Chapter 2

Seismic Verification in the Context of National Security

CONTENTS

Boxes

Figures

Table

Chapter 2

Seismic Verification in the Context of National Security

Seismic monitoring is central to considerations of verification, test ban treaties, and national security.

THE ROLE OF VERIFICATION

For an arms control agreement to be successful, each participating country must feel that the provisions of the treaty will enhance its own national security. This requires an evaluation by each country of the costs and benefits to its national security of the treaty's restrictions. In the case of nuclear test ban treaties, the cost is accepting restrictions on the ability to test nuclear weapons. In return for paying this price, each country gains the direct benefit of similarly restricting the other country. In addition to the direct benefits of the agreement, participating countries can also gain the political and non-proliferation benefits of working for arms control, as well as the environmental benefits of reducing the hazards of radioactive contamination of the environment from testing.

In considering agreements that bear on such vital matters as a restriction on nuclear weapon's development, each country may assume as a cautious working hypothesis that the participants would cheat if they were sufficiently confident that they would not be caught. Verification is a process that is undertaken to confirm treaty compliance and, therefore, the ability of the United States to monitor Soviet activity is central to the value of any such treaty.

Verification is most often viewed as a process that improves confidence in a treaty. The converse, however, is also true. Establishing a treaty generally improves our ability to monitor Soviet activity. In this way, monitoring and treaties have a mutually beneficial relationship. The United States monitors Soviet weapons developments whether or not a treaty exists, because the information is important for our national security. Treaties make monitoring easier and more accurate, because they include provisions explicitly intended to aid verification. Additionally, treaties create paths of communication that can be used to resolve, clarify, or correct ambiguous situations. In this sense, treaties have national security value that extends beyond their direct purpose.

Verification of treaties is a complex process. The question of whether a treaty is "verifiable" cannot be answered in any absolute or technical sense. It can only be answered relatively by referring to values that are influenced by a wide range of political and philosophical viewpoints. Consequently, verification involves not only technical considerations, but also judgements as to how these technical considerations translate into the policy world.

In the past several years, Congress has been asked to consider proposals for treaties that prohibit testing above various thresholds. Each proposal has sparked controversy within both the technical and policy communities. For any given treaty, some within both communities will claim it is "verifiable," whereas others will assert that the Soviet Union would be able to cheat and hence that stricter verification provisions are needed to ensure our national security. This chapter is intended to provide a framework for understanding how to weigh the risks and benefits of such treaties.

76-584 0 - 88 : QL 3 - 2

THE DEFINITION AND VALUE JUDGMENT OF "VERIFICATION"

In the case of test ban treaties, measures are taken to ensure that the advantages of the treaty cannot be undermined by the other country testing clandestinely. These measures, assessments of the measures' capabilities, evaluations of the risks and benefits, and the political climate within which all of these judgments are made make up the process referred to as verification.

"To verify" means to establish truth or accuracy. Realistically, in the arena of arms control, verification can never be perfect or absolute: it necessarily involves uncertainty and this is often described in probabilistic terms. Because the process of verification involves determining acceptable levels of uncertainty, it is political as well as scientific. The degree of verification needed is based on one's perception of the benefits of the treaty compared with one's perception of the disadvantages and the likelihood of violations. Consequently, the level of verification required will always be different for people with different perspectives.

In U.S.-Soviet agreements, the concern about verification is exacerbated by societal asymmetries whereby monitoring compliance is usually achieved more easily in the United States than in the Soviet Union. These asymmetries may cause the United States to insist on stricter verification procedures than the Soviets would judge are needed. This difference makes negotiations difficult, and can create the impression that the United States is obstructing negotiations.

A country considering cheating would have to evaluate the risks and costs of being caught against the benefits of succeeding. A country concerned about preventing cheating has to guess the other country's values for making this decision and then evaluate them against their own estimations of the advantages of the treaty compared with the risk of violation. If the countries lack insight into each other's value systems and decision processes, this uncertainty will result in the perception that a high degree of verification is needed. As a result, the degree of verification needed to satisfy the concerned country may be higher than what is really needed to discourage cheating.

To illustrate this argument, it is useful to consider the analogy of a treaty restricting each party to one side of a river. If the river freezes over, one or both countries may consider crossing to the other side. If the water is deep and there is nothing worth having on the other side, then the ice does not have to be very thin to discourage a party from crossing. If, on the other hand, the water is shallow and there is something of great value to be obtained from the other side, than the ice must be very thin to discourage a party from crossing. The thinness of the ice combined with the depth of the water is the degree of deterrent available to dissuade a party from trying to cross the ice. How thin it has to be to actually deter depends on each party's perception of the risk and the reward. In arms control, crossing the ice represents cheating on a treaty. The level of verification capability needed to deter crossing (the thinness of the ice) depends on each side's perception of the risks and rewards of cheating. The attraction of cheating (getting to the other side) would be the belief that it could result in some sort of advantage that would lead to a significant improvement to the country's national security. The consequence of being caught (falling through the ice) would depend on the depth of the water. This would involve international humiliation, the possible abrogation of the treaty resulting in the loss of whatever advantages the treaty had provided, and the potential loss of all other present and future agreements.

ARE TEST BAN TREATIES IN OUR NATIONAL INTEREST?

Test Ban treaties are a seemingly simple approach to arms control, yet their impact is complex and multi-faceted. Determining the advantages of such a treaty depends upon weighing such questions as:

- Is testing necessary to develop future weapon systems? Do we want both the United States and the Soviet Union to develop new weapon systems?
- Is testing necessary to ensure a high degree of reliability of the nuclear stockpile? Do we want the nuclear arsenals of both the United States and Soviet Union to be highly reliable?
- Is continued testing necessary to maintain high levels of technical expertise in the weapons laboratories? Do we want to continue high levels of expertise in both the United States and Soviet weapons laboratories, and if so, for what purposes?
- Is testing necessary to ensure the safety of nuclear devices?
- Could more conservative design practices reduce the need for nuclear testing?
- Would the effects of a test ban impact the United States and the Soviet Union differently?
- Would a decrease in confidence in nuclear weapons' performance increase or decrease the likelihood of nuclear war?
- Would a test ban treaty discourage nuclear proliferation? Could it be extended to cover other nations?
- Would the effects of a treaty be stabilizing or destabilizing?
- Overall, do the advantages outweigh the disadvantages?

Due to the immense uncertainties associated with nuclear conflict, there are few definitive answers to these questions. One's opinion about the answers is largely dependent on one's philosophical position about the role of a nuclear deterrent, and the extent to which arms control can contribute to national security. None of these questions, moreover, can be considered in isolation. Disadvantages in one area must be weighed against advantages in another. Consequently, all aspects of a new treaty must be considered together and their cumulative impact evaluated in terms of a balance with the Soviet Union. Such a net assessment is difficult because even greater uncertainty is introduced when we try to guess how a given

Box 2-A.—First Interest in Test Ban Treaties

Interest in restricting the testing of nuclear weapons began with an incident that occurred over 30 years ago. On February 26, 1954, an experimental thermonuclear device, named *Bravo,* was exploded on the Bikini Atoll in the Pacific Ocean. The explosion was the United States' 46th nuclear explosion. It produced a yield equivalent to 15 million tons of TNT, which was over twice what was expected. The radioactive fallout covered an area larger than anticipated and accidently contaminated an unfortunate Japanese fishing boat named *Lucky Dragon.* When the boat docked at Yaizu Harbor in Japan, twenty-three of the crew had radiation sickness resulting from fallout. The captain of the vessel, Aikichi Kuboyana, died of leukemia in September 1954. In another such accident, radioactive rain caused by a Soviet hydrogen bomb test fell on Japan. These incidents focused worldwide attention on the increased level of nuclear testing and the dangers of radioactive fallout. Soon after, the first proposal for a test ban was put forth.* The 1954 proposal presented by India's Prime Minister Jawaharlal Nehru was described as:

> . . . some sort of what maybe called "standstill agreement" in respect, at least, of these actual explosions, even if the arrangements about the discontinuance of production and stockpiling must await more substantial agreements among those principally concerned.

Since that time over 1,600 nuclear explosions have occurred and at least four more countries (United Kingdom, France, People's Republic of China, and India) have successfully tested nuclear devices.

*See Bmce A. Bolt, *Nuclear Explosions and Earthquakes,* W.H. Freeman and Company, 1976.

treaty would affect the Soviet Union. Finally, the total net assessment of the effects of a treaty on our national security must be weighed against the alternative: no treaty.

The first formal round of negotiations on a comprehensive test ban treaty began on October 31,1958 when the United States, the Soviet Union, and the United Kingdom opened, in Geneva, the Conference on the Discontinuance of Nuclear Weapon Tests. Since then, interest in a test ban treaty has weathered three decades of debate with a level of intensity that has fluctuated with the political climate.[1] During this time, three partial nuclear test limitation treaties were signed. Nuclear explosions compliant with these restrictions are now conducted only underground, at specific test sites, and at yield levels no greater than 150 kilotons (kt). These three treaties are:

1. 1963 Limited Nuclear Test Ban Treaty (LTBT). *Bans nuclear explosions in the atmosphere, outer space, and under water.* This treaty was signed August 5, 1963. Ratification was advised and consented to by the United States Senate on September 24, 1963 and the treaty has been in effect since October 10, 1963.
2. Threshold Test Ban Treaty (TTBT). Re*stricts the testing of underground nuclear* weapons *by the United States and Soviet Union to yields no greater than 150 kt.* This treaty was signed July *3,1974.* It was submitted to the United States Senate for advice and consent to ratification on July 29, 1976 and again on January 13, 1987. It remains unratified, but both nations consider themselves obligated to adhere to it.
3. Peaceful Nuclear Explosions Treaty (PNE). *This treaty is a complement to the TTBT. It restricts individual peaceful nuclear explosions by the United States and Soviet Union to yields no greater than 150 kt, and aggregate yields to no greater than 1,500 kt.* This treaty was signed May 28, 1976. It was submitted to the United States Senate for advice and consent to ratification on July 29, 1976 and again on January 13, 1987. It remains unratified, but both nations consider themselves obligated to adhere to it.

Although these treaties have fallen far short of banning nuclear testing, they have had important environmental and arms control impacts. Since 1963, no signatory country compliant with these treaties has tested nuclear weapons in the atmosphere, in outer space, or under water, thus eliminating a major environmental hazard. And from an arms control perspective, testing of warheads over 150 kt has been prohibited since 1974.

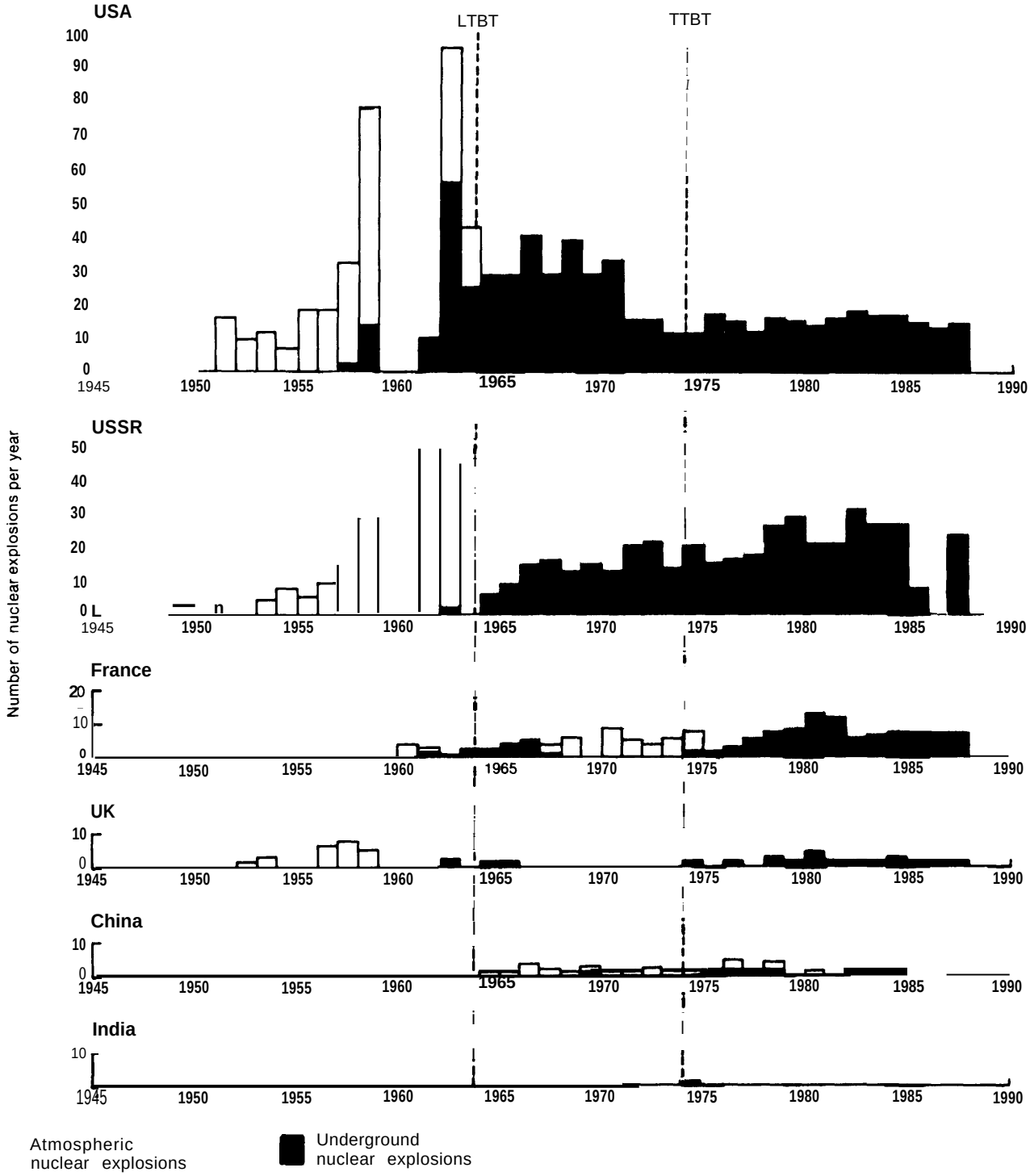
While these treaties have had important positive impacts, figure 2-1 illustrates that, in fact, they have not resulted in any decline in the amount of testing. The development of new types of warheads and bombs has not been limited by restricting testing, and so advocates of test ban treaties continue to push for more restrictive agreements.

A Comprehensive Test Ban Treaty was the declared goal of the past six U.S. Administrations, but remained elusive. The Reagan Administration, however, has viewed limitations on nuclear testing as not in the national security interests of the United States, both at present and in the foreseeable future. The stated policy of the Reagan Administration is as follows:

> A Comprehensive Test Ban (CTB) remains a long-term objective of the United States. As long as the United States and our friends and allies must rely upon nuclear weapons to deter aggression, however, some nuclear testing will continue to be required. We believe such a ban must be viewed in the context of a time when we do not need to depend on nuclear deterrence to ensure international security and stability and when we have achieved broad, deep, and verifiable arms reductions, substantially improved verification capabil-

[1] For an overview of the history of test ban negotiations, the reader is referred to G. Allen Greb, "Comprehensive Test Ban Negotiations 1958-1986: An Overview, " in *Nuclear Weapon Tests: Prohibition or Limitation?, edited* by Jozef Goldblat and David Cox, SIPRI, CIIPS, Oxford University Press, London, 1987.

Figure 2-1.— Nuclear Testing, July 16, 1945-December 31, 1987



SOURCE: Data from the Swedish Defense Research Institute,

ities, expanded confidence building measures, and greater balance in conventional forces.[2]

Despite this declared United States position, public interest in a test ban remains strong. Recently, a series of events have heightened worldwide interest in test ban treaties and a number of proposals have been brought before Congress to further limit the testing of nuclear weapons (see box 2-B). To evaluate these proposals requires an understanding of the desired and available levels of verification needed to monitor compliance.

[2]U.S. Department of State, Bureau of Public Affairs, "U.S. Policy Regarding Limitations on Nuclear Testing," Special Report No. 150, Washington, DC, August 1986,

Box 2-B.—Recent Interest in Nuclear Test Limitations

August 6, 1985 to February 26, 1987: The Soviet Union observes a 19-month unilateral test moratorium. Upon ending its moratorium, the Soviet Union declares that testing would be stopped again as soon as a U.S. testing halt was announced.

February 26, 1986: House Joint Resolution 3 (with 207 cosponsors) passes the House with a vote of 268 to 148, requesting President Reagan to resume negotiations with the Soviet Union towards a comprehensive test ban treaty and to submit to the Senate for ratification the Threshold Test Ban Treaty (TTBT) and the Peaceful Nuclear Explosions Treaty (PNE). The wording of the House Resolution is nearly identical to a similar proposal which previously passed the Senate by a vote of 77 to 22 as an amendment to the 1985 Department of Defense (DoD) Authorization Bill.

May 28, 1986: The Natural Resources Defense Council, a private environmental group, signs an agreement with the Soviet Academy of Sciences for the establishment of three independent seismic monitoring stations near the principal nuclear test sites of each country. The agreement specifies that the stations are to be jointly manned by American and Soviet scientists and the data is to be made openly available.

Summer 1986: UN Conference on Disarmament agrees to a global exchange by satellite of sophisticated seismic data.

August 7, 1986: The Five Continent Peace Initiative formed by the leaders of six nonaligned countries (India, Sweden, Argentina, Greece, Mexico, and Tanzania) urges a fully verifiable suspension of nuclear testing and offers assistance in monitoring the ban.

August 8, 1986: House of Representatives votes *234* to 155 in favor of an amendment to the Defense Authorization Bill that would delete funding in calendar year 1987 for all nuclear tests with yields larger than 1 kt, provided that the Soviet Union does not test above 1 kt and that the Soviet Union accepts a U.S. monitoring program. The House Amendment is dropped prior to the Reykjavik summit when the Administration agrees to submit the TTBT and PNET to the Senate for advice and consent.

May 19, 1987: House of Representatives votes 234 to 187 in favor of an amendment to the fiscal year 1988 DoD Authorization Bill to delete funding for all nuclear tests with yields larger than 1 kt during fiscal year 1988 provided that the Soviet Union does the same and, if reciprocal (in-country) monitoring programs are agreed on and implemented.

July 1987: At the expert talks on nuclear testing, Soviets propose calibration of test sites to reduce the uncertainty in yield estimates. The proposal invites U.S. scientists to the Soviet nuclear testing site to measure Soviet test yields using both the CORRTEX system and seismic methods. In return, Soviet scientists would measure a U.S. test at the Nevada test site using both methods.

November 9, 1987: Formal opening of negotiations in Geneva on nuclear test limitations.

January 1988: Teams of U.S. and Soviet scientists visit each other's test site to prepare for joint calibration experiments to reduce uncertainty in yield estimation.

EVALUATING THE CAPABILITY OF A VERIFICATION NETWORK

Making a decision on whether verification is adequate requires an understanding of the capability of the verification system, the significance of the potential violation, and a decision as to what is an acceptable level of risk. Developing the basis for the decision is difficult because it involves two different communities: those who can assess the system's capabilities (a technical question) generally do not form the same community as those who are officially responsible for assessing the overall risks and benefits (a policy question).

For policymakers to weigh the benefits of the treaty against the risks posed by the possibility of unilateral noncompliance, a clear understanding of the capabilities of the monitoring system *is* necessary. In the frozen river example, this understanding would result from measurements of how thin the ice is and a technical interpretation of how much weight the ice can bear. The decision as to what constitutes an acceptable level of risk is a policy decision because it is based on an assessment of the overall benefits of the treaty weighed against the risk. In the frozen river example, this assessment would represent the decision as to how thin the ice would need to be to deter crossing, how deep the river is, and how significant a crossing would be.

The burden on the policy-making community, therefore, is to understand technical descriptions of the verification system's capability and incorporate this knowledge into their risk-benefit decision. As we shall see, the difficulty is that monitoring capabilities are not certain, but rather they can only be described in probabilistic terms. For example: What are the chances that a clandestine nuclear test above a certain yield could go undetected? What are the chances that a detected seismic event of a certain magnitude could have been a nuclear explosion rather than an earthquake? If an underground nuclear explosion is recorded, how certain can we be that the yield of the explosion was below a specific threshold? The answers to these questions can be obscured by the manner in which they are portrayed. In particular:

- differences between verification systems can be made to look superficially either large or small,
- opportunities for Soviet cheating can be misrepresented, and
- the decision of what defines adequate verification can be made through an arbitrary process.

The next three sections illustrate the issues that arise in assessing a verification capability —and the misrepresentations that are possible–by considering a question that aroused much Congressional interest in early 1987: What is our ability to measure seismically the yields of Soviet explosions near the 150 kiloton limit of the Threshold Test Ban Treaty? The first section presents the statistical representations that are used to describe yield estimation. This includes the meaning of uncertainty and confidence levels, along with a comparative discussion that enables the reader to understand what changes in the uncertainty represent. The next section examines how these uncertainties translate into opportunities for Soviet cheating. And finally, the third section illustrates how the policy decision of what constitutes adequate yield estimation capability has changed in apparent response to variations in the attractiveness of particular monitoring systems.

Uncertainty and Confidence Levels

In determining the verifiability of the 1974 Threshold Test Ban Treaty, policymakers wanted to know the capabilities of a seismic monitoring system for estimating whether Soviet tests are within the treaty's limits. The description of such capabilities is accomplished through the use of statistics. While the statistical calculations are relatively straightforward, difficulties arise in correctly appreciating what the numbers mean. To illustrate how

such presentations can be misleading, we will first use an example from a common and comparatively well-understood event:[3]

> At the end of the 1971 baseball season, the San Francisco Giants were playing the Los Angeles Dodgers in a televised game. In the first inning, Willie Mays, approaching the end of his illustrious career, hit a home run. Now, one expects that hitting a home run in the first inning should be a rather unusual occurrence because the pitcher is at his strongest and the batter has not had time to get used to the pitcher. In any case, Willie Mays hit a home run and it triggered what every baseball fan would recognize as a typical baseball statistician's response. The calculations were made and it was discovered that, of the 646 home runs Mays had hit, 122 of them had been hit in the first inning: 19 percent! In the most unlikely one-ninth of the innings, Willie Mays had hit nearly one-fifth of his home runs. This realization captured the interest of the reporting community and was discussed extensively in the media. In response to the publicity, the Giants' publicity director explained it by saying that". . . Willie was always surprising pitchers in the first inning by going for the long ball before they were really warmed up. " The power of statistical analysis was able to draw out the hidden truths about Willie Mays' performance.

Although the data and calculations were correct, the interpretation could not have been more wrong. Throughout Mays' career, he had almost always batted third in the Giants' lineup (occasionally, he batted fourth). That meant he almost always batted in the first inning. Because he averaged about four at-bats per game, approximately one-quarter of his at-bats came in the first inning. Therefore, he only managed to hit 19 percent of his home runs during the first inning which comprised 25 percent of the time that he was at bat. Of the millions of people who must have heard and read about the item, not one pointed out the misinterpretation of the statistic. This included not just casual observers, but also experienced professionals who spend their careers interpreting just that kind of information.

The point here is that the interpretation of numbers is tricky and statistical presentations can often be misleading. The real challenge is not in calculating the numbers, but in correctly interpreting what different numbers mean. In an area with as many technical considerations and political influences as arms control verification, one has to be particularly careful that different numbers represent truly significant differences and not just arbitrary distinctions.

As with every real-world measurement, estimating the size of a nuclear explosion results in variation, or scatter, among the estimates. The use of different instruments at different locations, interpretations of the measurements by different people in slightly different ways, and unknown variations in signals being observed result in slightly different estimates. Similarly, if one were to measure the daily temperature outside using a number of thermometers located in several areas, there would be slight differences in the temperature depending on the particular thermometer, its location (surrounded by buildings and streets, or in a park), how each scale was read, etc.

In seismology, errors come from the instrumentation, from the interpretation of the data, from our incomplete knowledge about how well an explosion transmits its energy into seismic waves (the coupling), and from our limited understanding of how efficiently seismic waves travel along specific paths (the path bias). Some of these errors are random-they vary unpredictably from one measurement to the next. Other errors are not random, but are systematic and are the same from one measurement to the next.

In our example of measuring temperature, a systematic error would be introduced if each reading were made using an improper zero on the thermometer. With such a systematic error, the measurements would continue to be distributed randomly, but the distribution would be shifted by the difference between the true and incorrect zero. The distance from the incorrect value to the actual value would represent the size of the systematic error. In seismology, an example of a systematic error re-

[3] This example is paraphrased from David L. Goodstein, *States of Matter* (Englewood Cliffs, NJ: PrenticeHall) 1975.

suits from the failure to allow correctly for the difference in seismic transmission between the United States and Soviet test sites. The bias term added to the calculation corrects for this effect, although there still remains an uncertainty associated with the bias.

The distinction between random and systematic errors, however, is not a clear boundary. In many cases, random errors turn out to be systematic errors once the reason for the error is understood. However, if the systematic errors are not understood, or if there are lots of systematic errors all operating in different ways, then the systematic errors are often approximated as random error. In such cases, the random uncertainties are inflated to encompass the uncertainties in estimating the systematic effect. In monitoring the yields of Soviet testing near 150 kt, most of the uncertainty is associated with estimation of systematic error because the test site has not been calibrated. In describing the capability to measure Soviet tests, the estimate of the random error has been inflated to account for the uncertainties in the systematic error.

We will see in the next section that while systematic errors might be exploited if they happen to be to one country's advantage, random errors do not provide opportunities for cheating. Furthermore, the uncertainty in the estimates of the systematic errors can often be *significantly* reduced by negotiating into the treaty such provisions as the calibration of each test site with explosions of known yield. For this reason, calibration is important and should certainly be part of any future agreement. But, before we discuss how random and systematic errors affect monitoring, the method of statistically describing the capabilities needs to be explained. For this, it is useful to return to our temperature example.

While collecting measurements of the daily temperature by using many thermometers in many locations, we would find that some of the measurements were high and some were low, but most of them were somewhere in the middle. If all of the measurements were plotted, they would cluster around one number with roughly equal scatter distributed to either side. It would be most likely that the best actual value for the daily temperature would be near the central number and it would be increasingly less likely that the actual value would be off towards either end of the scatter distribution.

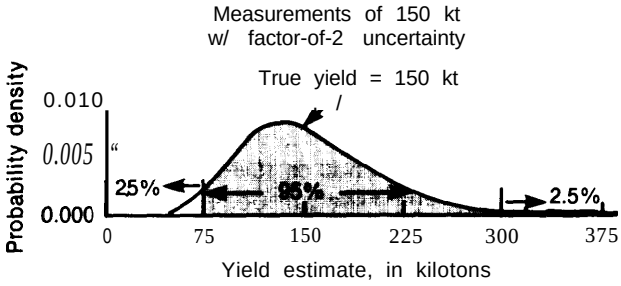
In seismology, that central number, measured using several techniques from many seismometers located in different locations, is referred to as the "central value yield. " No matter how accurately we could measure the yields of explosions known to be 150 kt, we would still expect—due to the normal random scatter of measurements and presuming there were no systematic error—that roughly half the explosions would be recorded as being below 150 kt and roughly half would be recorded as being above 150 kt. The width of the distribution above and below depends on the capability of the measuring system and can be described using a "factor of uncertainty. "

Unless otherwise stated, the factor of uncertainty for a given measurement is defined as that number which, when multiplied by or divided into an observed yield, bounds the range which has a 95 percent chance of including the actual (but unknown) value of the yield. There is only a 5 percent chance that the measurement would be off by more than this factor. For example, a "factor of 2" uncertainty would mean that the measured central value yield, when multiplied and divided by two, would define a range within which the true yield exists 95 percent of the time.

Naturally, the more confident one wants to be that the true value lies within a given range, the larger that range will have to be. The 95 percent range is used by convention, but there is no real reason why this should be the confidence level of choice for comparing monitoring systems. Using a different confidence level than 95 percent to define the factor of uncertainty would cause the factor to have a different value.

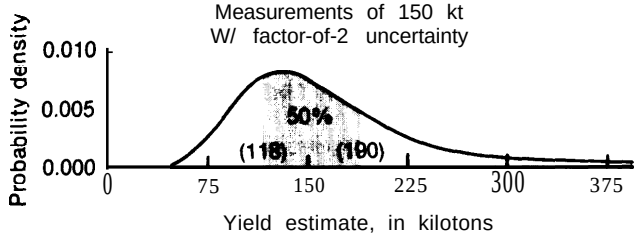
As an example, imagine that all Soviet explosions are detonated with an actual yield of 150 kt and that these explosions are measured

Figure 2-2.–Measurements of 150 kt With Factor-of-2 Uncertainty



Measurements of a 150 kt explosion using a factor-of-2 uncertainty would be expected to have this distribution. The probability that the actual yield lies in a particular range is given by the area under the curve. Ninety-five percent of the area lies between 75 and 300 kt.

Figure 2-3.—Fifty-Percent Confidence Level for Measurements of 150 kt With Factor-of-2 Uncertainty



Same distribution as figure 2-2 with 50 percent of the area marked. Over 50 percent of the measurements would be expected to fail within 118 and 190 kt.

by a monitoring system described as having a factor of 2 uncertainty. It would then follow that 95 percent of the yield estimates will be between 75 and 300 kt. Therefore, if the Soviets conduct 100 tests all at 150 kt, we would expect that about 5 of them would be measured with central yields either above 300 kt or below 75 kt. This is graphically illustrated in figure 2-2, where the probability that the actual yield lies in a particular range is given by the area under the curve over that range.

Although the range from 75 to 300 contains 95 percent of the measurements, we can see that for the distribution assumed (the normal distribution) most of the measurements are, in fact, much closer to the actual yield. For example, figure 2-3 showing the 50 percent confidence level for the same distribution illustrates that over half of the measurements will fall between 118 and 190 kt.

On the other hand, if one wanted to specify a range in which 99 percent of the measurements fall, the range would have to be extended to 60 to 372 kt. In real yield estimation, however, there is no meaningful distinction between 95 percent confidence and 99 percent confidence. The normal distribution is used as a convenience that roughly represents reality near the center of the distribution. The tails of the distribution are almost certainly not close approximations of reality. The general point can still be made: namely that, it takes ever greater increases in range for slight improvements in the confidence level. Table 2-1 illustrates this for the case of a normal distribution by showing how the yield range varies for given factors of uncertainty at various confidence levels.[4]

From the table, it is clear that a quoted range of values is highly dependent not only on the factor of uncertainty of the monitoring system, but also on the chosen confidence level. For example, the same quoted uncertainty range

[4]Table 2-1 can be read as follows: "If an explosion is measured at 150 kt using a system with a factor of [1.5] uncertainty, the [70 percent] percent confidence ranges from [121 to 186] kt."

Table 2.1.–Confidence Internals for an Explosion Measured at 150 kt

CONFIDENCE LEVEL	UNCERTAINTY		
	Factor of 2	*Factor of 1.5*	*Factor of 1.3*
99%	*61-372*	*88-255*	*106-211*
95%	75-300	100-225	115-195
90%	84-269	106-212	120-187
80%	96-236	116-195	126-179
70%	104-216	121-186	130-173
50%	118-191	130-173	138-164

SOURCE: Office of Technology Assessment, 1988.

that can be achieved using a factor of 1.3 monitoring system at the 95 percent confidence level can be achieved using a factor of 1.5 monitoring system at the 80 percent confidence level, or even using a factor of 2 monitoring system at the 50 percent confidence level.

In selecting an appropriate confidence level to use for quoting uncertainty, the purpose of the comparison must also be recognized. In the case of monitoring compliance with a threshold test ban treaty, one is concerned that intentional cheating could be missed or that unacceptable false alarms could occur. It must be remembered, therefore, that the uncertainty describes the likelihood that the actual value will not fall either above or below the range. For example, the 95 percent confidence level means that given repeated trials there is only a 5 percent chance the true yield was *either* above or below the range of the measured yield. In other words, there is a 2.5 percent chance that it could have been above the range and a 2.5 percent chance that it could have been below the range. The 2.5 percent below the range, however, is not a concern if it is below the threshold. The concern is only the 2.5 percent chance that it had a yield above the threshold. For the purposes of monitoring a threshold, this would only be the half of the 5 percent above the threshold, or in other words 2.5 percent. Consequently, the 95 percent confidence level really corresponds to a 97.5 percent confidence level for monitoring violations of a threshold. Following the same argument, the 50 percent confidence level really corresponds to a threshold monitoring at the 75 percent confidence level, and so on.

From the previous discussions, it is obvious that while it maybe convenient to chose a particular confidence level to compare monitoring systems (such as the 95 percent confidence level), it should only be done with great caution. In particular, it should be kept in mind that:

- the choice of any particular confidence level is arbitrary, in the sense of being only a convenience to allow comparison of the accuracies of different yield estimation methods;
- as seen from table 1-1, differences can be made to look small or large depending on which particular confidence level is chosen; and
- although very high confidence levels have large ranges of uncertainty, it is with decreasing likelihood that the actual value will be at the extremes of those ranges of uncertainty.

These considerations are important to ensure that common mistakes are not made, namely that:

- the range of uncertainty is not equated with a range in which cheating can occur, and
- the range of uncertainty chosen for comparative reasons does not evolve into the range of uncertainty that is used to determine what constitutes adequate verification.

The Relationship Between Uncertainty and Cheating Opportunities

The main reason for designing monitoring systems with low uncertainty is to reduce the opportunities for cheating. The relationship between uncertainty and opportunities for cheating, however, is not always straightforward. Even if the uncertainty of a particular monitoring system is large, this does not mean that the opportunities for cheating are correspondingly large.

As mentioned before, the uncertainty is created by two types of error: random and systematic. Although we try to estimate the systematic error (such as path bias) as accurately as possible, there is a chance that our estimates could be slightly too high or too low. For example, if our estimate of the bias[5] is too low, we would overestimate the yields of Soviet explosions, and Soviet testing near 150 kt would appear as a series of tests distributed around a yield value that was above 150 kt. If our estimate of the bias were too high, Soviet testing

[5]See chapter 7, "Estirnating the Yields of Nuclear Explosions," for an explanation of bias.

near *150* kt would appear as a series of tests distributed around a yield below 150 kt. The case where we systematically underestimate Soviet yields and they presume this underestimation is occurring is the only case that provides opportunity for unrecognized cheating. If this were happening, it could happen only to the extent that the systematic effect has been underestimated.

As chapter 7 discusses, the systematic part of the uncertainty can be significantly reduced by restricting testing to specific calibrated test sites. If such calibration were an integral part of any future treaty, the concern over systematic errors of this kind should be minimal. The majority of the error that would remain is random. A country considering violating a treaty could not take advantage of the random error because it would be unable to predict how the random error would act.

In early 1987, both the Senate Foreign Relations Committee and the Senate Armed Services Committee held hearings on verification capabilities in their consideration of advice and consent to ratification of the 1974 Threshold Test Ban Treaty and the 1976 Peaceful Nuclear Explosions Treaty. Members of these Committees wanted to know whether seismic methods could adequately measure the size of Soviet underground nuclear tests or whether more intrusive methods were required. The testimony was often confusing due to the various means of representing statistical uncertainties. For the time being, we will analyze only the use of statistics and take as given the underlying information. However, that acceptance is also controversial and is discussed separately in the chapter on yield estimation (chapter 7).

The Department of Defense presented the capabilities of seismic monitoring to the Senate Committee on Foreign Relations on January 13, 1987 and the Senate Committee on Armed Services on February 26, 1987 in the following manner:[6]

> The seismic methods that we currently must rely onto estimate yields of Soviet nuclear detonations are assessed to have about a factor-of-two uncertainty for nuclear tests, and an even greater uncertainty level for Soviet peaceful nuclear explosions.

This uncertainty was then explained as follows:

> This uncertainty factor means, for example, that a Soviet test for which we estimate a yield of 150 kilotons may have, with 95 percent probability, an actual yield as high as 300 kilotons-twice the legal limit-or as low as *75* kilotons.[7]

These statements are misleading in that they create the impression that there is a high probability, in fact, almost a certainty, that the Soviets could test at twice the treaty's limit but we would measure the explosions as being within the 150 kt limit. They imply that a factor of 2 uncertainty means that there is a high probability that an explosion measured at 150 kt could, in actuality, have been 300 kt. Yet as we have seen in the discussion of uncertainty, given a factor of two uncertainty, the likelihood of an explosion with a yield of 300 kt actually being measured (with 95 percent probability) as 150 kt or below is less than 1 chance in 40.

The chances decrease even further if more than a single explosion is attempted. For example, the chance of two explosions at 300 kt both being recorded as 150 kt or less is about 1 in 1,600; and the chance of three explosions at 300 kt or greater being recorded as 150 kt or less would be roughly 1 in 64,000. Thus, it is highly unlikely that explosions could be repeatedly conducted at 300 kt and systematically recorded as being 150 kt or less.

So far we have been looking only at the likelihood that a test will appear as 150 kt or less.

[6] Testimony of Hon. Robert B. Barker, Assistant to the Secretary of Defense (Atomic Energy) and leader of formal negotiations on Nuclear Test Limitations.

[7] This statement is nearly identical to the wording in the U.S. Department of State, "Verifying Nuclear Test Limitations: Possible U.S.-Soviet Cooperation," Special Report No. 152, Aug. 14, 1986, which states "A factor of two uncertainty means, for example, that a Soviet test for which we derive a 'central yield' value of 150 kt may have, with a 95 percent probability, a yield as high as 300 kt or as low as 75 kt."

From a practical point of view, it must be recognized that the test would not need to look like 150 kt or less; it would only have to appear as though it were within the error of a 150 kt measurement in order to avoid credible assertions of non-compliance. Some could misinterpret this as meaning that a test well above the threshold might have enough uncertainty associated with it so that its estimate might appear to be within the expected uncertainty of a test at 150 kt. They might then conclude that the opportunity to test well above the threshold cannot be denied to the Soviets.

Such a one-sided assessment of the uncertainty is extremely misleading because it assumes all of the errors are systematic and can be manipulated to the evader's advantage. A country considering violating the threshold would also have to consider that even if part of the systematic uncertainty could be controlled by the evader, the random part of the uncertainty could just as likely work to its disadvantage. This point was briefly recognized in the following exchange during a Senate Committee on Foreign Relations hearing:

> "
> . . . knowing these probabilities, if you really started to cheat, as a matter of fact you would take the risk of being out at the far tail. That would really show up fast. If you set out to do a 300 kt, you could show up on our seismographs as 450, right?"
>
> Senator Daniel P. Moynihan

> "The problem is, if you fired an explosion at 300 or 350 . . . it could very well look 450 or 500 and the evader has to take that into consideration in his judgment. "
>
> Dr. Milo Nordyke,
> Director of Verification
> Lawrence Livermore
> National Laboratory

Also, this analysis assumes that only one method of yield estimation will be used. Other methods of yield estimation are also available and their errors have been shown to be only partially correlated. The evader would have to take into account that even if the uncertainty is known and can be manipulated for one method of yield estimation, other methods might not behave in the same manner. Such considerations would severely diminish the appeal of any such opportunity.

In conclusion, it can be seen that although the statistical descriptions of the capabilities of various methods of yield estimation have been debated extensively, the differences they represent are often insignificant. There is both systematic and random uncertainty in the measurements of Soviet yields. The systematic error would provide only a limited opportunity for cheating, and then only if it was in the advantage of the cheater. Even in such a case, only the portion of the error that is systematic can be exploited for cheating. Furthermore, much of the systematic error would be removed through such treaty provisions as calibrating the test site. Once the systematic error had been nearly eliminated, the remaining uncertainty would be random. The random uncertainty does not provide opportunity for cheating. In fact, if a country were considering undertaking a testing series above the threshold, it would have to realize that the random uncertainty would work against it. With each additional test, there would be a lesser chance that it would be recorded within the limit and a greater chance that at least one of the tests would appear to be unambiguously outside the limit.

What Constitutes Adequate Verification?

After the accuracies and uncertainties of various verification systems have been understood, a decision must be made as to what constitutes an acceptable level of uncertainty. In 1974, when the Threshold Test Ban Treaty was first negotiated, a factor of 2 uncertainty was considered to be the capability of seismic methods. At that time, a factor of 2 uncertainty was also determined to constitute adequate verification.[8] Presently, the level of accuracy claimed

[8] Originally, the factor of 2 uncertainty was established for the 90 percent confidence level, whereas today it refers to the 95 percent confidence level. It should be noted that a factor of 2 at the 90 percent confidence level corresponds to about a factor of 2.5 at the 95 percent confidence level. Thus the accepted level of uncertainty in 1974 was really about a factor of 2.5 using the present confidence level. The insistence on a

(continued on next page)

for the on-site CORRTEX method is a factor of 1.3.[9] This level of 1.3 has subsequently been defined as the new acceptable level of uncertainty, although many believe it was defined as such only because it corresponds to the capabilities of this newly proposed system.

It appears that the determination of adequate compliance is a subjective process that has been influenced by the capabilities of specific monitoring systems. A decision as to what constitutes adequate verification should not be determined by the political attractiveness of any particular monitoring system, but rather it should represent a fair assessment of the protection required against non-compliance. In the frozen river analogy, this would be a fair assessment of how thin the ice must be to deter someone from crossing. Monitoring capability certainly influences our decision as to whether a treaty is worthwhile, but it should not influence the standards we set to make that decision. Also, the capability of a monitoring system is just one aspect to be considered, along with other important issues such as negotiability and intrusiveness.

What constitutes adequate verification may also vary for different treaty threshold levels. For example, a factor of 2 uncertainty for monitoring a 100 kt threshold would mean that 95 percent of the measurements at the threshold limit would be expected to fall within 50 and 200 kt (a total range of 150 kt), while a factor of 2 uncertainty for monitoring a 1 kt threshold would mean that 95 percent of the measurements at the threshold limit would be expected to fall within% and 2 kt (a total range of 1.5 kt). A range of uncertainty of 1.5 kt may not provide the same opportunities or incentives for cheating as a range of uncertainty of 150 kt. Consequently, at lower treaty thresholds, the significance of a given yield uncertainty will almost certainly diminish.

{continued from previous page)

higher confidence level occurred simply because it was more convenient to use the 95 percent confidence level which corresponds to 2 standard deviations.

[9]See appendix, Hydrodynamic Methods of Yield Estimation.

THE QUESTION OF DETERMINING COMPLIANCE

In addition to understanding the accuracy and uncertainty of the verification system, and deciding on an acceptable level of uncertainty, a decision will also have to be made as to what would constitute compliance and non-compliance. Violations of the treaty must be distinguished from errors in the measurements (both systematic and random) and errors in the test. This is of particular concern in light of findings by the administration that:

> Soviet nuclear testing activities for a number of tests constitute a likely violation of legal obligations under the Threshold Test Ban Treaty ."[10]

To examine the context in which this must be viewed, we can once again return to the frozen river analogy and imagine a situation where we come by and see marks on the ice. We must then determine whether the marks indicate that someone successfully crossed the ice. This could be misleading because all that we are doing is looking in isolation at the probability that a certain mark could have been made by someone crossing the ice. Thus the likelihood that a mark was made by a person becomes the likelihood that someone crossed the ice. This, however, is only part of the issue. If we knew for example that the ice were so thin that there was only a 1 in 10 chance it could have been successfully crossed, that the water was deep, and that there was no reason to get to the other side, these factors might weigh in our determination of whether a mark was mane-made or not. (Why would a person have taken such risks to gain no value?) On the other hand, if we knew the ice were thick and could be crossed with high confidence, that the water was shallow, and that real value was to be obtained by crossing, then we might

[10]"" The President's Unclassified Report on Soviet Noncompliance with Arms Control Agreements," transmitted to the Congress Mar. 10, 1987.

make a different judgment as to whether the marks were man-made because there would really be understandable motivation. Thus the question of compliance is also dependent on a judgment reflecting one's perception of the advantages that could be obtained through a violation.

In the case of test ban treaties, there are also "gray areas" due to the associated error of the measurements. For example, it must be assumed that a country will test up to the limit of the treaty, and therefore, some of the estimates would be expected to fall above 150 kt simply due to random error.[11]

Assuming that the errors are known and that apparent violations of the treaty due to such errors are recognized, there may also be other violations that cause concern but do not negate the benefits of the treaty. These include accidental violations, technical violations, and violations of the "spirit" of the treaty.

Accidental violations are violations of the treaty that may occur unintentionally due to the inexact nature of a nuclear explosion. It is possible that the explosion of a device with a yield that was intended to be within the limit of the treaty would produce an unexpectedly higher yield instead. This possibility was recognized during negotiations of the TTBT. The transmittal documents which accompanied the TTBT and the PNE Treaty when they were submitted to the Senate for advice and consent to ratification on July 29, 1976 included the following understanding recognized by both the United States and Soviet Union:

> Both Parties will make every effort to comply fully with all the provisions of the TTB Treaty. However, there are technical uncertainties associated with predicting the precise yields of nuclear weapon tests. These uncertainties may result in slight, unintended breaches of the 150 kt threshold. Therefore, the two sides have discussed the problem and agreed that: (1) One or two slight, unintended breaches per year would not be considered a violation of the Treaty; *(2)* such breaches would be a cause for concern, however, and, at the request of either Party, would be subject for consultations.

Technical violations are violations of the treaty that do not result in any sort of strategic advantage. An example would be a technical violation of the 1963 Limited Test Ban Treaty (LTBT) which prohibits any explosion that:

> . . . causes radioactive debris to be present outside the territorial limits of the State under whose jurisdiction or control such explosion is conducted.[12]

This prohibition includes the venting of radioactive debris from underground explosions. Both the United States and the Soviet Union have accused each other of releasing radioactive material across borders and of violations of the 1963 LTBT. These violations are "technical" if the treaty is viewed as an arms control measure. However, they are material violations if the treaty is viewed as an environmental protection measure.

Violations of the "spirit" of the treaty are also of concern. These include, for example, actions which are contrary to the treaty's preamble. The treaty's preamble declares the intentions and provides a context for the treaty. Such declarations, however, are nonbinding.

Another area concerns treaties that have been signed but never ratified. Both the 1974 TTBT and the 1976 PNE Treaty remain unratified, although they were signed over 10 years ago. Because neither the United States nor the Soviet Union have indicated an intention not to ratify the treaties, both parties are obligated under international law (Article 18, the 1969 Vienna Convention on the Law of Treaties) to refrain from acts which would defeat their objectives and purposes.

All of these types of violations contribute to the gray area of compliance versus noncompliance and illustrate why determining com-

[11] For example, if the Soviet Union tested 20 devices at 150 kt and we estimated the yields using a system that was described as having a factor of 2 uncertainty, the probability of measuring at least one of them as being 225 kt or greater is 92 percent.

[12] Article I,b.

pliance is a political as well as a technical decision. In the case of monitoring underground nuclear tests, the actual measured yield that would constitute clear evidence of a violation would always be higher than the yield limit of the treaty. Perhaps an analogy for uncertainties in yield estimates and Soviet compliance under the TTBT is in monitoring a speed limit of 55 mph. Under the present 150 kt limit, an observed yield of 160 kt is like comparing 58.7 mph to 55 mph. The police do not give tickets when their radar shows a speed of 58.7 mph because most speedometers are not that accurate or well calibrated, and because curves and other factors can lead to small uncertainties in radar estimates of speed. Similarly, although a 160 kt measurement maybe regarded by some as a legal lack of compliance, such a number can well arise from uncertainties in seismic estimates. At radar measurements over 65 mph the police do not question that the 55 mph limit has been exceeded, and the speeder gets a ticket. With this standard, it would take a calculated yield of about 180 to 190 kt to conclude that a violation had likely taken place. As mentioned before, however, this argument does not mean that the Soviets could test up to 180 to 190 with confidence, because the uncertainty could just as likely work against them. A 180 to 190 kt test might produce an observed yield well over 200 kt just as likely as it might produce a yield within the expected error range of a treaty compliant test.

It must be recognized, however, that the calculated yield for declaring a treaty violation will always be higher than the limit of the treaty. Consequently, one or two small breaches of the treaty could occur within the expected uncertainty of the measurements. A country intent on cheating might try to take advantage of this by risking one or two tests within the limits of the uncertainty range. Even if detected, a rare violation slightly above the permitted threshold could be explained away as an accidental violation due to an incorrect prediction of the precise yield of the nuclear test. This should be kept in mind when choosing a threshold so that small violations of the limit (whether apparent or real) do not fall in a range that is perceived to be particularly sensitive.

What is done about violations is an additional problem. In domestic law there are various kinds of violations. Traffic tickets, misdemeanors, felonies, capital crime–all are different levels. Similarly, in monitoring compliance, there are some things that amount to traffic tickets and some that amount to felonies. We must decide in which cases violations or noncompliance are at the heart of a treaty and in which cases they area marginal problem.