example, we might decide to limit the probability of each type of error to be no greater than 0.1. In this case, we formulate a testing plan such that if hypothetically performed for a huge number of good systems, we would reject no more than 10% of them and for a huge number of bad systems we would accept no more than 10%. When such a plan is actually implemented for a particular safeguards system, the results are characterized by the phrase "90% confidence" in the decision. Using traditional statistical approaches, numerous tests are required of the same unmodified system configuration for each scenario in order to have sufficient confidence in total system performance.

Need for Alternative Validation Techniques

Safeguards system validation concerns arise when:

- tests have high monetary, political, or operational impacts;
- tests are feasible for only a few limited elements of the system under less than ideal circumstances (e.g., simulated exercises);
- almost always, the system is modified in response to test results or with time; and
- there are critical consequences if inferences about system performance are incorrect.

Under these circumstances, the use of traditional statistical techniques to make inferences about system effectiveness may be neither feasible (too many tests required) nor appropriate (system is modified after each test).

The difficulties discussed above require alternative approaches and quantitative techniques that address these interrelated questions:

- How can we keep the number of tests reasonably small and still achieve meaningful validation?
- What scenarios make the best use of testing resources?

With these issues in mind, we investigated several promising alternative techniques and approaches. The techniques include: (1) minimizing the number of required tests using sequential testing; (2) combining data from models inspections and exercises using Bayesian statistics to improve inferences about system performance; and (3) using reliability growth and scenario modeling to help specify which safeguards elements and scenarios to test.

These approaches incorporate available knowledge to maximize learning from each test result. Each approach emphasizes a different aspect of how we might make better use of each test result as it occurs. In the remainder of this paper, we will briefly discuss the techniques of sequential testing, Bayesian statistics, and scenario and reliability growth modeling. Simplified examples will be used to illustrate these techniques.

Alternative Validation Approaches

Sequential Testing

For validating system performance we can consider sequential testing [1] plans as opposed to plans with fixed numbers of tests. For fixed testing (or sampling, as it is sometimes called), it is necessary to specify ahead of time the number of tests to be performed. For sequential sampling, a "stopping rule" can be applied after each test result is examined. Under certain assumptions, sequential testing can reduce by a factor of 2 the number of tests required to validate (i.e., accept or reject) a system.

Figure 1 illustrates a sequential testing plan for validating safeguards effectiveness. In this example, we define the desired probability of defeating an adversary to be 0.8 or better, the undesired performance level to be 0.5 or worse and a 90% confidence that we won't accept a bad system or reject a good system. For each scenario, "exercises" are conducted which are assumed to be independent and realistic tests of safeguards performance.

In sequential sampling, after each exercise, we follow the instructions of the region in Fig. 1 (FIX, CONTINUE testing or ACCEPT) in which we find ourselves based on the number of exercises and failures at that point. For example, if we had three failures in a row, we would stop and declare that the system must be fixed. If we had five

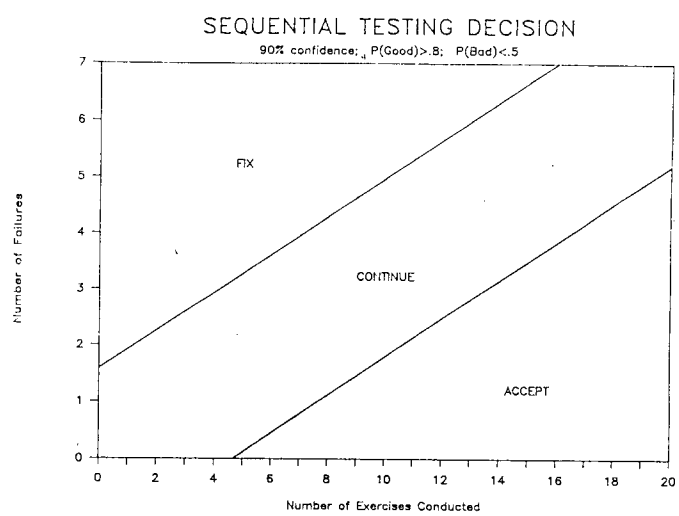

Fig. 1. Illustrative sequential testing plan.

successes in a row, we would stop and declare that the system is acceptable. But if we failed 2 of the first 6 tests we would continue testing the system. For this example, the expected number of tests required to validate a good system is about 10 [1]. With traditional fixed sampling, it would be necessary to plan for 16 tests and we would require no fewer than 11 successes before accepting the system with 90% confidence [1]. Thus, for a given scenario, the sequential plan would save on the average about 6 tests. If 3 scenarios were tested (such as day shift, night shift and weekends), the expected savings would multiply to about 18 tests. Sequential sampling plans usually have conventions for cutting short an unlucky "run" in the CONTINUE testing region (see Fig. 1) while still achieving the desired confidence.

To recap, the inputs for sequential testing (as with traditional testing) include the desired confidence in avoiding erroneous action, and the specification of performance levels for acceptable (good) and unacceptable (bad) systems. Higher confidence requirements and less differential between acceptable and unacceptable systems imply a larger number of tests to achieve validation. However, practical definition of acceptability and the use of a stopping rule can help limit the number of tests required and still result in a reasonable validation program.

Bayesian Statistics

As discussed in the introduction, tests such as force-on-force exercises tend to be limited in number due to their impact on facilities. Thus, while system exercises are clearly relevant, it should be borne in mind that, by themselves, they can be insufficient for adequate validation.

One solution is to supplement specific system exercises with results from other sources: e.g., internally conducted drills, safeguards inspections, exercises conducted at other facilities of a similar nature, expert judgments, and safeguards models. Ideally, we want an approach to validation that can combine such supplemental data with exercise results. The Bayesian statistics approach [2] provides a quantitative framework for using other relevant information in addition to the exercises themselves.

An important concept in implementing a Bayesian approach is to formalize the "relevance" of supplemental information. This is done by characterizing the supplemental data as being "equivalent" to having observed a number of system exercises and their outcomes. When the supplemental information is considered to be sparse or not especially relevant in distinguishing between relatively good and bad systems, the equivalent number of tests representing this information is small. When supplemental data is considered extensive and relevant, the equivalent number of tests would be relatively large. The analyst or decision maker controls the effect of supplemental data on the analysis by making judgments regarding the applicability of auxiliary information. There are a variety of techniques for formally assessing or "encoding" such probability judgments [2]. The Bayesian approach is very flexible and it is important to be aware of the implications of such judgments on the analysis. On the other hand, Bayesian techniques provide mechanisms for using supplemental information when it is available. When the number of possible system tests is very limited, using supplemental information may still permit reasonable validation.

Another important concept of the Bayesian approach is the probability of being good versus being bad given available information. This is in contrast to strictly classical statistics where the system must be good with probability 0 or 1 and all that we infer is the "confidence" of not making a Type I or Type II error. Thus, a Bayesian approach begins by specifying a probability of being good versus being bad before any testing is performed. The test data "updates" this "prior" probability. The prior probability, however, can include consideration of supplemental information.

Appropriate criteria for acceptance or rejection in a Bayesian approach also depend on judgments of the decision maker. However, in a manner somewhat corresponding to classical statistics, we can specify a desire to accept a system when there is no more than a 10% chance of it being bad and to reject a system when there is no more than a 10% chance of it being good. By using such criteria, Bayesian statistics and sequential sampling schemes can give de facto the same results even though they have different philosophical interpretations. Specifically, if we assign a prior 50% chance to the hypothesis that the system has 0.8 probability of defeating an adversary, and a 50% chance to the hypothesis that the system has a 0.5 probability of defeating an adversary, then Fig. 1 can represent a Bayesian sequential sampling plan.

It is at this point, however, that supplemental information can be introduced. Suppose, for example, that the system being tested passed a "third party" or auxiliary exercise before a formal internal validation testing program was begun. Then in Fig. 1, instead of starting at the origin, we could act as if we had already passed one test and start at the point (1,0). In this case, only 4 more consecutive successes rather than 5 would be required to accept the system. Similarly, if the system failed the previous auxiliary exercise, (and no upgrades were done) 2 more consecutive failures rather than 3 would be enough to reject the system.

As an alternative example, suppose we have a safeguards model (or an expert) claiming that the facility safeguards "effectiveness" is 0.8. If the model were judged as not very relevant, we might not change our initial prior probability (e.g., 50-50 as illustrated above) much. On the other hand, if we thought the model was exact and would only give a 0.8 result if the system were good, and would never give a 0.8 result if the system were bad, we would hardly need to test further. More generally, models or expert evaluations have some usefulness but do not negate the need for testing. If the model indicates that the safeguards system is more likely to be good rather than bad we would like to take advantage of this in our testing. Suppose we judge (e.g., encoded via assessments) that the model is 4 times more likely to give a result of 0.8 if the system is

good rather than bad. In Fig. 1, this is equivalent to starting at the point (3,0) rather than the origin. In this case, 2 consecutive successes would suffice to accept the system rather than the original 5. Conversely, 3 successive failures would not lead to rejection if we believed the model to the extent specified.

As a final point, we note that the simple example in Fig. 1 is actually somewhat conservative. The plan in Fig. 1 is designed to satisfy the error criteria when a good system is just barely good and also when a bad system is just barely bad. It may be more realistic in a Bayesian analysis to characterize the prior probability about safeguards effectiveness as not exclusively 0.8 or 0.5 but as some continuous distribution between 0 and 1. The use of continuous prior distributions can result in fewer tests required to satisfy validation criteria similar to those described above.

To summarize, Bayesian sequential testing can reduce still further the number of tests required for validation. The decision maker or analyst can control the "conservatism" of the analysis by how much significance is given to supplemental data. If it seems reasonable to give even modest credit to supplemental data then fewer tests are required for given validation criteria or more stringent validation criteria can be met with the same testing resources.

Scenario Modeling and Reliability Growth Modeling

In constructing a reasonable validation program, it is important to decide what scenarios and system elements should be tested. Common-sense notions would lead us to test scenarios which are:

- concerned with key vulnerabilities that could result in significant consequences to the facility and the public at large,
- relatively more likely to occur, and
- more likely to help identify correctable vulnerabilities

We may also want to characterize aspects that are common to many scenarios and aspects of scenarios that can perhaps be tested by less expensive or disruptive methods. Formal quantitative approaches involve modeling the relationship among scenarios (their variations and what knowledge of system performance for one scenario tells us about performance for another scenario) and identifying vulnerabilities and corrective actions.

Test data obtained regarding one scenario can be used as supplemental information regarding another scenario. For example, suppose we had 5 consecutive successes of a system for a night shift scenario against terrorists. This information might tell us something about system effectiveness for the day shift scenario. At one extreme, if it were judged always easier to defeat an adversary during the day versus the night, we might not need to test at all for the day shift after validating the system for the night shift. Without going to such an extreme, it could still be reasonable to treat the night shift successes as auxiliary information in a Bayesian analysis and require fewer tests to validate the day shift.

Another way of modeling relationships among scenarios is to break up a test of the overall system into tests of system components or stages. To illustrate, consider a theft scenario by a nonviolent insider who must acquire and remove a target from a material access area (MAA) and then exit the facility out the protected area (PA). It may be that these two stages can be tested independently vis-a-vis system effectiveness; that is, the system ability to prevent removal from the PA does not depend on how the material was acquired or removed from the MAA. Suppose further that safeguards for removal from the PA are essentially the same for all targets but safeguards for removal from the MAA could differ. If we obtained test data regarding removal from the PA for one target, we could formally use the same data for evaluating another target scenario. A modeling technique used to accomplish this is called series modeling. Series modeling is practical using a Bayesian approach [3] and could result in savings of testing resources by not repeating testing of one stage that has been analyzed sufficiently for other scenarios.

The concept of component modeling is also applicable when only limited elements of the system are amenable to "realistic" tests. Testing the limited elements results in more information about their effectiveness. This information is then combined with information (e.g., from models or experts) for the other system elements using Bayesian techniques analogous to those above.

A technique dealing with the process of identifying and correcting vulnerabilities is called reliability growth modeling [4]. A prime objective of testing is not merely to accept or reject a system but to improve the system by finding and correcting vulnerabilities. A testing program having good diagnostic and corrective action features should be encouraged over a program lacking such characteristics. To illustrate using Fig. 1, if we fail the first 2 tests, we must pass the next 9 in a row to accept the system and do the fewest tests. But what if we upgraded the system? Do we still need the same number of additional tests to validate? One procedural alternative is to begin validation over again when certain system modifications are made. This procedure, however, does not distinguish sufficiently between a testing program with good diagnostics and upgrade solutions and one without. For the latter, validation testing is simply repeated whether or not vulnerabilities have been systematically analyzed or corrected. Reliability growth models can give credit to the diagnostic/corrective capabilities of a testing program. In reliability growth models, judgments are required by decision makers as to diagnostic effectiveness. These judgments relate to the probability of finding "correctable" flaws, and also to the effectiveness of upgrades in improving performance. Analytically, initial failures due to correctable flaws are given reduced emphasis in making inferences about performance. Even modest credit ("where due") to such capability can result in fewer tests required for validation and in the use of all test data related to the validation program. (The alternative of throwing out data

and starting anew without a formal way of analyzing how one has learned from failures can impact the defensibility of a validation program.) As with other techniques discussed in this paper, the decision-maker or analyst can control the credit given to reliability growth.

Both scenario modeling and reliability growth modeling can give guidance as to which scenarios or elements of the system to test. We have seen that scenario modeling can help set priorities on scenarios to help reduce the number of required tests (e.g., night shift before day shift in the example above). Similarly, reliability growth modeling gives credit for starting testing with those scenarios that will provide the greatest diagnostic benefit. These techniques promote two objectives of testing which are to improve the system and show it is acceptable. Because "difficult" scenarios or initial failures are often perceived to detract greatly from the latter objective, techniques which permit planners to look aggressively for improvements and yet promote validation by requiring ultimately fewer tests can be very helpful.

Conclusion

In summary, it is appropriate for safeguards planners to consider: a) how to specify pragmatically an acceptable versus an unacceptable system, b) how to interpret the status of validation as each test result occurs, c) how to use supplemental information to assist during validation, and d) how diagnostic/corrective actions improve the system as testing continues. We have investigated techniques that would allow planners to apply such considerations in a formal quantitative manner in coping with the problem of validation when testing is limited.

The techniques make use of the concepts of sequential testing, Bayesian statistics, and scenario and reliability growth modeling. In this paper simple examples of safeguards validation were presented to illustrate how these techniques can be used. Currently, we are developing algorithms for helping planners use these techniques in designing validation programs. An important challenge is to provide aids for the decision-makers in deciding and controlling the appropriate degree of "conservatism" in giving credit to auxiliary information and reliability growth. The flexibility of the techniques allow for a variety of implementation strategies. Conceptually, each technique can be used individually (e.g., auxiliary models without use of reliability growth models) or they can be used collectively. For planners who must validate safeguards effectiveness given inherently limited data, we believe there is a set of approaches and techniques that can provide practical alternatives to traditional statistics.

References

1. Mace, A. E., Sample-Size Determination, New York: Reinhold Publication Corporation, 1964, pp. 131-143.

2. Berger, J. O., Statistical Decision Theory and Bayesian Analysis, New York: Springer-Verlag, 1985, pp. 77-82, 273-278, 481-500.

3. Martz, H. F. and Waller, R. A., Bayesian Reliability Analysis, New York: John Wiley & Sons, 1982, pp. 487-541.

4. Barr, D. R., "A Class of General Reliability Growth Prediction Models," Operations Research, Vol. 18, 1970, pp. 52-65.