Verifying the Authenticity of Nuclear Warheads Without Revealing Sensitive Design Information

Steve Fetter and Thomas B. Cochran

Verifying the dismantlement of nuclear warheads will require reconciling two conflicting objectives: the desire of the monitoring party to insure that the objects slated for dismantlement are *bona fide* warheads of the declared type, and the desire of the monitored party to protect sensitive information about the design of the warhead.[1]

A possible solution would involve visiting a deployment site on short notice and randomly selecting a given number of warheads for dismantlement. The warheads would then be placed in tagged, sealed containers for transport to the dismantlement facility, where the integrity of the tags and seals would be verified. If the number of warheads to be dismantled is a small fraction of the entire inventory, then the monitoring party would be reasonably sure that the warheads are genuine, for the only way the monitored party could defeat the scheme would be to deploy large numbers of fake warheads. Still, the process of on-site tagging and sealing for each warhead is tedious, and the monitored party would have no assurance that all the warheads were genuine, since the monitored party could easily replace 10 or 20 percent of the warheads slated for dismantlement with decoys.

A much better solution would involve gathering only a small sample of warheads during an initial random on-site inspection and establishing a unique "fingerprint" or signature for this warhead type. An excellent example of a fingerprint would be the warhead's characteristic gamma-ray emissions.[2] Subsequent warheads could simply be brought to the dismantlement facility at the convenience of the monitored party, where a warhead's fingerprint would be compared with those of the reference warheads to establish its authenticity.[3]

---

[1] For a more general discussion of verifying reductions in nuclear warheads, see ***Ending the Production of Fissile Materials for Weapons; Verifying the Dismantlement of Nuclear Warheads: The Technical Basis for Action*** (Washington, DC: Federation of American Scientists, June 1991).

[2] Other types of fingerprints can be imagined, but we have identified none that have convincing advantages over intrinsic gamma-ray emissions. One could, for example, radiograph the warhead, but the algorithm to compare radiographs would be far more complex than the algorithm to compare the intensity of gamma-ray emissions, the information contained in the radiographs would be far more sensitive than that contained in the gamma-ray spectra, and the possibility of cheating by substituting non-weapons-grade uranium and plutonium for weapons-grade uranium and plutonium, which would lead to identical X-rays, would require the use of gamma-ray and/or neutron detection as a supplement in any case.

[3] Alternatively, one could compare the fingerprints of disassembled components leaving the facility, to verify that a given number of a given type of warhead had been dismantled. One could, for example, place the plutonium pit in a given container and verify this fingerprint, although the gamma-ray fingerprint described below would presumably reveal more sensitive information in this case if the design of the container was known.

All nuclear warheads contain radioactive isotopes of uranium and/or plutonium.[4] Most of these isotopes emit gamma rays and neutrons at rates that can be detected outside the warhead. For example, plutonium-239—the most important plutonium isotope—emits over two dozen gamma rays that can be detected outside of typical warheads. Since the energies of the gamma rays are determined by the characteristics of the emitting nucleus, detecting just one or two of these gamma rays permits the unambiguous identification of the radioactive material. And since the intensity of each gamma ray depends on the quantity and geometry of the radioactive material and the thickness, geometry, and atomic number of all surrounding materials, it would be extremely difficult—if not impossible—to replicate the gamma-ray spectrum with a smaller amount of material in a different configuration.

The fingerprint should be based on emissions that would not be expected to vary by more than a few percent from warhead to warhead. Excellent choices would be the gamma rays emitted during the decay of uranium-235 and plutonium-239. The amount of uranium-235 and plutonium-239 in a given type of warhead, and the intensity of their gamma rays outside the warhead, should vary very little from warhead to warhead, and would not depend on the age of the material.[5] Gamma rays are emitted during the decay of other isotopes of uranium and plutonium (e.g., uranium-232, uranium-238, plutonium-241, etc.), but the concentration of these isotopes could easily vary by more than 20 percent from warhead to warhead. Copious neutrons are emitted by plutonium-240, but the percentage of this isotope in weapon-grade plutonium might vary by more than 50 percent from warhead to warhead (e.g., from 3 to 6 percent or more).

The energy and intensity of the gamma rays emitted by plutonium-239 are such that one can expect to receive a detectable signal outside any warhead that contains plutonium. It is, however, possible to build warheads that use uranium-235 for fissile components, and which contain no plutonium. The gamma rays emitted by uranium-235 are considerably less energetic than those emitted by plutonium-239; if the uranium is surrounded by thick materials with high atomic numbers, these gamma rays may be undetectable outside the warhead. Shielded counting chambers and long counting times will help, but may not guarantee detection of these low-energy gamma rays in some conceivable warhead designs. We do not believe that such designs are common in the superpower arsenals, but they are a possibility that must be considered.

To deal with the possible failure of gamma-ray detection, one could illuminate the warhead with bursts of high-energy neutrons. The neutrons would penetrate to the fissile material, causing a certain number of fissions and the consequent emission of prompt and delayed neutrons and gamma rays. Like the gamma rays emitted during radioactive decay, these emissions would depend on the quantity and geometry of the fissile material, the thickness, geometry, and isotopic composition of all surrounding materials, and the incident neutron energy. Inelastic absorption of

[4] For a more general and complete discussion of warhead detection, see Steve Fetter, Valery A. Frolov, Marvin Miller, Robert Mozley, Oleg F. Prilutsky, Stanislav N. Rodionov, and Roald Z. Sagdeev, "Detecting Nuclear Warheads," ***Science and Global Security***, Vol. 1, pp. 225-302 (1990).

[5] An exception might be non-fissile uranium components, in which the concentration of uranium-235 is low (between 0.2 and 0.7 percent). The concentration of uranium-235 in such components could vary by more than a factor of two from warhead to warhead if components were fabricated from depleted uranium of different assays, or if some components were fabricated from natural uranium and others from depleted uranium. One could deal with such cases by measuring the much stronger gamma rays emitted during the decay of uranium-238.

neutrons by surrounding materials would also lead to the emission of characteristic gamma rays that could be used in the warhead signature as well. Although a neutron source would greatly increase the cost the system, it would guarantee obtaining a fingerprint that would be extremely difficult to counterfeit.

A major objection to radiation fingerprinting is that it may reveal sensitive information about the design of the nuclear warhead. By working backwards from the gamma-ray spectra, it may be feared that the monitoring party could reconstruct previously unknown aspects of the warhead design, which could then be used to improve warhead design or strategic defenses. Although it seems highly unlikely to us that valuable information can be derived from such fingerprints (in the sense that it would allow meaningful improvements in nuclear offense or defense), it is reasonable to anticipate and deal with such objections whenever possible.

In the following example, we outline an automated system that would return a simple "yes" or "no" answer to the question: is the radiation fingerprint of this warhead significantly different from the fingerprints of the reference warheads?

First, the monitored party would provide a list of all locations where warheads of the type to be dismantled are deployed or stored. The monitoring party would then visit several of these sites on short notice, and randomly select a small number (e.g., a total of ten) warheads. (Selecting more than one warhead allows the natural variability in the warheads to be estimated, which will be valuable in minimizing false accusations of cheating.) Each warhead would be placed in a tagged and sealed container, which would be shipped to the dismantlement facility. Upon arrival, the authenticity of the tags and seals would be checked to ensure that the arriving warheads are the same as those selected from the declared sites.

Next, the warheads would be placed, one by one, on a turntable in a shielded chamber. (Rotating the warhead with respect to the detector would make it unnecessary to reproduce the exact orientation of the warhead.) While the warhead is rotated in the chamber, gamma rays emitted by the warhead are detected by high-purity germanium detectors located at different heights above the turntable.[6] The counting time should be long enough so that statistical counting errors are small (e.g., 1 percent, or smaller than the natural sample variability).

[As mentioned above, a neutron source could also be used to produce prompt fission, fission-product, and neutron-absorption gamma rays and prompt and delayed fission neutrons. For simplicity, only gamma rays from radioactive decay are discussed below, although the techniques are easily generalized to include neutron-induced emissions.]

Immediately before and/or after each measurement, background measurements and calibration measurements would be made. The calibration sources would ideally include a small quantity of uranium-235 and plutonium-239. Note that it is only necessary to measure changes in the relative efficiency of the detectors. These measurements would also serve to indicate the proper functioning of the equipment.

[6] Alternatively, a single detector could be used and the elevation of the turntable varied. Also note that warheads and bombs will probably require different shape counting chambers.

Each spectrum would then be analyzed by a peak-finding program to determine the energy and intensity of each gamma-ray emission (e.g., significant at the 3-standard-deviation level), the uncertainty in these quantities, and to identify those emissions that are due to the decay of uranium-235 and plutonium-239.[7] At this point, uncertainties are due mainly to counting statistics. For example, if 100 gamma rays of a given energy were detected, then (neglecting background) the uncertainty in the measurement would be the square root of the number of counts—in this case, 10 counts. A 1 percent uncertainty would require 10,000 counts (assuming the background is small). This number of counts per peak might be achievable in less than an hour for plutonium-bearing warheads.[8]

After the spectra are corrected for variations in detector efficiency, we have a library of gamma-ray energies E_{ij}, observed gamma-ray count rates x_{ij}, and uncertainties in these count rates s_{ij} (including uncertainties from counting statistics, background, and relative detector efficiency) for each warhead i and gamma ray j at each detector location. The mean count rate $\overline{x}_j$, averaged over the group of reference warheads, is given by

$$\overline{x}_j = \frac{\sum_{i=1}^{m} \frac{x_{ij}}{s_{ij}^2}}{\sum_{i=1}^{m} \frac{1}{s_{ij}^2}} \approx \frac{1}{m} \sum_{i=1}^{m} x_{ij} \tag{1}$$

If statistics are poor, the simple average may be a more robust measure of the mean than the weighted average. The variance in the count rate is given by

$$\sigma_j^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left(x_{ij} - \overline{x}_j \right)^2 \tag{2}$$

Before proceeding, we should ask two questions about the group of reference warheads: (1) are these warheads nearly identical, and (2) if they are not identical, are they all of the same type? Answering the first question is important because it determines the type of statistical analysis that should be used in comparing fingerprints. If the warheads are nearly identical (i.e., if variations in the count rates due to variations in manufacture are much smaller than variations due to counting statistics), then we can assume that deviations from the mean count rates are statistically independent, and we can use the much simpler (and more powerful) χ^2 statistic to test whether fingerprints match. If, however, manufacturing tolerances lead to variations in the count rates that exceed those due to counting statistics, then differences in count rates probably

[7] An example of such program is HYPERMET. See G.W. Phillips and K.W. Marlow, ***Program HYPERMET for Automatic Analysis of Gamma-ray Spectra from Germanium Detectors*** (Washington, DC: Naval Research Laboratory, Report NRL-3198, 1976).

[8] In measurements made on a Soviet warhead, count rates for 14 different lines varied from 0.05 to 1.6 counts/s. [Steve Fetter, Thomas B. Cochran, Lee Grodzins, Harvey L. Lynch, and Martin S. Zucker, "Gamma-Ray Measurements of a Soviet Cruise Missile Warhead," ***Science***, Vol., 248, pp. 828-834 (18 May 1990).] The measurements were, however, made about 75 cm from the center of the warhead, through a steel launch tube about 7 cm thick. At a distance of 50 cm and without the steel launch tube, count rates would have been 5 to 400 counts/s—enough for over 10,000 counts per peak in less than an hour.

will be correlated and the more complicated and less powerful T^2 statistic must be used to test the equivalence of two fingerprints. Nuclear warheads are perhaps the most carefully constructed items on the planet, and we doubt that manufacturing and assembly tolerances are so lax that they would result in large variations in gamma-ray emissions.[9] It is, nevertheless, a possibility that must be anticipated in advance.

One way to determine whether the warheads are nearly identical is to compare the variance calculated in equation (2) with the variance the would be expected based solely on counting statistics, which is given by

$$S_j^2 = \frac{1}{\sum_{i=1}^{m} \frac{1}{s_{ij}^2}} \tag{3}$$

If the observed variance in the count rate is much greater than what would be expected from counting statistics alone (i.e., if $\sigma_j^2 \gg S_j^2$), then warhead-to-warhead variations are important and must be taken into account. Unlike errors from counting statistics, which are independent of gamma-ray energy, variations in manufacture will lead to differences in count rates that are correlated with each other. For example, a small decrease in the amount of plutonium in a warhead will cause all count rates to decrease, though some will decrease more than others.

A more robust method of determining whether differences in count rates are correlated is to test the hypothesis that the correlation is zero. The coefficient for the correlation between the count rates for gamma-ray emissions j and k ($j \neq k$) is given by

$$r_{jk} = \frac{\sum_{i=1}^{m} \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{(m-1)\sigma_j \sigma_k} \tag{4}$$

The statistic z_{jk}, which is given by

$$z_{jk} = \frac{1}{2} \log_e \left(\frac{1 + r_{jk}}{1 - r_{jk}} \right) \tag{5}$$

is normally distributed with a standard deviation of about $(m - 3)^{-½}$. Therefore, the probability that the statistic is greater than z_{jk} given that the correlation is zero is given by

[9] To note an extreme example, the mean-free-path of a 186-keV gamma-ray (the most intense emission from U-235) is only 0.36 mm in uranium. If the thickness of a uranium component between the source and the detector varied by as little as one-thousandth of an inch (1 mil) from warhead to warhead, the intensity of the gamma ray at the detector would vary by over 10 percent. In the U.S., heavy-metal weapon components are typically machined to tolerances of ±1 to ±2 mils, and there might be five to six heavy-metal layers in a typical assembly. (Richard Hatfield, Lawrence Livermore National Laboratory, personal communication, 6 August 1991.)

$$\mathrm{P}\left(z_{jk}\,|m\right)=\mathrm{erfc}\left[z_{jk}\sqrt{\frac{m-3}{2}}\right] \tag{6}$$

where

$$\mathrm{erfc}\left(x\right)=\frac{2}{\sqrt{\pi}}\int_{x}^{\infty}\mathrm{e}^{-t^{2}}dt \tag{7}$$

There will be a total of $n(n–1)/2$ covariances; if q is the probability that no one of these covariances appear non-zero from random variations, then $P(z_{jk} \mid m)$ should satisfy the condition

$$\mathrm{P}\left(z_{jk}\,|m\right)\le 1-\left(1-q\right)^{\frac{2}{n(n-1)}} \tag{8}$$

If $q = 0.05$ and $n = m = 10$, then $P(z_{jk} \mid m) \le 0.001$ and $z_{jk} \ge 0.87$ for any j and k to reject the null hypothesis that the count rates are uncorrelated. A better procedure might be to use a more lax criterion (e.g., $P(z_{jk} \mid m) \le 0.05$, $z_{jk} \ge 0.66$), and simply require that more than, say, 20 percent of the z_{jk} exceed this criterion in order to reject the null hypothesis.

If the null hypothesis is accepted (i.e., the reference warheads are nearly identical), then we can safely assume that they are all of the same type. If they are not identical, then we must check to see if the data are consistent with the assumption that the reference warheads are all of the same type (i.e., that none of the warheads are dramatically different from the others). This is important because it could reveal an attempt by the monitored party to fool the system into accepting a large sample variability, which would make cheating much easier. To account for this, one could compare the count rates for (m–1) warheads with the count rates for the remaining warhead, for all m warheads; then compare the count rates for (m–2) warheads with count rates for the remaining two warheads, for all combinations of two warheads; and so on. The probability that the two groups of warheads are different is given by the T^2 statistic developed below. The computer program could alert the monitoring party of a significant difference between two subgroups of warheads, which could then request additional information from the monitored party before proceeding.

Let us assume that we accept the hypothesis that all the references are of the same type, and that the observed variations between warheads are due to counting statistics and variations in manufacture. When the monitored party brings a new warhead to the dismantlement facility, the monitoring party would attempt to verify its authenticity by comparing its radiation fingerprint to that of the reference set of warheads. Once again, the peak-finding and peak-fitting program would identify all lines associated with the decay of uranium-235 and plutonium-239 and would estimate the count rate of each emission y_j and the associated uncertainty s_{y_j}.

If we have also accepted the hypothesis that the warheads are nearly identical, then a simple chi-square test can be used to determine if a new warhead belongs to the same set as the reference warheads:

$$\chi^2 = \sum_{j=1}^{n} \frac{\left(\bar{x}_j - y_j\right)^2}{\frac{\sigma_j^2}{m} + s_{y_j}^2} \tag{9}$$

[Equation (9) assumes that the same gamma-ray emissions are detected in each case. If this is not true, the analysis could be restricted to only those peaks found to be significant in both spectra, or the spectra could be reanalyzed to determine the possible magnitude of "missing" peaks. The preferred technique probably depends on the number of such cases and the quality of the data in each case.]

The probability that the chi-square observed for the new warhead will exceed the value χ^2 by chance (i.e., the probability that the warhead is drawn from the population represented by the reference warheads and that the observed deviations in the spectra are due to random variations) is given by the incomplete gamma function:

$$\mathrm{P}\left(\chi^2 \middle| n\right) = \frac{1}{\Gamma\left(\frac{n}{2}\right)} \int_{\frac{\chi^2}{2}}^{\infty} \mathrm{e}^{-t} t^{\frac{n}{2}-1} \mathrm{d}t \tag{10}$$

Before making use of this equation, we must decide what probability of a false alarm is tolerable. If, for example, we reject warheads for which $\mathrm{P}(\chi^2 \mid n) < 0.01$, then on average one out of every 100 legitimate warheads would rejected. Since dismantlement campaigns might involve over 1,000 warheads, this is clearly unacceptable, since false accusations of cheating could be very damaging.

The total number of warheads that will be dismantled by the superpowers in the coming decades is probably between 10,000 and 50,000. If we require that the probability of a false alarm during the entire dismantlement campaign is fairly remote (e.g., 1 to 5 percent), then the criteria for rejecting a warhead should be $\mathrm{P}(\chi^2 \mid n) < 10^{-6}$. The following table gives the required χ^2 for a given n and $\mathrm{P}(\chi^2 \mid n)$:

	$\mathrm{P}(\chi^2 \mid n) = 10^{-5}$		$\mathrm{P}(\chi^2 \mid n) = 10^{-6}$		$\mathrm{P}(\chi^2 \mid n) = 10^{-7}$	
n	χ^2	χ^2/n	χ^2	χ^2/n	χ^2	χ^2/n
2	23.0	11.5	27.6	13.8	32.2	16.1
4	28.5	7.12	33.4	8.34	38.1	9.52
6	33.1	5.52	38.3	6.38	43.2	7.20
8	37.3	4.67	42.7	5.34	47.9	5.99
10	41.3	4.13	46.8	4.68	52.3	5.23
12	45.1	3.76	50.8	4.23	56.3	4.69
14	48.7	3.48	54.6	3.90	60.3	4.31
16	52.2	3.27	58.3	3.65	64.1	4.01
18	55.7	3.09	61.9	3.44	67.9	3.77
20	59.0	2.95	65.4	3.27	71.5	3.58
25	67.2	2.69	73.9	2.96	80.3	3.21

To see that false alarm rates as low as 10^{-6} may be possible while preventing significant diversion of material, assume, for the sake of simplicity, that $s_{y_j}^2 \approx y_j$, $s_{ij}^2 \approx x_{ij}$ and $\sigma_j^2 \approx \overline{x}_j / m$, where x and y are now the total number of counts; then

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(\overline{x}_j - y_j\right)^2}{\frac{\overline{x}_j}{m} + y_j} \tag{11}$$

Assuming that $y_j \approx \overline{x}_j$, the allowable root-mean-square (rms) relative difference in the count rate is approximately equal to $\left(\chi^2/n\right)\left(1+m^{-1}\right)/\overline{x}$; if we detect an average of just 1,000 counts in each of four lines, the allowable rms error would be less than 1 percent for a false alarm rate of 10^{-6}.

If, on the other hand, we rejected the hypothesis that the warheads are nearly identical (i.e., that warhead-to-warhead variations lead to correlated differences in count rates), the statistic we should compute is T^2, which in vector notation is as follows:

$$T^2 = \frac{m}{m+1}\left(\overline{\mathbf{x}} - \mathbf{y}\right)^{\mathrm{T}} \mathbf{c}^{-1} \left(\overline{\mathbf{x}} - \mathbf{y}\right) \tag{12}$$

The elements of the vector $\left(\overline{\mathbf{x}} - \mathbf{y}\right)$ are $\left(\overline{x}_j - y_j\right)$. Note that this differs from the normal two-sample problem in that the second sample contains just one set of observations, **y**. Therefore, the covariances are estimated from the observations of the reference group of warheads **x** (and are assumed to equal the covariances from the group that **y** was drawn from). The elements of the covariance matrix **c** are as follows:

$$c_{jk} = \frac{1}{m-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)\left(x_{ik} - \overline{x}_k\right) \tag{13}$$

It can be shown that the T^2 statistic is distributed according to the F distribution:

$$F_{n,m-n-1} = \frac{m-n-1}{n\left(m-1\right)} T^2 = \frac{m\left(m-n-1\right)}{n\left(m-1\right)\left(m+1\right)}\left(\overline{\mathbf{x}} - \mathbf{y}\right)^{\mathrm{T}} \mathbf{c}^{-1} \left(\overline{\mathbf{x}} - \mathbf{y}\right) \tag{14}$$

The probability Q that the F observed will exceed the value F' by chance (i.e., the probability that the new warhead is drawn from the population represented by the reference warheads) is given by

$$Q\left(F' \middle| n, m-n-1\right) = I\left(\frac{m-n-1}{m-n-1+nF}, \frac{m-n-1}{2}, \frac{n}{2}\right) \tag{15}$$

where the function I is equal to

$$I(x,a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t \tag{16}$$

Note that the number of gamma-ray emissions used computing the T^2 statistic (n) must be less than the number of warheads in the reference set (m). If $n \geq m-1$, then n should be reduced by choosing only the most significant emissions, taking care to space them as evenly as possible with gamma-ray energy. It should be fairly straightforward to design an algorithm to automatically select the best emissions for such an analysis.

Without actual data on different warheads of a given type, it is impossible to estimate the power of the T^2 statistic to discriminate between phony warheads and warheads with significant correlated differences. We believe that it should be possible to limit the acceptable difference in the count rates to less than 10 percent, at least in plutonium-bearing warheads. If this is insufficient, discrimination power could be increased considerably by storing the spectra from each warhead and using this information to refine the estimates with each new warhead. By adding the spectra of each acceptable warhead to the set of "reference" warheads, estimates of the covariances could be improved substantially, and more lines could be added to the analysis. In addition to asking whether the new warhead is statistically different from the set of reference warheads, the computer could ask whether any subgroup yet examined is statistically different from any other subgroup. This would eliminate the possibility of, for example, removing 10 percent of the plutonium from a large fraction of the warheads to be dismantled; even though a 10 percent difference may not be noticed in single warhead, a disproportionately large fraction of warheads showing a 10 percent difference *would* be noticed. Such techniques may be especially valuable in cases where few gamma-ray emissions are detectable (e.g., uranium-only warheads).

Note that all the data analysis discussed above could be performed automatically by a computer, without human intervention. Algorithms have been developed that can automatically and reliably find and identify statistically significant peaks in gamma-ray spectra; it would be relatively straightforward to analyze and store this information in a secure manner for comparison with future data. The computer could be programmed to provide a simple "yes" or "no" answer to the question, "Is the warhead under examination statistically different from group of reference warheads?", or it could provide the probability that the observed difference is due to random variations. The computer could also answer the question, "Is any subgroup of warheads examined so far statistically different from the other warheads, and if so, which warheads and by how much?" There would be no need to reveal any aspect of the spectra themselves, although such revelations would aid greatly in building confidence in the system and resolving ambiguities. At a minimum, the computer could indicate the presence of uranium-235 or plutonium-239 and the number of statistically significant peaks detected. In addition, the energy of the peaks might be given, along with the count rates for a few of the strongest emissions.

A great amount of attention obviously would have to be given to "red-teaming," or finding out if and how the system could be fooled into accepted illegitimate warheads. While it seems to us highly unlikely that the gamma-ray fingerprint of a legitimate warhead could be matched by a dummy warhead containing less fissile material, this possibility could be better evaluated by those knowledgeable in weapon design.