Authenticating Nuclear Warheads With High Confidence

Moritz Kütt,[*,**] Sebastien Philippe,[*]
Boaz Barak,[†] Alexander Glaser,[*] and Robert J. Goldston[*,‡]

[*] *Princeton University* [**] *Technische Universität Darmstadt, Germany*

[†] *Microsoft Research New England* [‡] *Princeton Plasma Physics Laboratory*

ABSTRACT. Negotiated deeper cuts in the nuclear arsenals may place limits on the total number of nuclear weapons in states' stockpiles, which could require inspections on hundreds or thousands of warheads or warhead components currently in storage or in dismantlement queues. The process of authenticating treaty limited items, e.g. with passive or active radiation measurements, is likely to be a critical element of verified disarmament. We have been developing a template-matching approach under which, for maximum security, classified information is never measured or stored electronically during the inspection. While one-on-one measurements (template versus inspected item) are feasible, such a system would strongly benefit from optimized sampling strategies authenticating as many items as possible with minimum probabilities for cheating. This paper proposes and examines the effectiveness of a strategy for authenticating nuclear warheads with a particular emphasis on controlling false positive and false negative rates of the inspections.

Background

Future reductions in the nuclear arsenals may place limits on the total number of nuclear weapons in states' stockpiles. Verification of such agreements could require inspections of hundreds or perhaps even thousands of warheads or warhead components currently in storage or in dismantlement queues. Under these circumstances, strategies to authenticate large number of items efficiently may become highly desirable. Inspections could be based on both the attribute or the template approach,[1] and both may have to deal with false positives (i.e., valid items declared bad) and false negatives (i.e., invalid items declared good). So far, proposed authentication systems have envisioned a single basic test that an item can either pass or fail.[2] Such basic tests, however, may have inherent technical limitations to achieve sufficiently low false-positive and false-negative rates in line with potential treaty verification objectives. To address this issue, we propose in this paper a formal protocol that determines how many times an item needs to be retested before it can be declared good or bad based on confidence levels chosen by the host and the inspector.[3] Below, we develop and discuss the method, illustrate sample results for a honest and a dishonest host with a particular emphasis on a template approach using a zero-knowledge protocol.

Methodology

We seek to distinguish between valid items and invalid items offered for inspection. We will do so by repeatedly applying a basic test to each item, until every item is declared either good or bad. We assume that the basic test is an (α, β) test, which means that the test will output pass with probability of at least α if a valid item is presented, while it will output pass with probability of at most β (with $\beta < \alpha$) if an invalid item is presented. We want to design a repetition strategy for this basic test such that a valid item will be declared good with probability of at least $\alpha^* > \alpha$, while an invalid item will be declared good with probability of at most $\beta^* < \beta$. For a viable inspection protocol, we will want α^* to be very close to 1 so that, in the honest-host case, with high probability all items are declared good. Similarly, in the dishonest-host (diversion) case, we will want that with high probability $(1 - \beta^*)$ an invalid item is declared bad. In sum, the four critical parameters are:

α : probability that valid item passes basic test (e.g. 0.95)
β : probability that invalid item passes basic test (0.05–0.10, but can be higher)

α^*: probability that valid item is declared good (host wants high, e.g. 0.99)
β^*: probability that invalid item is declared good (inspectors wants low, e.g. 0.10)

We build our formalism on statistical inference theory and proceed as follows. Suppose that $X_1, X_2, \ldots, X_n$ are n independent and identically distributed Bernoulli trials, i.e., random experiments with exactly two possible outcomes with constant probabilities p and $q = 1 - p$, representing a sequence of n (α, β) tests. We start by defining two statistical hypothesis tests.

First, to declare an item good: Let $H_0 \colon \theta = \beta$ and $H_1 \colon \theta = \alpha$ and $\beta < \alpha$, we reject the null hypothesis, $H_0 :$ the item is invalid, in favor of $H_1 :$ the item is valid and declare the item good, if $P \geq k_P$ where P is the number of pass in the sequence of n trials and k_P is a threshold chosen so that the probability of declaring the item good when it is in fact invalid is $Prob(P \geq k_P | H_0) \leq \beta^*$.

Similarly, to declare an item bad: Let $H_0 \colon \theta = (1 - \alpha)$ and $H_1 \colon \theta = (1 - \beta)$, we reject the null hypothesis, $H_0 :$ the item is valid, in favor of $H_1 :$ the item is invalid and declare the item bad, if $F \geq k_F$ where F is the number of fail in the sequence of n trials and k_F is a threshold chosen so that the probability of declaring the item bad when it is valid is $Prob(F \geq k_F | H_0) \leq 1 - \alpha^*$.

Tests are designed such that an item cannot be declared good and bad at the same time. It can be proven that both hypothesis tests are equivalent to performing likelihood-ratio tests on a sequence of n Bernoulli trials,[4] and that both are the most powerful tests available to accept valid items and reject invalid ones.[5]

We continue our analysis by specifying a set of $(\alpha, \beta, \alpha^*, \beta^*)$ values to generate a simple scorecard (Figure 1, left), showing the number of passed and failed basic tests. For every (P, F) position (with $n = P + F$), we check the following two inequalities based on the cumulative binomial probability:

$$\sum_{i=P}^{n} \binom{n}{i} \beta^i \, (1-\beta)^{n-i} \leq \beta^* \quad \text{and} \quad \sum_{i=F}^{n} \binom{n}{i} (1-\alpha)^i \, \alpha^{n-i} \leq 1 - \alpha^*$$

If the left inequality is true, we assign good; if the right inequality is true, we assign bad; all other positions remain inconclusive. We then search for the optimum number n_{opt} so that every sequence of basic tests with length n_{opt} will unambiguously declare items good or bad, i.e., never inconclusive. We choose $n_{opt} = n_{inc} + 1$, where $n_{inc} = P_{inc} + F_{inc}$ is the number of tests needed to reach the inconclusive field with highest n, for example, the position $(P_{inc} = 2, F_{inc} = 3)$ in Figure 1, left.[6] Finally, we use $n = n_{opt}$ to generate a reduced scorecard. On this scorecard, an item will be declared good as soon as $P \geq k_P = P_{inc} + 1$ and bad as soon as $F \geq k_F = F_{inc} + 1$ (see Figure 1, right, and Figure 5 in the Appendix). This method guarantees that an invalid item is never declared good with probability higher than β^* and that a valid item is never declared bad with probability higher than $(1 - \alpha^*)$.

Results

We distinguish two general situations: in the first case, the host is honest and all items are valid; in the second case, the host is cheating and at least one of the items offered for inspection is invalid. Of course, the inspector does not know which type of game is being played.

To illustrate the results below, we use a default parameter set to illustrate typical inspection outcomes. Any viable inspection system should be characterized by a basic test that passes a valid item with high probability α, but the value of β depends on the particular diversion scenario. In general, the higher the degree of similarity between the valid and the invalid item, the higher the value of β, and the more difficult to correctly declare the item bad in a basic test. Ultimately the value of β will depend on the agreed least detectable diversion of interest and the discrimination capability of the basic test. We use the reference values $\alpha = 0.95$ and $\beta = 0.05$.

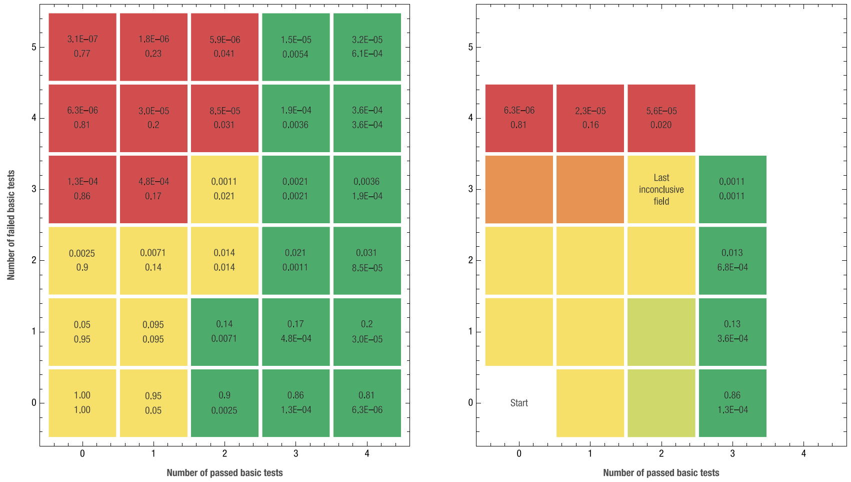

Figure 1: Left: Simple scorecard with results of hypothesis testing for every position (P, F) with $n = P + F$ (green is good, red is bad, and yellow is inconclusive). The probabilities to reach each position are given for both valid (top number) and invalid (bottom number) items. Right: Reduced scorecard obtained by using $n = n_{opt}$, the optimum number of basic tests, for every position. Probabilities to reach positions, where item are declared good and bad are given for both valid (top) and invalid (bottom) items.

Stopping criteria for the inspection depend on specific requirements set by the host and the inspector. In general, the honest host will require a high value for α^* to make sure that valid items are not incorrectly declared bad; similarly, the inspector will require a low value for β^* to make sure that a very low fraction of invalid items slip through. We use the reference values $\alpha^* = 0.999$ and $\beta^* = 0.01$ for stopping an inspection, but other values can be agreed upon by host and inspector.

Honest Host

In the case of the honest host, all items are valid, and the inspection should generally proceed smoothly. Figure 2 shows a typical realization for one hundred items and using the default values for $(\alpha, \beta, \alpha^*, \beta^*)$ during the inspection. Naturally, the honest host wants to avoid a situation where a valid item is declared bad. The odds of this outcome can be decreased by increasing the value of α^*, which requires additional testing. Table 1 summarizes the sensitivity of the inspection effort, i.e., the number of basic tests per item, for different stopping criteria. For $\beta < 0.40$, it is possible to reduce the probability of valid items being declared bad (e.g. from 0.001 to 0.0001) using only one additional inspection per item. For $\beta = 0.05$, only 3.2 tests are required to achieve 0.01% valid items being declared bad.

Dishonest (Cheating) Host

If the host is dishonest and tries to cheat by introducing one or more invalid items, it is best to analyze the situation from the inspector's perspective. The inspector is particularly worried about invalid items that are possibly declared good. Figure 3 shows a typical realization for the default values, and Table 2 shows the sensitivity of the inspection effort for different stopping criteria determined by the value of β^*. Similar to the case of the honest host, 1–2 additional basic tests per item can make a significant difference, especially for $\beta < 0.2$. Note that β^* represents the probability of missing one individual item; if more than one invalid item is in the batch, then the chances of finding at least one of them increases significantly.

		Probability that invalid item passes basic test (β)						
		0.01	0.05	0.10	0.20	0.30	0.40	0.50
Stopping criterion $(1-\alpha^*)$ (Valids declared "bad")	0.01	2.1	3.2	3.2	5.3	8.4	10.5	15.8
	0.001	2.1	3.2	4.2	6.3	8.4	13.7	17.9
	0.0001	2.1	3.2	4.2	7.4	9.5	14.7	21.1

Table 1: Lookup table for honest host listing the average number of tests per item before the inspection stops. Columns indicate performance of the basic test (β, probability that an invalid item passes basic test, which depends on the similarity of the invalid item compared to the valid item); rows indicate the probability $(1-\alpha^*)$ that a valid item is declared bad, i.e., the chances of an undesirable outcome for the host. Results are based on Monte Carlo simulations using the default values $\alpha = 0.95$ and $\beta^* = 0.01$.

		Probability that invalid item passes basic test (β)						
		0.01	0.05	0.10	0.20	0.30	0.40	0.50
Stopping criterion (β^*) (Invalids declared "good")	0.1	1.05	2.1	2.1	3.2	5.3	7.4	10.5
	0.01	2.1	3.2	4.2	6.3	8.4	13.7	17.9
	0.001	2.1	4.2	5.3	8.4	11.6	16.8	25.3

Table 2: Lookup table for inspector listing the average number of tests per item before the inspection stops. Columns indicate performance of the basic test (β, probability that an invalid item passes basic test, which depends on the similarity of the invalid item compared to the valid item); rows indicate the probability that an invalid item is declared good (β^*), i.e., the chances of an undesirable outcome for the inspector. Results are based on Monte Carlo simulations using the default values $\alpha = 0.95$ and $\alpha^* = 0.999$.

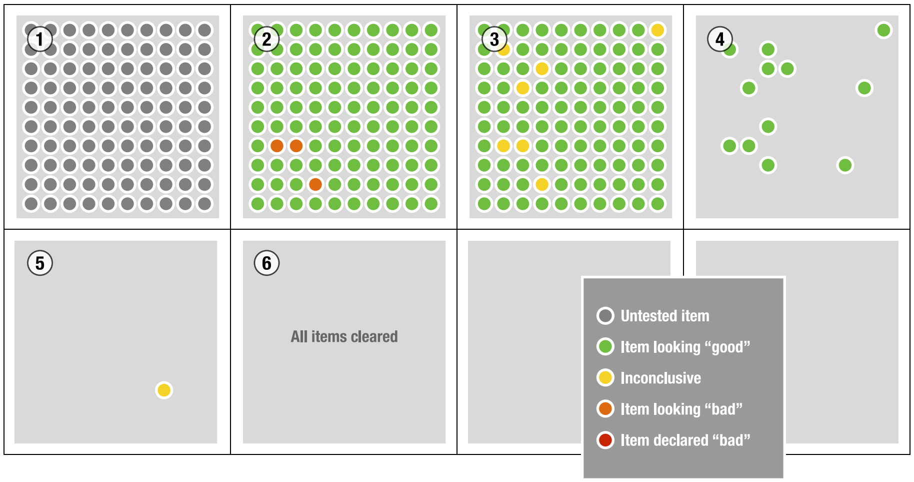

Figure 2: Testing a batch of one hundred valid items. In this particular realization, after five rounds and 313 basic tests, the inspection stops and clears all items. Parameters used for this realization: $\alpha = 0.95$, $\beta = 0.05$, $\alpha^* = 0.999$, $\beta^* = 0.01$.

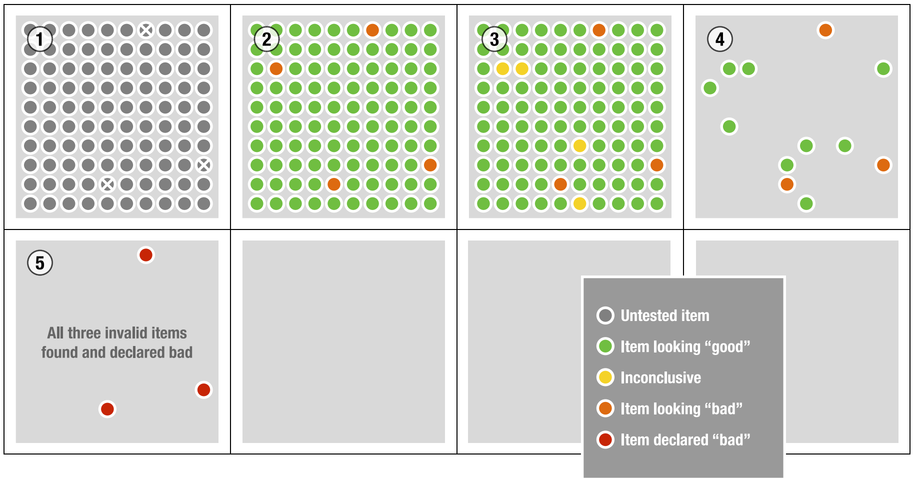

Figure 3: Testing a batch of one hundred items that includes, unknown to the inspector, three invalid items, which are highlighted (x) in the first panel. In this particular realization, after four rounds and 312 total tests, the inspection stops identifying all three invalid items correctly. Parameters used for this realization: $\alpha = 0.95$, $\beta = 0.05$, $\alpha^* = 0.999$, $\beta^* = 0.01$. For this value of β^*, the probability of catching all three invalid items is about 97%; the probability of catching at least one of them is greater than 99.999% (assuming uncorrelated statistical errors).

Special Case: Zero-knowledge Template Approach

We have previously proposed a template-matching approach under which, for maximum security, classified information is never measured. The details of our basic (one-on-one) approach have been described elsewhere.[7] As with all template approaches, prior to the inspection, a template could be selected at a deployment site in order to have high confidence in the authenticity of the item. The template is then placed in a storage container and, with strong chain-of-custody measures, brought to a dedicated dismantlement facility, where the containerized warheads slated for dismantlement are also located. For the basic one-on-one approach, the host prepares pairs of preloaded detectors, which are arranged in two arrays. Critically, the inspector then chooses which array is used on which item. Preloads are secretly determined by an honest host such that, after the irradiation, the detector count obtained by any measurement on the template or on any valid submitted item is distributed according to the Poisson distribution with mean equal to a previously agreed-on value. This value is known in advance by both sides. Therefore, neither the measurement nor its noise reveals any new information.

Given that, for this basic one-on-one approach, the template is directly compared against each inspected item individually, we would have to generate the reference signature in every single authentication process. This could pose various challenges, in particular for the situation described above, where it would be advantageous to authenticate many identical items as efficiently as possible. Using the methodology developed above, we propose a strategy to inspect a batch of N items simultaneously to check if all these items, including the template, are identical.

We assume that $K \geq 1$ templates are available and that, by definition, confidence in the authenticity of these templates is high. With regard to the inspection protocol, there is no qualitative difference between templates and the items offered for inspection. In fact, as we will see, the template may be eliminated from the batch of N items early on. Critically, the host prepares in advance preloads for all the basic tests we will need to perform. For example, if N items are in the batch and, on average, each item will not be tested more than four times, then $(4 \times N)$ preloaded detectors (or detector arrays) will need to be provided at the outset of the inspection. Depending on the type of detector, this preparation could be challenging.[8] Through this commitment, the host does not gain anything, and can in fact lose, from using unequal preloads (with the intention of concealing one or more invalid items in the batch), and hence we can assume the preloads are all identical. Preloads could also be validated through testing randomly selected detectors on the template(s) prior to inspection. Figure 4 shows the outcome of a typical inspection, which includes 99 items offered for inspection and one template.

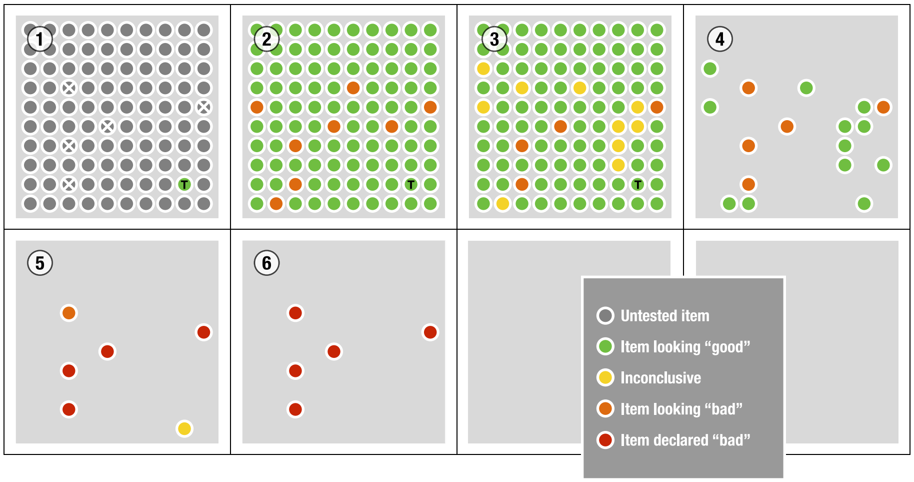

Figure 4: Testing a batch of one hundred items that includes one template (T) and, unknown to the inspector, five invalid items (x). In this particular realization, after six rounds of testing, all five invalid items have been found and declared bad. Overall, 324 basic tests are needed for this realization. Parameters used for this realization: $\alpha = 0.95$, $\beta = 0.05$, $\alpha^* = 0.999$, $\beta^* = 0.01$.

Conclusion

In this paper, we have proposed a formal protocol that can be used to control the probabilities of inspection failures for efficient nuclear warhead authentication. Based on the likelihood-ratio method, the protocol can be used for both attribute-type and template-type measurements; it is particularly useful when large numbers of identical items have to be authenticated and when measurement outcomes are communicated in a pass/fail manner, for example through an information barrier. The protocol allows the host and the inspector to adjust testing parameters such that both achieve their desired performance requirements with a minimum inspection effort: generally, the honest host wants few if any valid items declared bad, while the inspector wants few if any invalid items declared good. For an efficient inspection process, the probability β of an invalid item passing a basic test needs to be as low as possible. This probability, however, depends on the particular diversion scenario and cannot be defined independently. We find that values for β of up to 0.2–0.3 may be acceptable; above this range, the number of required test repetitions increases sharply. Further research needs to determine the potential impact of systematic errors, which have not been considered in this analysis, and the tradeoffs between repeated tests and longer individual inspections (i.e., with better statistics for radiation measurements), which also reduce the false positive and false negative rates.

Appendix

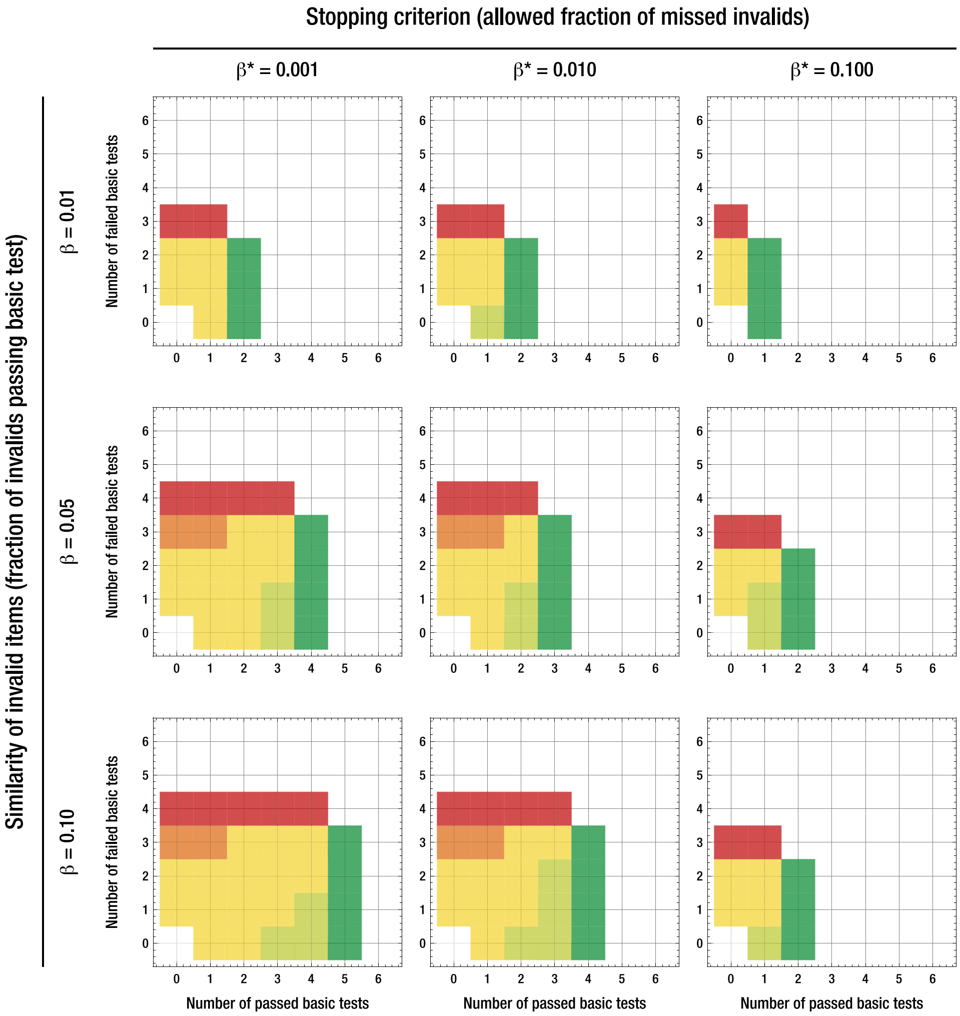

Figure 5: Scorecards for selected β and β^* values.

($\alpha = 0.95$ and $\alpha^* = 0.999$)

Endnotes

[1] David Spears (ed.), *Technology R&D for Arms Control,* U.S. Department of Energy, Office of Nonproliferation Research and Engineering, Washington, DC, 2001.

[2] A basic test could, in principle, provide more information than a simple pass/fail result. If information barriers are used to process measured data and remove more detailed information generated in the measurement, then it is impossible to know if a test failed by a small or large margin. In the absence of an information barrier, as in the zero-knowledge approach, this information could be available. Note also that inspection failures can be due random and systematic errors. This analysis only focuses on failures that are statistical in nature. Other types of tests can be devised to assure that the inspection system is working as designed and that the inspected item is properly prepared to a specified level of systematic uncertainty.

[3] Note that pass and fail are properties for the basic test while good and bad are properties for a repeated test sequence.

[4] For the proof, see http://nuclearfutures.princeton.edu/warhead-likelihood.pdf.

[5] This is equivalent to saying that both tests are the best at declaring an item good when it is valid and bad when it is invalid. For the proof see: J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character,* 231 (1933), pp. 289–337.

[6] See proof from Endnote 4 on how to obtain n_{opt}.

[7] A. Glaser, B. Barak, and R. J. Goldston, "A Zero-knowledge Protocol for Nuclear Warhead Verification," *Nature,* 510, 26 June 2014, pp. 497–502.

[8] R. J. Goldston, F. d'Errico, A. di Fulvio, A. Glaser, S. Philippe, and M. Walker, "Zero Knowledge Warhead Verification: System Requirements and Detector Technology," 55th Annual INMM Meeting, Atlanta, GA, July 2014.

Revision 1, July 2014