

**WHY A GOOD VERIFICATION SYSTEM  
CAN GIVE AMBIGUOUS EVIDENCE**

Barry O'Neill

York Centre for International and Strategic Studies, and  
Departments of Political Science and Economics  
York University

YCISS Working Paper #8  
April 1991

Centre for International and Strategic Studies  
York University  
4700 Keele Street, Downsview, Ontario, M3J 1P3

## ABSTRACT

The recent debate on ambiguity provides lessons for future arms control verification. American officials sought verification systems that returned unambiguous evidence about Soviet compliance, but a simple model of verification suggests that a verification scheme can be more ambiguous yet better. It may be more effective in deterring violations and avoiding false alarms. The reason is as follows: should the inspecting party come upon suspicious evidence, it will, on the one hand, have a reason to trust that evidence more, as it was returned by a more reliable verification system. On the other hand it will have a reason to be more sceptical that the other is violating since the other would probably not dare to cheat in the face of the improved verification technology. In some situations a reasonable inspector will regard the second factor as weightier than the first, and give lower credence to the evidence.

Ambiguity in verification is a tricky notion and misunderstandings about it arise from two sources: from the vocabulary of verification, which suggests that one dichotomously "detects" or does "not detect" violation, when in fact evidence comes in gradations, and from the human tendency not to look at the situation from the other's viewpoint. The model uses game theory's logic to represent the strategic aspects of the situation, and has a mathematical feature different from past models, the notion of continuous degrees of evidence, to give a proper account of ambiguity. It also clarifies past technical studies of verification by locating them within the model's structure.

*Acknowledgement:* I am grateful to John Brown, Neta Crawford, William Durch and James Schear for their suggestions, and for support from an SSRC/MacArthur Fellowship in International Security. This paper was prepared while I was a visiting scholar at the Center for International Studies, Massachusetts Institute of Technology.

Submitted to *International Studies Quarterly*

## INTRODUCTION

In the verification context, *ambiguity* means one treaty partner's uncertainty about the other's compliance.<sup>1</sup> During the 1980s Reagan Administration officials and advisers asserted that a verification system that leaves ambiguity will undermine arms control. In 1981, Eugene Rostow, then Director of the Arms Control and Disarmament Agency, explained the Reagan Administration's new emphasis, "We will seek verification provisions which . . . limit the likelihood of ambiguous situations developing. Ambiguity can never be eliminated but we shall do our best to keep it to a minimum. (Rostow, 1981; quoted by Schear, 1982)." The report *Discriminate Deterrence*, from the President's Commission on Integrated Long-Term Strategy (1987), called for agreements "capable of yielding evidence, in the event of a major violation, that will be sufficiently unambiguous to enable the U.S. Government to decide on an adequate response." The context of arms control has changed since then, but ambiguity is still an unsettled problem, likely to reappear. The political climate is warmer now, but may change. Members of the Reagan Administration group that objected to "ambiguity," whom Schear (1989) termed the verification "revisionists," hold some key positions in Bush's arms control bureaucracy, and agreements may come under general attack.

It makes sense to clarify this past debate. Specialists in verification technology see some hard obstacles for future treaties, which may aggravate the problem of ambiguity. Past treaties have tended to focus on weapons that were easiest to verify, leaving, for example, sea-launched cruise missiles for later. The type of problems that might arise may be foreshadowed by some difficulties with the INF Treaty. In one instance, U.S. sources worried that Soviet declared inventory did not match CIA records of production (Geertz, 1990). In another case, Soviet missiles were apparently sighted outside their legal deployment area (1990). These verification ambiguities are more likely as verification becomes more thorough and extends over more weapons.

Calling for effectively no ambiguity in a verification system is a strict demand. Current technology cannot guarantee certainty for any of the agreements now being contemplated, and improvements in verification technology are often cancelled by weapons developments. If somehow verification methods gained a lead, the two sides might use up much of the advance by signing wider treaties, with the result that confidence in compliance would still be incomplete. Combined with a further position of some officials that all sections of an agreement are important in themselves<sup>2</sup>, the call to eliminate ambiguity is an attractive slogan that opponents of arms control could use to block worthwhile agreements.

If we try to analyze the ambiguity debate, we note that proponents of minimal ambiguity have left a number of issues unclear. Does the overall thoroughness of a verification system reduce ambiguity, or is ambiguity determined by some other feature? Director Rostow's statement went on to suggest that more intense verification was the answer, but a paradox is lurking here. A more sensitive test might pick up smaller events that would previously have been ignored, a point raised by Meyer (1984). For example, originally one side might

---

<sup>1</sup> "Ambiguity" here means uncertainty about objective facts of compliance due to an imperfect detection system, rather than semantic fuzziness in the treaty language. When the agreement presents a dichotomy such as compliance or non-compliance, ambiguity is maximum when a treaty partner holds a 50/50 likelihood.

<sup>2</sup> President Reagan's Report to Congress on Compliance, Feb. 1, 1985, stated "In order for arms control to have meaning . . . it is essential that all parties to agreements fully comply with them. Strict compliance with all provisions of arms control agreements is fundamental and this Administration will not accept anything less."

have accused the other of having too many missiles, but with better surveillance, it complains about the details of the missiles. Old ambiguities would be replaced by new ones, and it is not clear that uncertainty would go down overall. In this example, the better scheme shifts the ambiguity to the finer details, and so to the less important parts of the treaty. If this were the only problem surrounding ambiguity, the practical issue would be settled -- ambiguity would be unequivocally bad and correctable by more extensive verification. However, as this paper will show, even when comparing two verification systems for a *given* provision of the treaty, the better verification may yield more ambiguity.

Defining "better" requires some care, since the performance of verification is not purely technical, but depends in part on a human decision, the inspecting government's judgement about how much evidence it will require before it accuses. In a context involving simple violate/comply, accuse/not accuse decisions, one can define a system B as *better* than A, if, for *any* policy the user adopted with A, B would allow the user to adopt some policy that both increases the likelihood of an accusation given the other is guilty, and lowers the likelihood of a accusation given the other is innocent. It is easy to show that a more thorough system, that is, one that collects all the original information plus some more, is always better in this sense.

This paper will show that low ambiguity is not always desirable. It gives a formal model that clarifies the difference between a unambiguous verification system and a good one, and allows this difference to be made intuitive. It yields an example in which the user has a choice of two verification systems, and wisely prefers a system that would lead to more uncertainty when the evidence suggests that the other has violated the agreement. During successive arms control negotiations, the Kennedy, Nixon and Carter administrations set the desideratum for verification provisions that they allow a timely response to any militarily significant cheating (Krepon, 1985). This definition of quality correctly focused on the benefits of the *consequences*, and I will argue that the emphasis initiated during the Reagan administration on the inspecting government's *mental state*, its degree of uncertainty, is off track.

To summarize why low ambiguity is not per se desirable, if a government refines its verification system, and the other side knows of this improvement, the latter will generally be less inclined to violate. The observing government will indeed know that the other is less likely to violate. If the verification scheme returns evidence suggesting guilt, the inspecting government has to weigh this evidence against the implausibility that the other really would violate, "Would they dare do such a thing in the face of our better verification?" In certain cases, as the model will show, an observing government will attribute the possible evidence of guilt to the vagaries of its verification system, and decide that the prior expectation of innocence outweighs the evidence of guilt. The government will be sensible to do this. It will be less sure of a violation under the better system, and hold a likelihood of the other's guilt that is closer to 50/50. The observing government is confident of the higher deterrent power of the better test; its greater uncertainty stems from a positive quality of the improved system, not a weakness.

The role of ambiguity in verification has seldom been discussed, even informally<sup>3</sup>, in part, I believe, because two barriers hinder reasoning and talking about it. The first, termed here "myopic thinking," involves ignoring that the other side will behave differently in response to changes in your system, and that indeed you yourself will behave differently in expectation of their altered responses. Myopic thinking ignores that the other

---

<sup>3</sup> One exception is Meyer (1984). See also Katz (1980).

side is an aware calculating entity, and is related to the "fallacy of the last move" (York, 1970), so common in strategic analysis.

The second misconception involves the language typically used to analyze verification -- I will call it "detection-talk" -- which gives the impression that evidence of a violation comes back as a dichotomous "Violation!" or "No violation!" when in fact confirmation is usually a matter of degree. Analysts talk of "determining" compliance, "detecting" a violation, or "catching" the other cheating, and even the term "verification" itself suggests a clear finding on the truth. Perhaps these words were chosen to convince rather than to describe, but in important ways they are misleading. Detection-talk ignores the importance of false alarms, since it suggests that the only fault of a system is failure to "detect" a violation. If violations were simply detected or not detected, one could forego any discussion of ambiguity in verification, and this is in fact what one sees in the literature.

Two formal techniques will correct each of these misconceptions. Myopic thinking will be avoided through game theory. A game model will calculate how the observer's assessed probability of a violation should change with a new verification system. Detection talk will be replaced by a conception of gradual evidence, explicated by "identification curves." The curve for a given system states its tradeoff between false alarms and correct accusations. So far their arms control applications have been mainly in the subculture of the seismological discrimination of nuclear explosions (Dahlmann, 1977; Fetter, 1988), but they can clarify puzzles in many other types of arms agreements since they neatly summarize all the decision-relevant properties of verification systems involving choices that are dichotomous (cheat or not, accuse or not).

Game theory and identification curves are two essentially mathematical techniques, and deriving certain results will require some formalism. However, the assumptions and conclusions below will always be restated in words, and the logic leading to the conclusions is understandable without mathematics. Paraphrasing Wittgenstein, the attitude is that mathematics is a ladder that helps us ascend to a level of understanding. Having climbed it we can throw it down behind us.

The model can clarify some other puzzles. One can prove that better systems sometimes increase "ambivalence," which refers to a situation where *what to do* is unclear, as opposed to "ambiguity" which implies uncertainty about *what to believe*. Also, better systems sometimes increase the likelihood of a violation, but the observing government prefers them nevertheless. A third result that is intuitive but not trivial, is that increasing the difficulty of cheating lowers its likelihood. In addition the model can organize past research -- the appendix gives a survey of technical research on verification and how it relates to the political decisions of whether to comply and whether to accuse.

#### *The Influence of Ambiguity on the Decision to Accuse*

Freeing ourselves from detection-talk lets us ask how uncertainty interacts with other considerations in the compliance and accusation decisions. Ambiguity complicates the decision in several ways, which I first describe informally and later express in the model. First and most straightforward, the partial nature of evidence generates a discussion on just how strong the evidence really is. This was exemplified by the recent U.S. debate on Soviet compliance, summarized later in the Conclusions section.

Second, even if the Reagan Administration and arms control advocates had been of one mind on the degree of evidence, another factor might have led them to disagree whether the Soviet Union was complying. That is, even though two groups come to opposite conclusions on the same evidence, they might both be thinking

rationally. The reason is a simple fact of probability assessments: one's final beliefs should depend in part on one's prior beliefs before examining the immediate evidence. In the verification context the decision to accuse should reflect not only the degree of intelligence information pointing to guilt, but also the observer's previously held likelihood that the other side would violate the treaty. The influence of previous beliefs, of *general information*, as I will call it here, is not an irrational prejudice or stubbornness, since the beliefs may be based on real experience about the other's goals and political culture. (These ideas have been discussed by Jervis, 1976, and Meyer, 1984.)

Just how one's conclusion should depend on prior expectations of a violation, can be stated in a simple formula. We want to assess this probability:

P1: Probability that the other violated based on information from verification, as well as on our general information about the other.

This probability P1 can be calculated by using:

P2/P3: Probability P2 in this ratio is the likelihood that we would be getting this evidence from our verification system, based on the premise that the other violated. It is divided by P3, the probability of observing this evidence conditioned on the other's not having violated. (Both probabilities will require us to consider our general information.)

and

P4: Our prior probability that the other would violate, i.e., that based only on our general information. Elementary probability theory (e.g., Feller, 1960), gives this formula for P1:

Analyses purely of the verification scheme supply P2/P3 since they follow the lines: If the other side were violating what would our intelligence show, compared to the picture if the other side were complying? The ratio P2/P3, called the *likelihood ratio*, can be regarded as the strength of evidence from the verification system. However, a crucial variable is P4, our expectation of a violation based on our prior political and social views and not conditioned on the immediate evidence. An analyst who has always been convinced that the other side wishes to keep the treaty will be relatively reticent about concluding guilt, while one who distrusts the other anyway will take very low evidence as sufficient. Each is being consistent. If the verification system had sent back decisive evidence, the resulting probability P1 of a violation would be 0 or 1, but otherwise parties' prior attitudes should influence their conclusions.

The decision to accuse should also reflect the observer's costs or benefits for the possible outcomes. For example, a government that regards the treaty's advantages as small and violations as dangerous will be more ready to accuse. The simplest case involves four values attached to each of the four outcomes (cheat and accuse, cheat and not accuse, etc.) According to standard decision analysis, one should accuse when  $P1 \times V1 + (1-P1) \times V2 > P1 \times V3 + (1-P1) \times V4$ . Here the notation for one's values is

V1: making an accusation when the other has violated,

V2: making an accusation when the other has complied,

V3: not making an accusation when the other has violated,

V4: not making an accusation when the other has complied.

*Strategic versus Myopic Thinking in the Decision to Accuse*

In summary the following factors are important for a government deciding whether to accuse: its prior expectations, P4; the strength of evidence returned by the verification scheme P2/P3; and the values the government attaches to different outcomes, V1 to V4. From the description above it seems simple in principle to calculate P1 from P4 and P2/P3, estimate V1 to V4, apply the formula and then act. The problem, however, is that all the factors are entangled. They cannot be considered independently because the two decision-makers know each other's outlooks, and know that their interests are in partial opposition. The verification system's quality, for example, influences the decision of the other to violate and thus our prior probability of a violation (we know that the other is aware of that quality.) If the governments tried to decide by this usual cost/benefit/likelihood approach there would be no place to start.

Some of these circular influences can be seen in Figure 1. The decisions of both parties lie in sequences of arrows that form closed loops, meaning that the governments cannot expect to judge some set of quantities as exogenously given, and make a decision resting content with these original assessments. Circularities like these are the stock-in-trade of game theory, and this approach will be able to untangle them.

FIGURE 1 HERE

*A FORMAL MODEL OF OBSERVING AND ACCUSING*

The model's purpose is to give insight into the structure of verification decisions, and to state a framework for abstract reasoning. It is not to generate numerical predictions. The model should be just complicated enough to let us formulate some significant dilemmas within its structure, but no more so than that. Accordingly, it is one-sided in that only one government observes and judges, while only the other chooses whether or not to violate. That is, the observer does not have the option of violating. Another assumption is that each party makes a dichotomous decision: the inspectee cannot for example "partially" violate the treaty, nor select among several types of violations, and the observer cannot choose among hedged accusations of different strengths. Also the model portrays the observer as looking at a single test, so the term "test" here should be taken to mean the entire system of the government's direct intelligence, its complete ensemble of verification means. The three parts of the model are then an imperfect observation system, a decision by the inspectee to violate or not, and decision by the observer to accuse or not. Each of these will now be described.

*Detection-talk Versus Identification Curves*

Detection-talk suggests that verification evidence is typically cut and dry. This usage caused no problems when the Partial Test Ban was the major treaty in force and a violation would have been obvious, but experience from the ABM Treaty and SALT II has shown that verification is a matter of degree. Much of the recent debate in the United States was over factual claims about Soviet compliance.

Detection-talk is common even specialists in verification and has led them to adopt inadequate indices of quality. One measure has been a system's likelihood of "detecting" a violation given the other side has committed one. In 1979, for example, United States Secretary of Defense Harold Brown described the rating system in the U.S. intelligence community. Verification that could detect violations with probability 90% to 100% were termed "high confidence," with categories extending down to "low" (10% to 50%) and "very low" (0% to 10%) confidence. However, any approach involving only one number must be deficient because no single quantity can

convey the informational value of all the possible shades of evidential strength -- each body of evidence would require its own probability value. Another inadequacy in the probability-of-detection approach is that one can fudge a "perfect" detection scheme by adopting the policy of declaring a detection no matter how weak the evidence is. This would score a perfect record of no misses. Of course, given innocence, the likelihood of a "false alarm" would be unity and the treaty would collapse. Thus second fault of one-number probability-of-detection methods is that they ignore the false alarm rate.

What is a proper way to state a system's quality? The description should be stated in an abstract way, beyond the details of technical means of intelligence, for example, and the description should be independent of the expectations and values of the user? In the abstract, the user chooses the level of evidence required before accusing, then examines the evidence and accuses if it is higher, or does not accuse if it is lower. The quality of the verification system *for that particular criterion of evidence* is given by two numbers: the hit probability and the false alarm probability. The *hit probability* is the probability of an accusation when a violation has occurred, and the *false alarm probability* is the probability of an accusation given no violation. The user could change the criterion for accusing to require a more or less proof (be a more cautious or aggressive, due either to changes in prior expectations or values). Demanding a higher level would reduce the false alarm probability, which is good, but decrease the probability of a hit, which is bad. To specify the informational quality of a detection system *independent of the observer's choice of threshold of evidence*, one must graph the functional relationship between the two probabilities, to state how they trade off against each other.

Figure 2 is such a function, called the *Identification Curve*<sup>4</sup>. The elegant feature of the identification curve, the reason for our interest in it, is that it is a pure and complete description of test quality. It contains all one needs to know about the test to make a decision whether to accuse, but nothing more. The other information needed goes beyond the test: our prior likelihood of a violation, our values for the outcomes, and the evidence actually observed. The curve reflects *only* information about the test itself, in the sense that two observers may have different estimates of the opponent's disposition to violate, or different perceptions of the costs of their missing a violation or falsely accusing, but if they are using the same verification system they have the same identification curve.

---

FIGURE 2 HERE

---

The test determines the identification curve as a whole, but the observer's prior probability (P4) and values (V1 to V4) fix the particular point on the curve the observer is using. The threshold of evidence that an observer chooses based on expectations and attitudes, determines a point on the curve. The coordinates of that point are the false alarm and hit probabilities the observer would display when the evidence is just enough to trigger an accusation. Someone who chooses a threshold at a point on the identification curve is said to be *operating* at that point. A government with a low prior expectation of a violation and an attitude that misses are not as serious as false alarms, is operating at the lower left end of the curve and evincing a general tendency not to accuse. Conversely, a government disposed to accuse operates at the upper right end. Changing an observer's

---

4

The term comes from the seismological identification of nuclear explosions. Dahlman (1977) and Fetter (1988) give examples. The method seems to have originated in engineering, where the curve was known as the Receiver Operating Characteristic, and is used in several other fields such as the analysis of medical diagnoses and perception psychology.

costs, benefits and prior expectations causes the best policy to move around to different positions on the curve, but the curve itself does not change, since the latter depends only of the evidential qualities of the verification system.

For a given verification system, how would one derive its identification curve? The answer depends on the context. In seismological detection, the curve can be constructed from data from real explosions and earthquakes. (Of course past detonations involved no attempt at concealment, so more realistic identification curves would require a physical model of the evasion technique.) In areas such as arms limitation agreements, evidence comes from many sources, and sometimes is "soft," based on human judgement as well as technological hardware. One must assess the likelihood ratios that a given policy generates, so the actual function is harder to produce. Some general techniques for teasing such judgements out of specialists are described by Spetzler and Stael von Holstein (1975), but it might be more realistic to regard the curves as tools for clarifying concepts, rather than making real decisions.

#### *Properties of Identification Curves*

Only certain functions can be identification curves, and three necessary features are listed now. Since an observer has the options of never or always accusing, an identification curve must *start at (0,0) and end at (1,1)*. It should *lie above the diagonal*, since it should produce a higher probability of accusation when there is cheating than when there is none. A less obvious property is that *the curve's slope decreases moving to the right*. It can be shown that the slope at the point determined by a certain threshold of evidence is the likelihood of this degree of evidence given cheating, divided by the likelihood given non-cheating, i.e., the slope is the likelihood ratio  $P2/P3$ , defined earlier (Green and Swets, 1966, p.38). The explanation for the decreasing slope is that if one requires less evidence of cheating at the threshold for accusation, and then observes that exact amount of evidence, that should mean that cheating is more likely. Therefore moving rightwards on the identification curve should correspond to a decreasing likelihood ratio, and thus a decreasing slope.

A further consequence is: *If one curve lies entirely northwest of another, the former is a better test*. "Better" was defined above as allowing one to choose an accusation policy that yields a lower false alarm and a higher hit rate. A point on the better test that lies northwest of the point chosen on the worse test will satisfy this. A worthless test would have the diagonal line as its identification curve, and lead the user to accuse with a likelihood independent of the evidence. A perfect test, on the other hand, would involve a line starting at (0,0), proceeding vertically to the upper left corner (0,1) then horizontally to (1,1). It would allow the user to accuse when and only when there was cheating.

#### *The Observer's and Inspectee's Values*

The next step in constructing the model is to specify the goals of the two parties, that is, their degrees of preference for a continuing agreement, successful cheating, etc. I assumed verification is one-way, so the actors can be called the *Observer* and the *Inspectee*. The Observer's values for the outcomes are denoted as follows:

- No cheating and not accuse: t
- No cheating and accuse: w
- Cheating and accuse: y
- Cheating and not accuse: z.

It is assumed that  $t$  is greater than  $z$  and that  $w$  and  $y$  lie in between:  $t > w > z$  and  $t > y > z$ , i.e., that  $w$  or  $y$ , which involve a breakdown of the agreement are preferred to  $z$ , a secret violation.

The values of the Inspectee are

Cheat and not be accused:  $A$

Not cheat and not be accused:  $b$

Not cheat and be accused:  $c$

Cheat and be accused:  $d$ .

I assume that  $A$  is greater than  $d$ , that  $b$  and  $c$  lie between these two values:  $A > b > d$  and  $A > c > d$ . These orderings show a perceived conflict of interest since the Observer's ideal outcome is not the same as the Inspectee's. Someone outside the distrust and rivalry of the two countries might object that the Inspectee would gain more by complying than by cheating successfully, but these goals are evaluated from the participant's perspective, not necessarily objectively. The purpose here is to analyze the Observer's decision, and it seems plausible that each power might see the other as motivated to cheat.

What does each side know about the other's goals? I assume that each knows everything, except that the Observer is uncertain about the Inspectee's degree of motivation to cheat,  $A$ . (Capitalizing  $A$  shows its special status.) The Inspectee is aware of  $A$  but the Observer has only a probability distribution  $H$  over it: the Observer's probability that  $A$  is lower than a given number  $a$  is given by the value  $H(a)$  of the function.

There is a final quantity,  $P1$ , about which I make no assumptions. It is the Observer's prior expectation that the other will cheat, and is determined within the model as a consequence of the game-theoretical assumption that each is maximizing its gain, given its expectation of the other's behaviour. This is a more realistic depiction of the Observer's thinking, better than regarding the Inspectee as a random entity.

### *The Course of Events in the Model*

Events proceed as follows:

- 1) the Inspectee assesses  $A$ , its value for successful cheating,
- 2) the Inspectee decides whether to cheat,
- 3) the Observer calculates an evidence threshold above which it will accuse,
- 4) the Observer examines the strength of evidence that the Inspectee is cheating,
- 5) the Observer accuses or does not accuse,
- 6) the two parties receive the values appropriate to the outcome.

Game-theoretical methods will determine a best strategy for each side. What would a strategy look like? A strategy means as a complete plan of action telling what to do, perhaps as a function of something learned during the play. The Observer's must choose a level of evidence necessary to accuse. This threshold has a one-to-one relationship with the hit and false alarm probabilities, and the hit probability is itself a function of the false alarm probability (through the identification curve function  $f$ ). Therefore one regard the Observer's strategy as a false alarm probability  $x^*$ .

The possible strategies for the Inspectee are more complicated: the latter chooses whether to cheat or not depending on the motivation  $A$ , so a strategy must specify a choice for every value of  $A$ , that is for all "types" of Inspectees. To simplify the task one can make a justifiable assumption that there is a cutoff value  $a^*$  such an Inspectee with a motivation at or beyond the cutoff will cheat, and one with lower motivation will not. Specifying a single  $a^*$  will give a rule for every type of Inspectee. From now on I will talk as if  $a^*$  were the Inspectee's "strategy," although more precisely it is a rule for determining a strategy.

The goal is then to determine the pair  $a^*$  and  $x^*$  that results when each tries to maximize its benefits given a sensible expectation of the other's behaviour. The criterion of rationality used is an equilibrium: *an Observer's  $x^*$  and Inspectee's  $a^*$  are such that even if one party knew the other's choice, it would have no incentive to change its own behaviour.* (The pair I calculate actually satisfies stronger criterion, they are *strict* and *subgame perfect* equilibrium in that either side would be losing by deviating from the policy, and no one acts irrationally even at positions in the game that have zero probability of being reaching.)

To calculate  $x^*$  and  $a^*$  one writes an expression for the expected value to the Observer as a function of the choice of threshold  $x$ , given that the other has selected  $a^*$ , and chooses  $x$  to maximize this expression. This maximization is expressed by setting the derivative of the expression with respect to  $x$ , equal to 0 at the unknown optimizing value, which is designated  $x^*$ . Likewise one writes an expression for the Inspectee's payoff as a function of  $a$ , given that the observer uses  $x^*$ , and requires that cutoff motivation  $a^*$  be set to maximize it. These operations yield a pair of nonlinear simultaneous equations, given in Appendix 2, in the two starred variables. The following examples were solved on a computer using simple approximation methods.

#### *How a Better Test can be More Ambiguous*

Here I give an example to show that a better test can sometimes induce more ambiguity. It uses a single indicator test, a single observation that simply signals "guilty" or "innocent" to show greater or lesser evidence of guilt<sup>5</sup>. One might object that a real Observer would be uncertain about all four of the other's values, not just  $A$ . Introducing such uncertainty would involve slightly more complicated formulae but would not change the basic results.

After the Observer switches to a better indicator, when the indicator says "guilty," the Observer will be less sure that the Inspectee has violated. The results are the probabilities in Table 1. The Improved Indicator is more likely to be correct than the Original Indicator whether there is a violation or not, and therefore it is indeed better. (If the other is guilty, it says so with probability .76 rather than .30; if innocent, it says that with probability .90 rather than 80.) The identification curves of the two tests are shown in Figure 3. Substituting the

---

<sup>5</sup> A *single-indicator test* is defined here as one that sends back just two states of evidence, "guilty" and "innocent." The indicator is probabilistically associated with a violation or compliance, and is more likely to say "guilty" when a violation has occurred than when one has not. What will be the form of identification curve? If one defines  $x = \text{Prob}(\text{Indicator says "guilty," given no violation})$  and  $y = \text{Prob}(\text{Indicator says "innocent," given a violation})$ , the observer could choose a threshold-of-evidence-for-accusing corresponding to hit probability  $x$  and false alarm probability  $y$ , simply by accusing when and only when the indicator says "guilty." It follows that the identification curve will contain the point with horizontal coordinate  $x$  and vertical coordinate  $y$ . Alternatively the observer could choose to use the indicator to operate at any point on the line segment joining  $(x,y)$  to  $(1,1)$  or the segment joining it to  $(0,0)$ . For example, if the observer adopted the policy of always accusing if the indicator says "guilty," and accusing with some random probability if it says "innocent," then this would yield a false alarm and a hit probability lying on the straight line joining  $(0,0)$  and  $(x,y)$ , as in Figure 3. Although unrealistic, single-indicator tests are easy to handle mathematically, and are useful in proving claims that certain phenomena can occur in principle.

formulae for these curves in the two non-linear equations of Appendix 2, gives optimal solutions  $a^*$  and  $x^*$  for each player for each function, as in Table 1.

---

TABLE 1 AND FIGURE 3 HERE

---

With the Original Indicator, the Inspectee should cheat whenever motivation is greater than  $a^* = .467$ , thus has a probability .533 of cheating; the Observer should always accuse when the indicator says "guilty" and accuse with probability .026 when it says "innocent." (This follows from calculating  $x^*$  using the two nonlinear equations in the Appendix applied to the identification curve of Figure 3.) With the Improved Indicator the Inspectee cheats when motivation is greater than .884, giving probability .116 of cheating, and the Observer should accuse with probability .617 when the indicator says "guilty" and never accuse when it says "innocent." The point of the result here is that when the Observer sees a guilty indicator, its opinion is more definite with the Original than the Improved (it holds probability .632 that the other is cheating rather than .500). Although Director Rostow advocated lower ambiguity, the Observer would experience greater ambiguity under the better test: indeed it would then face the maximum possible ambiguity, assessing a 50/50 likelihood that the other had complied.

Although the new test yields greater ambiguity, its superiority manifests itself in a number of ways: there will be less cheating (12% of the time versus 53%); the Observer experiences lower ambiguity (probability of a violation is .03 versus .50) when the indicator says "innocent"; and most important, the Observer has a higher expected benefit (.89 versus .73). So the new test really is better, but "better" does not mean "less ambiguous."

The explanation is that the new test is so much more thorough than the old that the Observer is very confident before examining the evidence that the Inspectee will not cheat. (The Observer has prior confidence .884 for the new versus .467 for the old. These are simply the values of  $a^*$ , which equal the probabilities that the Inspectee's motivation is below the threshold.) Even though the Improved Indicator gives more informative evidence, the Observer's revised opinion in the face of guilty evidence is still lower than with the Original Indicator, because the Observer quite sensibly does not shift far away from the prior view that cheating is unlikely.

### CONSEQUENCES OF THE MODEL

#### *How a better test can cause more ambivalence*

The Commission on Integrated Long-Term Strategy worried about verification that makes it difficult to decide how to act. This concept, which I term *ambivalence*, is slightly different from ambiguity in the following way. One's aversion to some outcomes might predispose one to accuse as long as the probability of a violation were above some very low value. In that situation a 50/50 assessment would leave full ambiguity about what to believe, but low ambivalence about how to act.

The example of Table 1 illustrates that ambivalence can be higher under a better test. When evidence comes back from the poorer Original Indicator that suggests guilt, the decision is clear. Not accusing achieves an expectation of .132 versus accusing's expectation of 0. (These are calculated by determining the equilibrium strategies, which give, in turn, the probability of cheating, of seeing various strengths of evidence, and of accusing. One can then determine the likelihood of the four possible outcomes, and weight them by the proper

probabilities and thus calculate expectations.) With the Improved Indicator the expectation of both courses of action is 0 in the face of guilty evidence. The better test has generated more ambivalence.

*How a better test can mean more cheating*

The model sets up a paradigm situation in verification: one side deciding whether to cheat, the other watching and weighing whether to accuse. It provides a framework to clarify a number of other questions. One can ask, in particular, whether a better test might induce more cheating. The answer is yes. With the two single-indicator tests and with values held by the governments as given in Table 2, the improved indicator is a better test, and Tables 1 and 2 show that the Inspectee's threshold motivation for cheating goes down in the improved test, resulting in a rise in the probability of cheating from .125 to .143.

---

TABLE 2 AND FIGURE 4 HERE

---

One can glean the explanation by examining what other strategies and beliefs result when the Observer uses the new test. The reasons behind the increase in cheating illustrate an interesting general point about verification system: some are especially effective when used by an observer cautious in accusing, while others are better in the hands of an aggressive observer. A system whose identification curve is skewed to the lower left is relatively useful to a cautious observer since its quality is high just in that region where the observer will operate. It is one where much guilty evidence is necessary before the indicator signals "guilty," but should the indicator be tripped, then the test is highly credible. This type of scheme can be termed "sluggish." On the other hand a system with a curve skewed to the upper right would be acceptable to an aggressive observer since its poor behaviour is confined to a area in the square the observer will not use, and one can term it a "hair-trigger" system.

An example of making a system better and more sluggish would be the following: the observer has a seismic network that returns accurate evidence on the position, depth, and location of a seismic "event" along with an imperfect "yes" or "no" indicator of whether it was a nuclear explosion. A better test would be to add a treaty provision for drilling for radioactivity at the site whenever the indicator shows "guilty." This would give a near-zero false alarm probability although the system might miss some explosions. Thus its identification curve would be strongly skewed to the left.

The Improved Indicator in Figure 4 is especially sluggish, i.e., relatively more effective in the hands of a cautious observer. Switching to the Improved Indicator system induces the Observer to become more cautious in order to exploit the indicator's positive properties. The Observer will be less likely to accuse falsely, at the slight expense of the scheme's deterrent potential. If evidence arrives that is just at the Observer's threshold of accusation, the values of  $f(x^*)$  and  $x^*$  show that the optimal false alarm probability is reduced by .135 at a cost of only about .005 in the hit rate. The Inspectee will anticipate the Observer's greater caution and will be slightly more willing to cheat. It is interesting that both Observer and Inspectee prefer the Improved Indicator to the old.

*What happens when cheating becomes more costly?*

A further question that the model can address involves the effects of increasing the cost of cheating. Would we be more justified in believing cheating will occur, or less so? The answer is not obvious -- sometimes considerations that suggest innocence may just as plausibly imply guilt. In 1979 an American surveillance

satellite recorded a light flash with a time course similar to a nuclear explosion (Walters and Zinn, 1985). The flash observed was consistent with a nuclear test in the Atlantic Ocean near Southern Africa and some commentators suggested that the Republic of South Africa, or perhaps Israel, had exploded a first weapon. A contrary argument was that the ocean surface at that location would be especially rough, an unlikely place to choose for a nuclear test. But would not the implausibility of this site give a concealer more confidence that it will not be accused? Should it not support the case that a nuclear test occurred? It is clear that our assessment of the cost of cheating should alter our judgement, but unclear in which direction?

I can incorporate an increase in the cost of cheating by altering the Inspectee's payoffs to  $A+k$  and  $d+k$ . The constant  $k$  is negative, and represents the change in benefit to the cheater associated with cheating under new circumstances, in the example, the added costs of testing on a rough ocean. Caught or not, the cheater has to pay that extra charge. I can then examine how optimal behaviour changes with  $k$ . The assumption here is that the increased cost is given, not a strategic choice of the Inspectee. Calculations given in Appendix 2 imply a general conclusion, true for all identification curves and values for outcomes within the model's assumptions: if an illicit nuclear test becomes more costly to conduct, the Inspectee will be less likely to cheat and the Observer will demand more evidence before accusing.

### *CONCLUSIONS*

One can judge the impact of ambiguous detection on the recent U.S. debate by looking at a list of allegations of Soviet non-compliance. A report from the Reagan Administration claimed twelve Soviet violations or "likely" violations of agreements (Office of Press Secretary, the White House, 1987). The Arms Control Association, a pro-arms control public education group, released a counterdocument replying to each point (1987), and together these are a precis of the debate. Comparing the allegations and responses, one can classify the issues as involving semantic ambiguity in the treaty language, or factual ambiguity, or both, or neither. Fully six points of difference center on factual ambiguity, two are semantic, two others are a combination of factual and semantic, and in one instance both American parties saw a clear violation.<sup>6</sup> Soviet accusations of U.S.

---

<sup>6</sup> For these six accusations the debate was primarily factual: whether new Soviet ABM systems were mobile, whether the yields of underground tests exceeded the Threshold Test Ban Treaty limit, whether certain Soviet Bear bombers were dismantled, whether the "yellow rain" of Cambodia was Soviet-supplied toxins, whether chemical weapons were used in Afghanistan, and whether the 1979 anthrax outbreak in Sverdlovsk indicated a biological warfare facility there. Two of the issues were mainly semantic: whether the ban on "rapid" reloading of ABM interceptors proscribes a new Soviet system, and whether extensive encrypting of missile test telemetry "impedes" verification. Two further issues involved both factual and semantic uncertainties: whether the SS-25 intercontinental missile carried too high a "throwweight" and whether the SA-12 anti-aircraft missile had ABM "potential." Two issues cannot be categorized as either semantic or evidential uncertainty: the first was the radar site then under construction at Krasnoyarsk which both the Arms Control Association and the Administration regarded as a violation, and the other was the concurrent operation of air and ABM defenses where the validity of the Reagan Administration's charge depended on the terms of a secret Soviet-American understanding.

noncompliance showed a reasonably similar pattern.<sup>7</sup> Uncertainty about the facts of compliance has thus been significant in promoting doubts about arms treaties.

Many items on lists like these seem rhetorical, and opponents of arms control have sometimes raised accusations hoping to weaken the agreements. Their success in generating doubts about treaties shows that uncertainty in verification has been fundamentally misunderstood. Uncertainty does not signal that the provisions are lax, and it should not trigger excessive hindsight about how the treaty was poorly negotiated. It can arise from one's prior expectations, and even a positive quality of the verification system, its deterrent power.

*Effective* verification, in proper usage, means that the parties enjoy a high likelihood of accusing if the other violates and of not accusing otherwise, these two goals being balanced according to the user's values. The aim of verification is to deter violations and reduce false alarms, not to provide governments with easier decisions. The gist of this analysis is that ambiguity and ambivalence are inevitably part of the process.

---

<sup>7</sup> The Soviet press and television have emphasized these points (and others) that involve factual questions about U.S. behaviour (dates refer to the Foreign Broadcast Information Service *Daily Report, Soviet Union*; other references are given by Duffy, 1988, and Koulik, 1991): violation of the Non-Proliferation Treaty by supplying nuclear materials to Israel (April 10, 1987); violation of the Biological Weapons Treaty (March 31, 1987); use of Minuteman missiles for antimissile purposes; transferring ABM components outside US territory. Other issues involve the meaning of treaty language: the Fylingdales and Greenland radar (April 7, 1987) where the meaning of permissible "modernization" is at issue, and the wide interpretation of Article V of the ABM Treaty, violation of confidentiality of the SCC. A further accusation, which the U.S. acknowledged as a deliberate choice, was the conversion B-52's to carry more cruise missiles than permitted by SALT II (Dec.8, 1986). Several other allegations fall into the category of "undermining," "circumventing" or acting "contrary to the spirit of" certain agreements, outside our categorization.

*Appendix 1. FORMAL STUDIES OF THE VERIFICATION DECISION-MAKING*

The model's structure covers each part of the verification decision, and allows us to locate other formal studies to show how they relate to each other and ultimately to the political decision to comply or accuse. I will include only works that are formal and decision-oriented, that is, only mathematical studies involving probabilities and costs. There have been surprisingly many, but often writers were unaware of each other's research. Often they did not elaborate on their work's place in the structure of factors leading to a policy decision about compliance.

*Probability assessment studies*

The first group deals with *probability assessment*, measuring evidence of a violation, and fits in the "VERIFICATION SYSTEM" box of Figure 1. Some use Bayesian statistics to calculate the strength of evidence  $P2/P3$  directly (e.g., Heckrotte and Moulthrop, 1984; Ciervo and Hall, 1987; Nicholson and Heasler, 1984; Bryson, 1984). Other studies use hypothesis testing methods following the logic of classical statistics to derive only  $P3$  (Hall, Nicholson and Heasler, 1984; Davis, 1984; Shumway and Rivers, 1984; Westervelt, MacKay and Bryson, 1984; Berg, 1986; Lewis, 1990). If the value of  $P3$ , the evidence given no violation, is low, they conclude guilt. This inference is valid only if one regards  $P2$  as substantial and more or less constant with respect to the evidence --  $P3$  thus functions as an index for the likelihood ratio  $P2/P3$ . An example would be a one-sided test of the null hypothesis that a certain explosion had a yield of the treaty limit of 150 kilotons. Another indirect way to get at the likelihood ratio would be to find an estimate of the uncertain physical magnitude, and several studies apply statistical methods of estimation to the problem of explosive yield (Berg and Deemer, 1977; and Ciervo, Hall and Thomas, 1977). Again the estimate per se does not give us the likelihood of a violation since an estimate of, say, 200 kilotons might have arisen by chance from a 150 kiloton explosion, but if one could show that the variability of the estimation procedure is low, then the likelihood ratio indicating a violation would be high. Some studies describe discrimination techniques, e.g., optimal methods for distinguishing earthquakes from explosions (Douglas, 1981; Tjostheid, 1981.) Discrimination methods apply to dichotomies (explosion versus earthquake) whereas estimation methods apply to continuous quantities (size of explosion), but otherwise the two are alike, and one can categorize both as ways to generate indirect indices for the likelihood ratio  $P2/P3$ .

Much of the evidence relevant to the value of the likelihood ratio comes from intelligence sources that cannot be quantified, but formal methods may still play a role. The intelligence analyst may have a natural skill in assessing likelihoods of individual events, but not at articulating them or combining them. Probability assessment methods try to tease the probabilities of elementary events out of the observer so as to construct the probability of the more complicated event of interest. Intelligence agencies currently use these methods, although, as one would expect, details are not announced. One discussion of possible approaches is given by Cohen, Schum, Freeling and Chinnis (1985).

Also within the first group, probability assessment studies, are those that discuss the choice among different ways of collecting data. Instead of accepting the data as given and showing how to squeeze maximum information out of it, these compare different procedures of gathering the data and discuss how to acquire the necessary evidence for the least cost of collection -- how many units must be observed and in what pattern. For example, the problem of verifying quotas, numerical limits on missiles or troops, has been treated by Meyer (1979), Richelson (1979), DeVolpi (1987), Wittekindt (1984), and Fetter and Rodionov (1991). Avenhaus (1977, 1986) has published very thorough work on the monitoring of nuclear materials to prevent diversion to weaponry, and two papers from the United Nations Conference on Disarmament (Netherlands Delegation, 1984; and Japan

Delegation, Disarmament, 1986) discuss a chemical weapons treaty. The above studies derive mathematical formulae and so consider only a few variables, but complexity can be added and analysed with the help of computer programs. Examples are the codes SNAP/D (Ciervo, Sanemitsu, Snead, and Suey, 1980) and NETWORTH, which calculate how much confidence a certain array of seismic stations can convey on the yield and location of a Soviet nuclear test.

#### *Value Assessment Studies*

The second group, *value assessment*, tries to determine formally the seriousness of cheating, and thus these studies lie within the box "OURSELVES" of Figure 1. The group has meagre representation outside government agencies. One example is Hafemeister's use of strategic exchange models (1986) to study the cost to the United States of a Soviet treaty breakout. Alfred Lieberman, as Chief of the Operational Analysis Division of the Arms Control and Disarmament Agency, surveyed the use of exchange models to estimate the cost of successful violation (1984). A description of SIRMEN, a computer code used by the agency for this analysis, was published by Academy of Interscience Methodology (1978).

#### *Decision Analysis Studies*

The third group combines probabilities and values to the observer or inspectee to determine an optimal inspection procedure. They fit in the "OURSELVES" and "VERIFICATION SYSTEM" boxes. Some of these are detailed interactive computer programs aimed at letting the user explore different schemes. In the Regional Seismic Verification System computer program (Younker et al., 1985; Judd et al., 1986, 1988; Strait and Sichertman, 1986; Hannon, 1972), one can alter assumptions about the verification system and the other's motivation to evade. In the same vein but less complex is Ulvila and Brown's discussion of heuristic rules of allotting resources in materials safeguards inspection (1981). Ciervo and Watson (1983) give mathematical calculations for verifying a treaty on intermediate nuclear forces and Ciervo (1974) discusses the problem in regard to a test ban. Other studies in the decision-theoretic mould investigate seismic identification curves (Ericsson, 1970; and Weichert and Basham, 1973). The third group also includes Wiesner curves which describe how the requirements of verification grow as the treaty becomes tighter, i.e., sets stronger limits on strategic weapons (Wiesner, 1961; Karkoszka, 1977).

#### *Game-theoretical Studies*

The final group, the game-theoretical studies, have been more abstract, aimed at elucidating general principles (Bellany, 1982) or proving theorems. These involve all the boxes in the diagram. They differ from the decision theory group in that each party makes a choice that takes into account the other's thinking and motives. The first studies date from the early 1960's when the test-ban negotiations were reaching an impasse on the issue of seven yearly inspections, the American demand, versus three, the Soviet offer. The great concern at the time over the value of each on-site inspection was reflected in the work of three prominent game theorists Harold Kuhn (1963), Melvin Dresher (1962) and Michael Maschler (1966, 1967), who treated very elegantly the problem of the best use of a quota of inspections. The Inspector faces the dilemma of expending one inspection from the quota on a suspicious event knowing that there will then be fewer inspections left to deter a violation. Maschler's papers showed the advantages of having the inspector make an open commitment to a strategy to the other before

the evidence is examined.<sup>8</sup> Kilgour and Brams (1990) expand on this theme. Rapoport (1986) gave a simple example in this group. Moglewer (1973), Brams, Davis and Kilgour (1988) and Kilgour (1990) extend quota verification theory.

Other game-theoretical studies have been more abstract, representing the evidence as a single indicator that alleges "innocent" or "guilty" (Avenhaus and Frick, 1983; Fichtner, 1985, 1986; Dacey, 1979; Wittman, 1987; Weissenberger, 1991). Brams and his associates have developed extensively a number of such models (Brams, 1985; Brams and Davis, 1987; Brams and Kilgour, 1986; Brams and Kilgour, 1987, 1988). Weissenberger's analysis (1990) jointly determines the theoretically verification criterion and the optimal treaty provisions, and in the latter feature it steps outside the scheme of Figure 1. Filar (1983) and Filar and Schultz (1983) treated the case of an inspector who has different travel costs between different sites. Some recent sophisticated game theory work has turned back to specific contexts, notably the monitoring of materials in nuclear energy plants in support of the Non-Proliferation Treaty (Avenhaus, 1986; Avenhaus and Canty, 1987; Avenhaus, Fichtner and Vachon, 1987; Avenhaus and Zamir, 1988; Bierlein, 1983; and Zamir, 1987.)

Looking at the whole survey points up the lack of studies in the upper righthand box of Figure 1, which would look at the verification question from the other side's point of view.

---

<sup>8</sup> Models like Maschler's "price-leadership" model face the problem of threat credibility. They imply that the observing government names the threshold irrevocably before examining the evidence and therefore will sometimes accuse, even when evidence of violating is so low that the it does not believe the other has violated, and in fact expects greater benefit by remaining silent and preserving the treaty. If it can make this threat credible it will benefit from doing so, but a sensible actor would not deliberately do this unless there were some mechanism forcing it to keep its commitment. Unless the modeller can identify such a mechanism in the political world, it seems safer to do as the present model has done, to not assume the ability to precommit to actions which at a future time would be harmful to oneself.

*Appendix 2: Derivation of the equilibrium strategies.*

To show that a pair  $x^*$  and  $a^*$  is an equilibrium, we must show that the Observer has no incentive to move from  $x^*$  given the Inspectee uses  $a^*$ , and we must show the same for the Inspectee with respect to  $a^*$  and  $x^*$ . Clearly the extreme strategy  $x = 0$ , never accuse, is not part of an equilibrium pair, since the Inspectee's equilibrium strategy would then be to cheat for any incentive level (i.e., to choose the minimum  $a^*$ ), and the Observer would have an incentive to always accuse, contradicting the assumption that  $x = 0$ . For analogous reasons  $x = 1$  cannot be part of an equilibrium pair.

To examine non-extreme values of  $x$  and  $a$  as candidates for an equilibrium, we calculate the two players' expectations as functions of their strategies and require that they be maximized simultaneously. The Observer's expectation  $E_O(x)$  as a function of an arbitrary threshold  $x$ , given the Inspectee uses  $a^*$ , is the value of the four possible outcomes weighted by their probabilities:

$$E_O(x) \equiv P(\text{NC})P(A \mid \text{NC}) V(\text{A\&NC}) + P(\text{NC})P(\text{NA} \mid \text{NC}) V(\text{NA\&NC}) + P(\text{C})P(A \mid \text{C}) V(\text{A\&C}) + P(\text{C})P(\text{NA} \mid \text{C}) V(\text{NA\&C}),$$

where A and C mean accuse and cheat respectively, NA means not accuse, etc., and P and V stand for the probability of the event and the value of the outcome. Then for  $x$  in  $[0,1]$

$$E_O(x) = H(a^*)xw + H(a^*)(1-x)t + [1-H(a^*)]f(x)y + [1-H(a^*)][1-f(x)]z.$$

Since the Observer will maximize this expression with respect to  $x$ , then  $dE_O(x)/dx=0$  at  $x=x^*$ . This, along with the assumed differentiability of  $f$ , implies that the derivative of  $E_O(x)$  is zero at  $x=x^*$ :

$$H(a^*)w - H(a^*)t + [1-H(a^*)]f'(x^*)y - [1-H(a^*)]f'(x^*)z = 0,$$

yielding  $H(a^*) = f'(x^*)/[(t-w)/(y-z) + f'(x^*)]$  which in turn gives

Given the Observer uses  $x^*$ , the Inspectee will choose to maximize

This expression is the Inspectee's expectation before learning the value  $a$  of its motivation to cheat.

Differentiating this expression with respect to  $a$  and equating the result to zero gives the value of  $a$  at which the expected benefit of complying is just equal to that of violating;

Next we show that equations (1) and (2) have exactly one solution for  $x^*$  in  $[0,1]$ . The procedure is to show that (1) defines  $a^*$  as a continuous and strictly decreasing function of  $x^*$  and (2) defines it as a continuous and strictly increasing function of  $x$ , then to show that these two curves are positioned such that they cross, and therefore the two equations have exactly one solution.

Since  $H(a^*)$  is strictly increasing for non-extreme values of  $a^*$ , equation (1) defines  $a^*$  as a function of  $x^*$ , which we will denote  $g_1(x^*)$ , and (2) defines another such function, denoted  $g_2(x^*)$ . Since  $f'$  is strictly decreasing,  $g_1$  is strictly decreasing, for  $x^*$  in  $[0,1]$ . To show that  $g_2$  is strictly increasing we calculate its derivative,

We wish to show that the numerator of (3) is positive for  $x$  in  $(0,1)$ . At  $x = 0$  it is  $b[f'(0) - 1] + c - f'(0)d$ , which is positive, since  $f'(0) > 1$  and  $b, c > d$ , and at  $x = 1$  it equals  $f'(1)(c - d)$  which is positive or zero. To show that the numerator is positive in between, we examine its derivative  $[b(1 - x) + c - d]f''$ . The first factor must be positive since  $b, c > d$ , and the second factor  $f''$  is negative for  $x^*$  in  $(0,1)$ . In all we know that the numerator of (3) is positive at  $x = 0$ , decreases strictly and continuously as  $x$  goes to 1, and is positive or zero at  $x = 1$ . Therefore (3), the derivative of  $g_2$ , is positive and  $g_2$  itself is strictly increasing, also continuous.

It is easy to check that  $g_1(0) > g_2(0)$  and that  $g_1(1) < g_2(1)$ . Therefore the two curves cross exactly once, and this intersection gives a strategy pair  $x^*$  and  $a^*$  that satisfies (1) and (2). These latter equations are necessary conditions for an equilibrium but not sufficient. We must also check that the two expectations are really maxima. Their second derivatives indeed turn out to be negative.

*Response to an increased cost for violating.*

To examine the change in the solution of equations (1) and (2) when we add an increment  $k$  to the value of cheating, we alter  $A$  to  $A+k$  and  $d$  to  $d+k$ . If cheating becomes more costly, as in the case of nuclear testing in rough ocean, then the corresponding event in the model is  $k$  changing from zero to some negative value.

Equations (1) and (2) can be denoted

$$C(x^*, a^*, k) = 0, D(x^*, a^*, k) = 0, \quad (4)$$

respectively, where

$$C(x^*, a^*, k) \equiv (y-z)[1-H(a^*)]f'(x^*) - (t-w)H(a^*),$$

$$D(x^*, a^*, k) \equiv (a^*+k)(1-f(x^*)) + f(x^*)(d+k) + x^*c + (1-x^*)b + a^* + k + b + f(x^*)(d-a^*) + x^*(c-b).$$

Since the equilibrium is unique, each value of  $k$  yields a pair  $x^*$  and  $a^*$  so we can regard  $x^*$  and  $a^*$  as two functions of  $k$ , which are determined by the equations (4). The latter can be written,

$$C(x^*(k), a^*(k), k) = 0 \text{ and } D(x^*(k), a^*(k), k) = 0.$$

Differentiating with respect to  $k$ , using the chain rule,

These are linear equations in  $dx^*/dk$  and  $da^*/dk$ , and have solutions,

We may not know the exact values of the derivatives on the righthand side but we can determine their signs by differentiating (4) and (5) and recalling the assumptions made earlier about  $f$ ,  $a$ ,  $b$ , etc.

In this case we can derive conclusions about the signs of the values on the left by substituting signs:

This means that with lower  $k$  (when the test becomes more costly), the inspector lowers the threshold, i.e., is willing to take less evidence to trigger an accusation. Thus when  $k$  falls, the threshold motive for cheating  $a^*$  rises and cheating, and the inspectee cheats only at higher levels of motivation.

Thus when  $k$  falls, the threshold motivation for cheating  $a^*$  rises, and cheating becomes less likely.

*Assumptions on values and expectations:*

$t = 1, b, c, w, y = 0, d = z = -1,$

A is drawn from a uniform distribution on  $[0,1]$ :  $H(a) = a.$

<i>Assumptions about Indicators</i>	Original Indicator (Solid Line Figure 3)	Improved Indicator (Dotted Line Figure 3)
Pr("guilty"/violation)	.30	.76
Pr("guilty"/no violation)	.20	.10
Pr("innocent"/violation)	.70	.24
Pr("innocent"/no violation)	.80	.90
<i>Results</i>		
Probability of violation before observing evidence	.533	.116
Probability of violation given a guilty indicator	.632	.50
Probability of violation given an innocent indicator	.500	.034
Observer's expected benefit before observing indicator	.727	.891

*Table 1.* A better test that induces more ambiguity and ambivalence when a violation is detected.

<i>Assumptions about Indicators</i>	Original Indicator (Solid Line Figure 4)	Improved Indicator (Dotted Line Figure 4)
Pr("guilty"/violation)	.44	.46
Pr("guilty"/no violation)	.20	.10
Pr("innocent"/violation)	.56	.54
Pr("innocent"/no violation)	.80	.90
<i>Results</i>		
Probability of violation before observing evidence	.125	.143
False alarm rate at threshold of accusation, $x^*$	.238	.103
Hit probability at threshold of accusation $f(x^*)$	.467	.462
Observer's expected benefit before observing indicator	.133	.154
Inspectee's expected benefit before observing indicator	.067	.077

*Table 2.* A better test that induces a higher probability of violation, but a much lower false alarm rate.

## REFERENCES

- ACADEMY OF INTERSCIENCE METHODOLOGY. (1978). Simplification of Strategic Model Utilization, Model Maintenance and Mathematical Support. United States Arms Control and Disarmament Agency Research Report AC7VC110.
- ARMS CONTROL ASSOCIATION. (1987). Analysis of the President's Report on Soviet Non-compliance with Arms Control Agreements, Washington, D.C.: The Arms Control Association, March 12, 1987. In *Arms Control Today*, 17:2, 1987.
- AVENHAUS, R. (1977). *Material Accountability: Theory, Verification and Applications*, New York: Wiley.
- AVENHAUS, R. (1986). *Safeguard Systems Analysis*, New York: Plenum.
- AVENHAUS, R., and M. CANTY (1987). Game Theoretical Analysis of Safeguards Effectiveness: Attribute Sampling, Spezielle Berichte der Kernforschungsanlage Juelich, Report 417.
- AVENHAUS, R., J. FICHTNER AND G. VACHON. (1987). The Identification of Factors and Relationships Relevant to a Routine Random Inspection Procedure under a Future Chemical Weapons Convention and the IAEA Experience. (Mimeo) Arms Control and Disarmament Division, Department of External Affairs, Canada.
- AVENHAUS, R., AND H. FRICK. (1983). "Analyse von Fehlalarmen in Ueberwachungssystem mit Hilfe von Zweipersonen Nichtnullsummenspielen" (Analysis of false alarms in inspection systems with the help of two-person non-zero-sum games), *Operations Research Verfahren* 16:629-639.
- AVENHAUS, R., AND S. ZAMIR. (1988). Safeguards Games with Application to Material Control. Working Paper 12. Zentrum fur interdisziplinare Forschung. Bielefeld University.
- BELLANY, I. (1982). "An introduction to verification," *Arms Control*. 3:1-13.
- BERG, R. (1986). Suite Hypothesis Testing: Why? United States Arms Control and Disarmament Agency, Washington, D.C.
- BERG, R., AND W. DEEMER. (1977). The Application of Statistics to the Estimation of Yield of Soviet Nuclear Tests. United States Arms Control and Disarmament Agency.
- BIERLEIN, D. (1983). "Game-theoretical modelling of safeguards of different types of illegal activities." In *Proceedings of the Fourth Formator Symposium on Mathematical Methods in the Analysis of Large-scale Systems*. Prague: Czechoslovakian Academy of Science.
- BRAMS, S. (1985). *Superpower Games: Applying Game Theory to Superpower Conflict*. New Haven: Yale University Press.
- BRAMS S., AND M. DAVIS. (1987). "The verification problem in arms control: a game-theoretic analysis." In *Communication and Interactions in Global Politics*, edited by C. Cioffi-Revilla, R. Merritt and D. Zinnes. Beverly Hills: Sage.
- BRAMS, S., M. DAVIS AND M. KILGOUR. (1988). Optimal Cheating and Inspection Strategies under INF. (mimeo). Department of Politics, New York University.
- BRAMS, S., AND M. KILGOUR. (1986). "Notes on arms control verification: a game-theoretic analysis." In *Modelling and Analysis of Arms Control Problems*, edited by R. Avenhaus and R. Huber. Berlin: Springer-Verlag.
- BRAMS S., AND M. KILGOUR. (1987). "Verification and stability: a game theoretic analysis." In *Arms and Artificial Intelligence*, edited by A. M. Din. Oxford: Oxford University Press.
- BRAMS, S., AND M. KILGOUR. (1988). *Game Theory and International Security*, New York: Basil Blackwell.
- BROWN, H. Statement to Senate Foreign Relations Committee Hearings on the Ratification of SALT II, 1979.

- BRYSON, M. (1984). "A primer on Bayesian and classical statistics as applied to nuclear test ban compliance issues." In *A Review of Statistical Aspects of Threshold Test Ban Treaty Verification*. Defence Advanced Research Projects Agency, Arlington, Virginia.
- CIERVO, A. (1974). A Dynamic On-site Inspection Strategy for Test Ban Verification, Pacific-Sierra Research Corporation Report 409, Santa Monica.
- CIERVO, A., AND G. HALL. (1987). TTBT Compliance Tests: an Application of Empirical Bayes. Pacific-Sierra Research Corporation Report 1650, Santa Monica.
- CIERVO, A., G. HALL AND F. THOMAS. (1977). The Reduction of Bias in TTBT Yield Estimates, Pacific Sierra Research Corporation Report 771, Santa Monica.
- CIERVO, A., S. SANEMITSU, D. SNEAD, AND R. SUEY. (1980). A User's Manual for SNAP/D -- Seismic Network Assessment Program for Detection. Pacific-Sierra Research Corporation, Santa Monica.
- CIERVO, A., AND A. WATSON. (1983). Optimal Allocation of On-site Inspection for INF Treaty Verification, Pacific-Sierra Research Corporation Report 1337, Santa Monica.
- COHEN, M., D. SCHUM, A. FREELING AND J. CHINNIS. (1985). On the Art and Science of Hedging a Conclusion: Alternative Theories of Uncertainty in Intelligence Analysis. Decision Science Consortium, Inc., Technical Report 84-6, Falls Church, Virginia.
- COMMISSION ON INTEGRATED LONG-TERM STRATEGY. (1988). *Discriminate Deterrence*, Washington: U.S. Government Printing Office.
- DACEY, R. (1979). "Detection and disarmament." *International Studies Quarterly*. 23:589-598.
- DEVOLPI, A. (1987). Statistical Methods Applied to Treaty Verification, Paper presented at the Meeting of the American Physical Society, Crystal City, Virginia. (mimeo) Argonne National Laboratory.
- DOUGLAS, A. (1981). "Seismic source identification -- a review of past and present effort." In *Identification of Seismic Source -Earthquake or Underground Explosion*, edited by E. Husebye and S. Mykkelveit. Dordrecht: Reidel.
- DAHLMAN, O. (1977). *Monitoring Underground Nuclear Explosions*, New York: Elsevier.
- DAVIS, M. (1984). "Inspection against clandestine rearmament: approximate solutions and analysis." In *A Review of Statistical Aspects of the Threshold Test Ban Treaty Verification*, Defense Advanced Research Projects Agency, Arlington, Virginia.
- DRESHER, M. (1962). The Sampling Inspector Problem in Arms Control Agreements, RAND Corporation, RM-2972-ARPA, Santa Monica.
- DUFFY, G. 1988. *Compliance and the Future of Arms Control*. Stanford: Center for International Security and Arms Control.
- ERICSSON, U. (1970). "Event identification for test ban control," *Bulletin of the Seismological Society of America*. 60:1521-1546.
- FETTER, S. (1988). *Toward a Comprehensive Test Ban*. New York: Ballinger.
- FETTER, S. AND S.N. RODIONOV. (1991). "Verifying START." 95-122 in F. Calogero, M. Goldberger and S. Kapitza, eds. *Verification, Monitoring Disarmament*. Boulder: Westview Press.
- FICHTNER, J. (1985). "Statistische Tests zur Abschreckung von Fehlverhalten - Eine mathematische Analyse von Ueberwachungssystemen mit Anwendungen" (Statistical tests for the Deterrence of Illegal Behaviour, a Mathematical Analysis of Verification with Applications), Ph.D. Dissertation, Universitat der Bundeswehr, Munchen.

- FICHTNER, J. (1986). "Concepts for solving two-person games which model the verification problem in arms control." *Modelling and Analysis of Arms Control Problems*, edited by R. Avenhaus and R. Huber, pp.421-441. Berlin: Springer-Verlag.
- FILAR, J. (1983). The Travelling Inspector Model. Technical Report 374, Department of Mathematical Sciences, The Johns Hopkins University.
- FILAR, J., AND T. SCHULTZ. (1983). Interactive Solutions for the Travelling Inspector Model and Related Problems. Operations Research Group Report Series #83-06. Department of Mathematical Sciences, The Johns Hopkins University.
- GREEN, D., AND J. SWETS. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.
- HAFEMEISTER, D. (1986). "Breakout for arms control treaties: a sensitivity analysis of the national security threat." In *Arms Control Verification: the Technologies that Make it Possible*, edited by K. Tsipis, D. Hafemeister and P. Janeway, pp.44-62. New York: Pergamon-Brassey.
- HALL, D., W. NICHOLSON AND P. HEASLER. (1984). Application of Statistical Hypothesis Testing to Soviet Compliance with the Threshold Test Ban Treaty. Pacific-Northwest Laboratory, PNL-X-294, Santa Monica.
- HANNON, W. (1972). Some Comments on Probability Models Associated with Test Ban Control Using On-Site Inspections for Event Verification, Lawrence Livermore Laboratory, UCID-16160.
- HECKROTTE, W., AND P. MOULTHROUP. (1984). "The Probability of High Yield Events in the Soviet Test Program by the Use of Bayes Rule." In *A Review of Statistical Aspects of the Threshold Test Ban Treaty Verification*. DARPA, Arlington, Virginia.
- JAPAN DELEGATION, U.N. CONFERENCE ON DISARMAMENT. (1986). Some Quantitative Aspects of a Chemical Weapons Convention. Document CD/713, Geneva.
- JERVIS, R. (1976). *Perception and Misperception in International Politics*. Princeton: Princeton University Press.
- JUDD, B., S. STRAIT AND L. YOUNKER. (1986). Decision Analysis Framework for Evaluating CTBT Seismic Verification Options. Lawrence Livermore National Laboratory. UCID-20853.
- JUDD, B., L. YOUNKER, W. HANNON, S. STRAIT, P. MEAGHER, A. SICHERMAN. (1988). Decision Framework for Evaluating Compliance with the Threshold Test Ban Treaty. Lawrence Livermore National Laboratory. UCRL-53830.
- KARKOSZKA, A. (1977). *Strategic Disarmament, Verification and National Security*, New York: Crane.
- KATZ, A. (1980). "The fabric of verification: the warp and the woof." In *Verification and SALT*, edited by William Potter. Boulder: Westview.
- KILGOUR, M. (1990). "Optimal Cheating and Inspection Strategies under a Chemical Weapons Treaty." *INFOR*. 28:27-39
- KILGOUR, M., AND S. BRAMS. (1990) Arms Control Inspection Strategies that Induce Compliance. Mimeo. Department of Mathematics, Wilfred Laurier University.
- KOULIK, S. "SALT." Pp.191-202 in R. Kokoski and S. Koulik, eds. *Verification of Conventional Arms Control in Europe*." Boulder: Westview.
- KREPON, M. (1985). "The political dynamics of verification and compliance debates." In *Verification and Arms Control*, edited by William Potter, pp. 135-152. Lexington, MA: Lexington Books.
- KUHN, H. (1963). "Recursive inspection games," in *Mathematica, Inc., The Application of Statistical Methodology to Arms Control and Disarmament*. United States Arms Control and Disarmament Agency Report ST3.

- LEWIS, P. (1990). "Implementation of verification methods." Pp.168-190 in R. Kokoski and S. Koulik, eds. *Verification of Conventional Arms Control in Europe.* Boulder: Westview.
- LIEBERMAN, A. (1984). "Models in national policy analysis." In *Military Modeling*, edited by W. Hughes. Alexandria, Virginia: Military Operations Research Society of America.
- MASCHLER, M. (1966). "A price leadership model for solving the inspector's non-constant sum game." *Naval Research Logistics Quarterly*, 13:11-33. Also in Mathematica, Inc., *The Application of Statistical Methodology to Arms Control and Disarmament.* United States Arms Control and Disarmament Agency Report ST3, 1963.
- MASCHLER, M. (1967). "The inspector's non-constant sum game: its dependence on a system of detectors." *Naval Research Logistics Quarterly*, 14:275-290. Also in Mathematica, Inc., *The Application of Statistical Methodology to Arms Control and Disarmament*, United States Arms Control and Disarmament Agency Report ST37, 1965.
- MEYER, S. (1979). "Verification and the ICBM shell game," *International Security*, 4:40-65.
- MEYER, S. (1984). "Verification and risk in arms control." *International Security*, 8:111-126.
- MOGLEWER, S. (1973). *A Game-theoretical Approach to Auditing Inspection.* Douglas Paper 6165. McDonnell-Douglas Corporation. Long Beach, California.
- NETHERLANDS DELEGATION, U.N. CONFERENCE ON DISARMAMENT. (1984). *Size and Structure of a Chemical Disarmament Inspectorate*, Document CD/455, Geneva.
- NICHOLSON, W., AND P. HEASLER. (1984). *Bayesian Evaluation of the Seismic Yield Estimator Methodology.* Pacific-Northwest Laboratory. Richland, Washington.
- OFFICE OF THE PRESS SECRETARY, THE WHITE HOUSE. (1987). *The President's Unclassified Report on Soviet Noncompliance with Arms Control Agreements*, Washington, D.C., March 10, 1987.
- PATTERSON, R., AND W. RICHARDSON. (1963). "A decision-theoretic model for determining verification requirements." *Journal of Conflict Resolution*, 7:603-607.
- RAPOPORT, A. (1966). *Two-person Game Theory: the Essential Ideas*, Ann Arbor: University of Michigan Press.
- RICHELSON, J. (1979). "Multiple Aim Point Basing: Vulnerability and verification problems," *Journal of Conflict Resolution*, 23:613-628.
- ROSTOW, E. (1981). Statement, U.S. Senate Armed Services Committee, July 24, 1981. Reprinted in United States Arms Control and Disarmament Agency, *Documents on Disarmament.*
- SCHEAR, J. (1982). "Verifying arms agreements, premises, practices and future problems," *Arms Control*, 3:76-95.
- SCHEAR, J. (1989). "Verification compliance and arms control, the dynamics of the domestic debate." In *Nuclear Arguments*, edited by L. Eden and S. Miller. Ithaca: Cornell University Press.
- SHUMWAY, R., AND W. RIVERS. (1984). "Testing the hypothesis of TTBT compliance and magnitude-yield regression for explosions in granite." In *A Review of Statistical Aspects of the Threshold Test Ban Verification*, DARPA, Arlington, Virginia.
- SPETZLER, C., AND C.-A. STAEL VON HOLSTEIN. (1975). "Probability encoding in decision analysis." *Management Science*, 22:340-358.
- STRAIT, R., AND A. SICHERMAN. (1986). *Comprehensive Test Ban Treaty Seismic Verification Decision Analysis and Computer Model*, Lawrence Livermore National Laboratory, UCID-20704.

- TJOSTHEID, D. (1981). "Multidimensional discrimination techniques, theory and application." In *Identification of Seismic Source -- Earthquake or Underground Explosion*, edited by E. Husebye and S. Mykkelveit. Dordrecht: Reidel.
- ULVILA, J., AND R. BROWN. (1981). Development of decision analysis aids for non-proliferation safeguards, Decision Science Consortium, Falls Church, Virginia.
- WALTERS, R., AND K. ZINN. (1985). The September 22, 1979 Mystery Flash: Did South Africa Detonate a Nuclear Bomb? Washington: The Washington Office on Africa.
- WEICHERT, D., AND P. BASHAM. (1973). "Deterrence and false alarms in seismic discrimination." *Bulletin of the Seismological Society of America*, 63:1119-1133.
- WEISSENBERGER, S. (1990). Deterrence and the Design of Treaty Verification Systems. Lawrence Livermore National Laboratory. UCRL-JC-105810.
- WEISSENBERGER, S. (1991). Treaty Verification with an Uncertain Partner. Lawrence Livermore National Laboratory. UCRL-JC-105885.
- WESTERVELT, D., M. MACKAY AND M. BRYSON. (1984). TTBT: Testing the Hypothesis of Soviet Compliance. LA-9852-MS. Los Alamos National Laboratory, Los Alamos, New Mexico.
- WIESNER, J. (1961). "Inspection for disarmament." *Arms Control Issues for the Public*, edited by L. Henkin. New York: Columbia University Press.
- WITTEKINDT, R. (1984). "Verification of MBFR Agreements - a systems analysis." In *Quantitative Assessment in Arms Control*, edited by R. Avenhaus and R. Huber. New York: Plenum.
- WITTMAN, D. (1989). "Arms control verification and other games involving imperfect detection." *American Political Science Review*, 83:923-945.
- YORK, H. (1970). *Race to Oblivion*. New York: Simon and Schuster.
- YOUNKER, L., J. HANNON, D. SPRINGER, R. AL-AYAT, B. JUDD, P. MORRIS, J. SANDLING. (1985). Evaluation Framework for Comprehensive Test Ban Treaty Seismic Verification. Lawrence Livermore National Laboratories.
- ZAMIR, S. (1987). A Two-period Material Safeguards Game. (Mimeo). Department of Economics, University of Pittsburgh.

who is the director of the ACDA and what has he written on verification

Three references on decision theory

Brams, D.J. Morton D. Davis and D. Marc Kilgour Optimal  
Cheating and Inspection Strategies under INF. Mimeo 1988

Figure out whether the appendix really answers the ocean testing question

Figure out the right figure number

Look up the Dutch book in Steve fetters piece and add it

Add idea of many technological studies versus systems analysis

three ideas of new verification problems

washington times report on projected production

seeing rockets out of area

also standard problems of mobile land based missiles and cruise missiles.

Do I want to make up a new diagram for the boxes using columns