Risk Analysis, Vol. 30, No. 4, 2010 DOI: 10.1111/j.1539-6924.2010.01370.x

Deceptive Detection Methods for Effective Security with Inadequate Budgets: The Testing Power Index

Paul Kantor* and Endre Boros

Detection of contraband depends on countermeasures, some of which involve examining cargo containers and/or their associated documents. Document screening is the least expensive, physical methods, such as gamma ray detection are more expensive, and definitive manual unpacking is most expensive. We cannot apply the full array of methods to all incoming cargoes, for budgetary reasons. We study the problem using principles of game theory, and find that best detection rates are achieved when the available budget is allocated between screening and definitive unpacking using a mixture of strategies that maximize detection rate and, further, serve to deceive opponents as to the specific tests to which contraband will be subjected. This yields increases of as much as 100% in detection, with essentially no increase in inspection cost.

KEY WORDS: Homeland security; mixed strategies; nuclear detection; screening; testing

1. INTRODUCTION

Securing our ports and harbors where container traffic enters the country is a major component of protecting our homeland. The possibility that terrorists might introduce catastrophic weapons, such as chemical or biological agents or nuclear devices, requires high vigilance. Inspection of containers involves inexpensive screening (document checks), more expensive scanning tests, and expensive unpacking procedures. The definitive examination of each container (that is unpacking it and exploring it) is prohibitively expensive. Many testing methods have been proposed for deciding which containers to unpack. Scanning procedures based on documents are relatively inexpensive. Other screens involve more expensive use of radiation detectors, or imaging tests. Estimates[1] have been made of the costs of providing totally secure inspection. This goal seems out of reach even for container traffic through ports and harbors. When we consider the further problem of protecting cities and high value industrial, cultural, and human targets, perfect "testing" is almost surely not cost effective.

As noted, previous work has distinguished among "scans," "screens," and "unpacking." For clarity in presentation we consider only two steps here, which we take to be "testing" and "unpacking." Thus, what we call "testing" may include some examples of what are elsewhere called "scans" and "screens." It is also important to note that our approach will consider not only the cost of each of the tests to be used, but also the cost of definitive unpacking of a container deemed "highly suspicious." We consider that this expense and the expense of the testing activity are comparable, and managed by the same decisionmakers. The best strategies will adjust the thresholds or decision rules used with later tests, to exploit information learned in earlier tests. Related work is reviewed in References 1 and 2. In particular, Saeger and Stroud[3] have assumed that each test is used to make a binary decision, so that the overall problem of developing a testing strategy

Rutgers University
*Address correspondence to Paul Kantor; tel: 732 932 7500 ext 8216; fax: 732 932 1504; paul.kantor@rutgers.edu.

becomes a binary decision tree problem. Continuing in this vein, Madigan *et al.*[4] implemented an efficient local search algorithm to identify the best binary decision tree, with a nonlinear search included to select the best thresholds for each test, at each point in the tree. Boros *et al.*[2] have removed the binary restriction, and have reformulated the problem as a very large linear programming problem, for the case of tests that produce continuous readings representing the risk, or some other characteristic.

Related issues arise in the study of passenger screening for airport security, as discussed by McLay *et al.*,[5,6] who have formulated integer programming models of the problem. Virta, Jacobson, and Kobza[7] and Jacobson, Karnani, and Kobza[8] recognized the benefits of reducing the use of expensive passenger baggage testing at airports by employing much cheaper and faster prescreening tests. While they argue that the *effectiveness* of the prescreening procedure is the most dominant factor in the effectiveness of the entire system, we find, by considering the costs of testing and of unpacking, that the best choice of testing methods depends on a *complex relation between effectiveness and cost*. We find that achieving the benefit may require an element of randomness in preselecting, prior to the testing itself.

All of these methods involve considerable computational sophistication, and yet they all find ways to place the problem within the reach of modern optimization software and powerful contemporary computers. The present article addresses a remarkable simplification arising in the all-too-likely event that the available budgets are simply inadequate to implement the optimal strategies found by these methods.

Remarkably, we find that the cost effectiveness of any particular screening test may be summarized by a single number, which we call the *Testing Power Index* (TPI). This index depends on the sensitivity and specificity (or operating characteristics) of the test, on known cost information, and on estimated probabilities. These numbers are, in reality, not widely available and in some cases held confidential. However, once they are known, it is easy to compute the index. The method can therefore be applied by operators of terminals, and sensitive information need not be shared with researchers. The TPI applies precisely when budget limitations dictate that *not all containers can be tested*. In this situation randomization strategies will improve the detection rate.[2] Such an approach, in the terminology of game theory, is called a *mixed strategy*. In addition to its optimality properties, it has the virtue of deception, as, properly implemented, it thwarts an opponent's efforts to circumvent it.

Table I. Probability that a Container of the Given Type Will be Flagged

	Contraband	Innocent
Flagged	d	f
Not flagged	$1-d$	$1-f$

1.1 Simple Tests have Fixed Thresholds

The results of this article stem from the fact that screening tests (such as the documentation, shipper, country of origin, factory of origin, etc.) and scanning tests (such as handheld radiation detectors, or imaging systems) are not perfect indicators of whether a container should be unpacked. This uncertainty is expressed as the probability that a particular test will "flag" a container with contraband and the probability that it will flag an innocent container, as shown in Table I. The number represented by d is the "conditional probability of detection," while f is the "conditional probability of false alarm" (and, as probabilities, these numbers must be between 0 and 1).

The information given in Table I may be summarized in a graph of the "ROC" or "(Radar|Receiver|Relative) Operating Characteristic"[9–11] as shown in Fig. 1(a). The case of a fixed threshold (cf. Figs. 1(a) and (b)), will clarify the key ideas. It is extremely important to understand that the linear segments in a curve such as Fig. 1(a) are determined by the one operating point $(f,d) = (0.1,0,8)$ but they are not merely drawn "to guide the eye." Rather, each point on the linear segments represents a level of performance that can be achieved by *randomly mixing* the operating point (f,d) with either of two "zero-cost" strategies. For the lower linear segment the operating point is mixed with the point (0,0), which simply means "flag everything as harmless, without even looking at it." And for the higher segment, the mixing is with the strategy that means "flag everything as a threat, without even looking at it." Technically, any intermediate point can be achieved by the proper random strategy. For example, if 70% of all the items are tested and 30% are simply flagged as harmless, then the detection rate will be $70\% \times 80\% = 56\%$, and the false alarm rate will be $70\% \times 10\% = 7\%$. It is easy to show that this point lies on the lower linear segment in

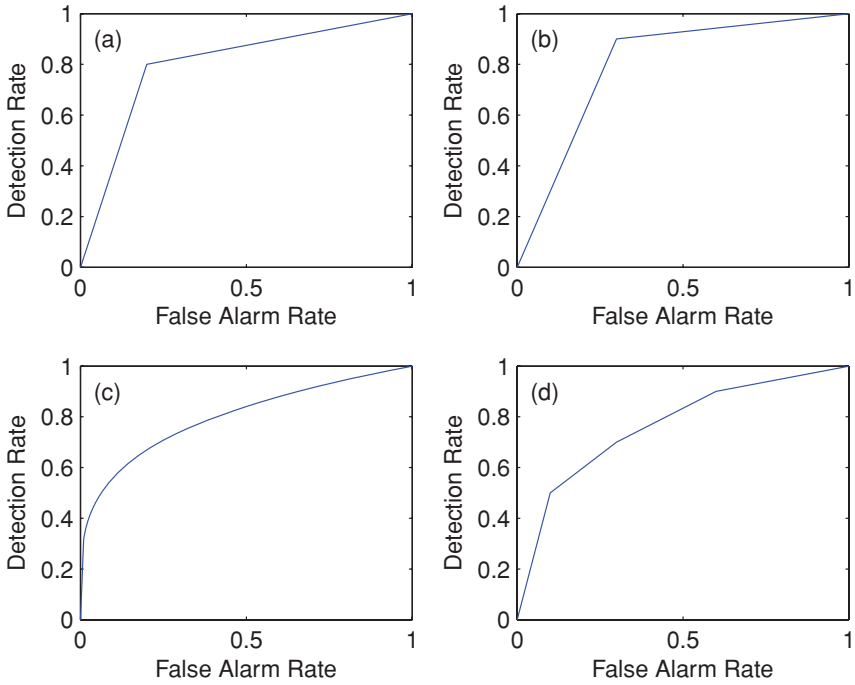

Fig. 1. The operating characteristic (ROC) shows the fraction of dangerous cargo detected, as a function of the fraction of harmless cargo that is tagged for costly inspection. If the system has a single threshold fixed in advance, the curve has two linear segments. Several models provide for a continuous dependence of detection on false alarms, as a threshold is varied. More complex discrete processes, such as document checks, may yield a curve with multiple linear segments. Operating characteristics are shown for sensors with: (a) ($f = 0.1$, $d = 0.8$) ; (b) ($f = 0.2$, $d = 0.9$); (c) $d = f^{0.25}$; (d) $\{(d,f)\} = \{(0.1, 0.5); (0.3, 0.7); (0.6, 0.9)$. The line segments represent mixed strategies (see text).

Fig. 1(a). A corresponding argument produces the upper linear segment, by mixing with the other zero-cost strategy.

Fixed thresholds arise when the screening test has no adjustable parameters. Optimization in this case depends on three costs: (a) the cost, C_T, of testing all the containers arriving in some agreed upon unit of time such as a month or a year; (b) the (much larger) cost C_U of unpacking all those containers, to verify that they contain no contraband; (c) the cost of impact on commerce, C_I, of unpacking (and thus delaying) all the containers. For a complete analysis we also need to know how many are expected to fall into each of the two categories: harmful (or contraband) and harmless. The *a priori* probability of a container being harmful is represented by π.

In principle, we would like to be able to test everything, and then open all the items that are flagged as suspicious.[1] The resulting total cost of *operation* b^* (per budget period) can be computed from the above parameters, as given by Equation (1):

$$b_T^* = C_T + C_U \pi d + (C_U + C_I)(1 - \pi) f. \qquad (1)$$

[1] While it might be argued that it would be even more desirable to completely unpack every container arriving on our shores, that is clearly not feasible,

We refer to b_T^* as the "natural operating budget" required to fully utilize the specific test T. The first term (C_T) is the cost of testing everything with test T. The second term is the expected or average cost of unpacking the detected dangerous containers, and the third term is the (wasted) expense of unpacking and delaying harmless containers. Note that if there were only enough money to do the tests, and not enough money to do any unpacking, it would be foolish to do the tests, and there would be no way to act on the information. Alternatively, one might propose to begin the fiscal year by testing everything, and unpacking all the suspicious items each day. Then, one day, there will be no money left for testing or unpacking. The opponent[12] would of course wait for that day to occur, and would attack some time during the remainder of the fiscal year.

It is common in engineering analyses to complete the description by adding a fourth term $+C_B \pi (1 - d)$ representing the expected cost of not detecting a dangerous content, where C_B denotes the cost of a smuggled nuclear device being blown up inside our country. The true value of C_B is, however, highly speculative, and as the 9/11 attacks on the United States in 2001 show, it is simply impossible to cast into a single number the true costs to the economy, to society, and to our future. It is clear, in any case,

that this fourth term would dominate our cost analysis, no matter how low we (reasonably) set C_B.

Thus, our best strategy for minimizing the combined total cost is to maximize the detection rate d, which leaves us with the same goal, regardless of the specific (enormous) value of C_B. Therefore, in subsequent steps of our analysis we do not explicitly represent the unknowably large fourth term as part of the cost. By cost we will henceforth mean the "real" cost of inspection as expressed by Equation (1). Subject to available funds for this activity, we concentrate on maximizing the detection rate d.

While we speak of a single "test T," in fact T could be a complex procedure involving a number of different testing technologies, in an array of complex inspection policies. For all useful testing technologies we can assume that the testing cost is substantially smaller than the ultimate inspection involving the manual unpacking of a container. Thus, in our cost analysis, Equation (1), we separate the "test T" from the unpacking, that is, we consider an inspection strategy as a two-component procedure, involving a (possibly complex) test T used to "flag" some of the containers, followed by unpacking of the "flagged" ones.

Another way to think of Equation (1) is to say that as we have more money to spend, we can increase the detection rate, but, because of the ROC curve, we must correspondingly increase the false alarm rate. In this way both the detection rate d and rate of false alarm f, as we argue below, can be viewed as functions of our real budget b. Our subsequent focus is on the realistic scenario where we have a fairly powerful testing policy T, whose cost b^*_T exceeds our available budget b. In this situation we can try to stay below our budget b by applying some components of our inspection policy to only some of the containers, producing various operating characteristics for the achievable detection rate as a function of our budget b. Some corresponding curves are shown in Fig. 2 for the specific choices of $C_T = 3, C_U = 16, C_I = 6, \pi = 1 \times 10^{-6}$, and the performance $(f,d) = (0.1,0.8)$ as given in Fig. 1(a).

Fig. 2(a) shows the result of first testing everything, and then applying the remaining funds to unpack as many of the items as we can. In these scenarios, all the items that come in each hour (say) are tested. One of the possibilities is to manually unpack only a fraction of the flagged containers. The corresponding achievable detection rate as a function of the real budget b is shown in Fig. 2(a). The striking feature of this figure is that until the amount C_T has been spent, there is *no detection at all* (i.e., for $b < C_T$

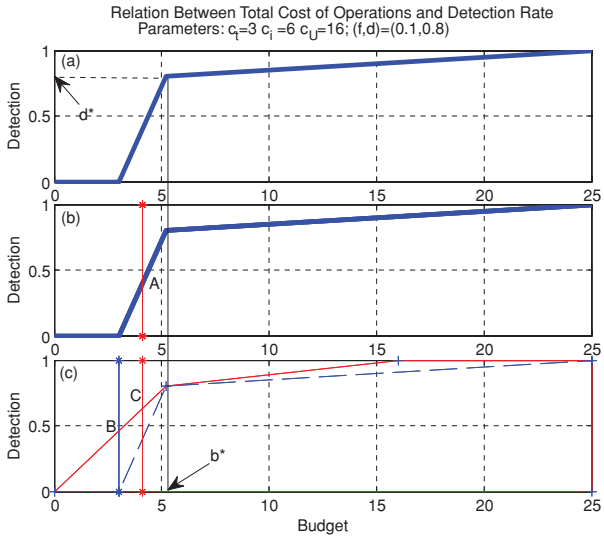

Fig. 2. Plots of the relation between the fraction of all dangerous cargoes detected, and the budget. Parameters are set at $(f,d) = (0.1, 0.8)$, using the detection scheme of Fig. 1(a). (a) For budgets less than C_T there is no detection at all. (b) For budgets between C_T and b^*, such as the solid vertical line through the points marked A, random selection of containers flagged by the test will produce the detection level shown by A. However, randomly selecting containers to be tested, and then opening all of the ones that are flagged, will produce a higher detection level shown by B. (c) This principle is even more effective when the budget C is less than C_T, and the mixed strategy produces substantial non zero detection as shown by D.

we cannot guarantee any detection because we cannot afford to do anything about the items that have been flagged). If we spent so much on testing that we cannot afford to unpack all the flagged items then we have to use a random choice on which ones we will unpack, so that the adversary cannot know in advance which ones we will not unpack. We do not stop the threats in the ones that are released, which is why the true detection is lower than the natural performance of the test itself. An example is shown by the point A in Fig. 2(b).

We have, however, better possibilities, using ideas from game theory. To provide a realistic scale, we draw on reports that current port operations in the United States are said to open or examine between 6 and 9% of all incoming containers.[1,13] With the assumed performance numbers, this corresponds to a real budget b somewhere between C_T and b^*_T. In this range, one may randomly select from among the flagged containers until the funds are exhausted. This brings performance to the point labeled A in Fig. 2(b). For example, if we would randomly examine half of the flagged containers, we would

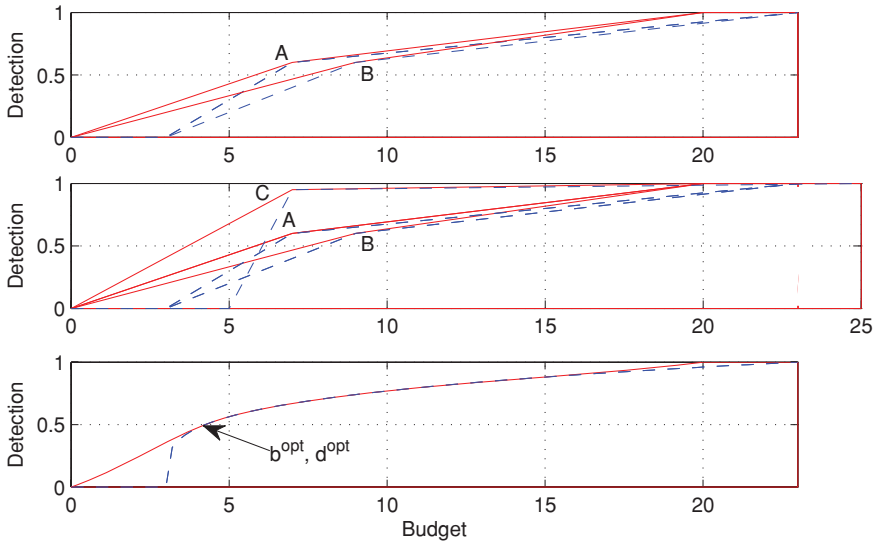

Fig. 3. Systems can be compared by examining the slope of the line from the origin to the "top" of the curve. The slope of the highest line represents the Testing Power Index (TPI). For curves with a single "corner" the coordinates of the corner are (b^*, d^*). (a) Test T is therefore better, for inadequate budgets, than is Test S. (b) But test B, which is much more expensive, is in fact better than either of them. (c) The same principle applies to operating characteristics such as the power law form shown here, but the optimal point is found either analytically or by numerical search. For a smooth curve (corresponding to a variable threshold), the point (b^*, d^*) is called (b_{opt}, d_{opt}) to emphasize that it is determined by computing the slope of the line to the origin, from every point on the curve, and selecting the best. The needed one-dimensional search is highly efficient in every case, as the TPI is a concave function of the budget.

find, on average, half of the contraband contained in all the flagged containers. Note that randomization (rather than systematic rules) provides an essential element of deception and is a key to the success of this strategy. If, for example, we only opened flagged containers at a particular terminal "on Mondays, Wednesdays, and Fridays" the rule would eventually be discovered by our opponents, who would then make every effort to arrive on the other days.

In fact, by recalling the meaning of the linear segments in the ROC plot, we see that it is possible to do even better, by adopting a mixed or randomized strategy even before applying the test. Thus, if we randomly select containers to be *tested*, we are now mixing the natural operating point of the test with the zero-cost option "flag everything harmless." This yields, at the same budget, a higher detection rate, corresponding to the point in Fig. 2(c) labeled C. This principle clearly applies with even more force to budgets less than C_T, where the naïve alternative would be to detect *nothing*. The mixing principle is shown by the straight line in Fig. 2(c), which connects the origin to the *natural operating point* $d^*(b_T^*)$. (In what follows, and in Fig. 2, we write b^* for b_T^* points on this straight line to represent the effectiveness and budget of *mixed testing strategies.*) For example, the point marked B is about 50% of the way from the origin to the optimal point. It represents a strategy in which only 50% of the containers are tested at all, while the others are untested. If the reading for a tested container falls above the threshold, it is unpacked. As is apparent from the geometry of Fig. 2(c), the cost of this strategy is $50\% b_T^*$, and the corresponding detection rate is $50\% d^*$. There is no way to do better at this budget level because the only available strategies will fall on lines connecting the points (0,0), (d,f) and (1,1). It is clear from the diagram that the highest value for d is on the first linear segment, and is $50\% b^*$.[2]

1.2 Tests Themselves May be Compared Using a Testing Power Index

Before estimating how much this approach might improve detection performance, let us see how *different* tests can be easily compared. Suppose there is an entirely different testing strategy (S), with its own ROC curve, as shown in Fig. 1(b), and its own cost C_S. We can perform exactly the same calculation, and obtain the corresponding point (d_S^*, b_S^*). The two lines connecting those points to the origin are strictly ordered (i.e., one is less steep than the other).

[2] Technically, the objective, detection, is a piecewise linear function of the cost of the procedure, which is the convex hull of the zero-cost strategies and the test strategy.

This order is determined by the ratios $\{d_X^*/b_X^*\}_{X=S,T}$, where X represents the particular testing method (see Fig. 3(a)). This ratio, as the construction makes clear, depends on several factors: the cost of testing, the cost of unpacking and interruption to commerce, the performance of the test, and the prior estimate of the probability of a harmful container.

Without revealing any sensitive information to researchers, a terminal operator may readily calculate this critical ratio, which we call the "TPI." This can be done for any method of testing whose cost and performance are known. Comparison of the TPI shows which method provides the highest level of protection, *for* **every** *"inadequate" budget*. The choice among testing protocols is then guided, in a rational way, by a calculation that is readily performed, and readily explained. This way of approaching the calculation makes the decision among testing methods, when budgets are inadequate, "as simple as possible, but not simpler."

As an example, with the values as shown in Fig. 3(b), the testing method B is clearly preferred to the other two methods, even though its "natural" operating expense would be the highest of the three, and well beyond the available funds.

1.3. For More Complex Tests, a Threshold Can be Selected Optimally

For more complex testing methods, such as those involving radiation detectors, both f and d can be decreased (or increased) by raising (or lowering) a parameter called the *threshold.* In the limits, one either unpacks none of the containers, or all of them. A typical continuous f–d relation is shown in Fig. 1(c). It might be determined from physical principles (e.g., by Monte Carle calculation of how much radiation leaks from a container, at various frequencies, with various kinds of shielding) or by experiments on a variety of container types.

An ROC with several linear portions might, on the other hand, result from empirical data (e.g., by learning what fraction of the containers coming from country X, and having documents Y, contain a contraband of interest). We are not able to work with real data in this report, but we illustrate the principles using model forms for the ROC. Fig. 1(c) shows an ROC produced by assuming that signals follow an exponential distribution. Since the detection and false alarm rates may now depend on a threshold t we write them as $(f(t), d(t))$. The choices for t may vary smoothly (for example, click rates on a radiation detector) or may fall at a few discrete points (for example, combinations of document check information). Fig. 1(d) shows a model example in which there are only a few possible test results (the "discrete" case).

At any threshold setting, the cost of operation (per budget period) is calculated, as before, as the cost of testing plus the cost of unpacking, plus the cost of interruption to commerce.

$$b(t) = C_T + C_U \pi d(t) + (C_U + C_I)(1 - \pi) f(t) \qquad (2)$$

Since the same parameter t determines both the detection rate, $d(t)$, and the budget required, $b(t)$, and each of them is monotonically dependent on the threshold, we can combine the calculations and draw a single curve relating b and d. Because the points $(f(t), d(t))$ form a monotone increasing curve as a function of the threshold t, each possible cost value b is attained for a unique point (f, d) on this ROC curve. We can begin with the points $(f(t), d(t))$ for all $t \in T$ where T is the set of possible thresholds. Each value of t determines a unique cost, by Equation (2). Thus, as t varies across the set T, it determines a set of pairs $(b(t), d(t))$ that implicitly define the $d(b)$ curve. Note that Equation (2) is linear in $d(t)$ and $f(t)$ so that interpolation in the $d(b)$ curve is also linear. We have used these facts in the intuitive presentation of Section 1.1 above.

Technically, we can say that: the cost detection curve consists of all the pairs (x, y) such that the detection $y \in [0, 1]$ and the cost x is given by the condition: $\exists$ *false alarm rate*, $u : (u, y) \in$ *ROC and* $x = C_T + C_U \pi y + (C_U + C_I)(1 - \pi)u$.

In other words, for every point (u,y) on the ROC curve we can imagine a corresponding policy, the expected cost of which is $x = C_T + C_U \pi y + (C_U + C_I)(1 - \pi)u$. Changing u from 0 to 1, we see that the corresponding (x,y) values, as computed above, form a concave curve, which we call the cost-detection curve. With this fact established, we no longer need to explicitly indicate the corresponding value of the threshold, t.

An example is shown in Fig. 3(c). This curve gives precisely the information that a policymaker needs. The goal is to *minimize* the risk of missing something dangerous (that is to *maximize* d) and the constraint is the amount of money that can be spent. Note that there could be other constraints as well, in which case the same principles used here apply, but the calculation, while straightforward (see, e.g., Boros *et al.*[2]), becomes multi-dimensional and cannot be easily represented by graphs.

Recalling that we do not have adequate budget to move arbitrarily far up the cost-detection curve,

we need to select a threshold. At any point on the curve, the slope of the line to the origin is given by $d(b)/b$. We choose the operating point so that the TPI is a maximum, and this defines the *optimal testing power point* on the $d(b)$ curve, with coordinates (b^{opt}, d^{opt}). To deal with any budget that is less than the budget required to reach that operating point, we use a mixed strategy, represented by connecting the origin to that point by a straight line.

The line is chosen so that no other line, from the origin to any point on the $d(b)$ curve, has a greater rate of increase. We can express this mathematically by saying that b^{opt} is the unique value of the budget that maximizes the TPI $d(b)/b$, where b is given by Equation (2). Recognizing that we can go from b to d, and thence (using the ROC curve) to f, we see that we can express both d and f as functions of b. Finally, using Equation (1), we find:

$$b^{opt} = \arg\max_b \{d(b)/(C_T + C_U \pi d(b) + (C_U + C_I)(1-\pi) f(b))\}. \quad (3)$$

We note that the selection of this point requires consideration of all of the cost parameters, and the presumed prior probability of contraband. We note also that in all cases of interest, the operating point for a test with a single fixed threshold is also the optimal operating point. (Otherwise, it would simply be better to open containers at random!)

In spite of the apparent complexity of the mathematics used to define it, the TPI has a simple interpretation. As long as the budget is not adequate to screen **all** containers, the TPI is the increase in the percentage (or fraction) of contraband that is detected, for a unit increase in the funds available for testing and inspecting, combined. In the preceding discussion we have measured funds in artificial units such that 25 is the cost of inspecting everything. With complete inspection, of course, preliminary testing and selection would not be needed at all.

A key finding of this work, then, is that there is a precise mathematical procedure to select the thresholds, which is based on the properties of the tests, the several costs of operation and interference, and the prior estimate of the chance of contraband. But this calculation does not require us to assign a value to potential catastrophes.

In practice, an operating point is sometimes selected by convention (for example, that the rate of missing dangerous cargoes (false negatives) be equal to the rate of false positives). In other studies it is selected by hypothesizing the relative cost of false negatives and false positives. The points selected in these ways are unlikely to be the correct point (b^{opt}, d^{opt}) *needed to optimize the effect of an inadequate budget.*

1.4 Complex Tests are Compared Using the Testing Power Index

As before, an entirely different testing strategy (S) will have its own ROC curve, and its own cost C_S. Repeating the calculations yields the corresponding curve $d_S(b)$. When they are examined together we may see that the curves cross more than once, in which case neither is "absolutely better than the other." However, each has its own unique optimal point b^* and two lines connecting those two points to the origin are strictly ordered as determined by the ratio $d_x(b^*)/b_x^*$ (where X labels the particular testing strategy). Thus tests with an adjustable threshold can be separately tuned so that each one gives its own best performance, and they can then also be ranked to select the highest level of protection, for a given "inadequate" budget.

2. IMPACT OF EFFECTIVENESS OF TESTING

2.1 Implications for Actual Port Security May be Quite Significant

Consider a testing method has a $f = 0.20$ $d = 0.80$ performance, which is generally considered a good level. This means that the single operating point (b^*, d^*) will have coordinates $(b^*, d^*) = (0.8C_U\pi + 0.2(C_U + C_I)(1-\pi), 0.8)$. Since π is quite small for many serious threats we can neglect it for purposes of discussion. Then $(b^*, d^*) = (0.2(C_U + C_I), 0.8)$. Now, with standard approaches, one tests every single container, and then unpacks until the budget is exhausted. If, for example, we are currently unpacking, say, 7% of all containers, two things may be happening. First, we may be adjusting our definition of "high threat" to conform to our budgetary capabilities. This may have been done by adjusting the thresholds applied, so that the operating point is somewhere around $(f, d) = (0.07, 0.28)$. In other words, 28% of the dangerous containers are being unpacked. However, with a mixed strategy we could achieve a higher level of detection. Suppose that the cost of unpacking is 20 times the cost of testing. Then the mixed strategy achieves a performance given by:

$$d_{optimal}(b_{now}) = 0.8(1 + 20 * .07)/(1 + 20 * .20) = 0.8 \times 2.4/5 = 0.384. \quad (4)$$

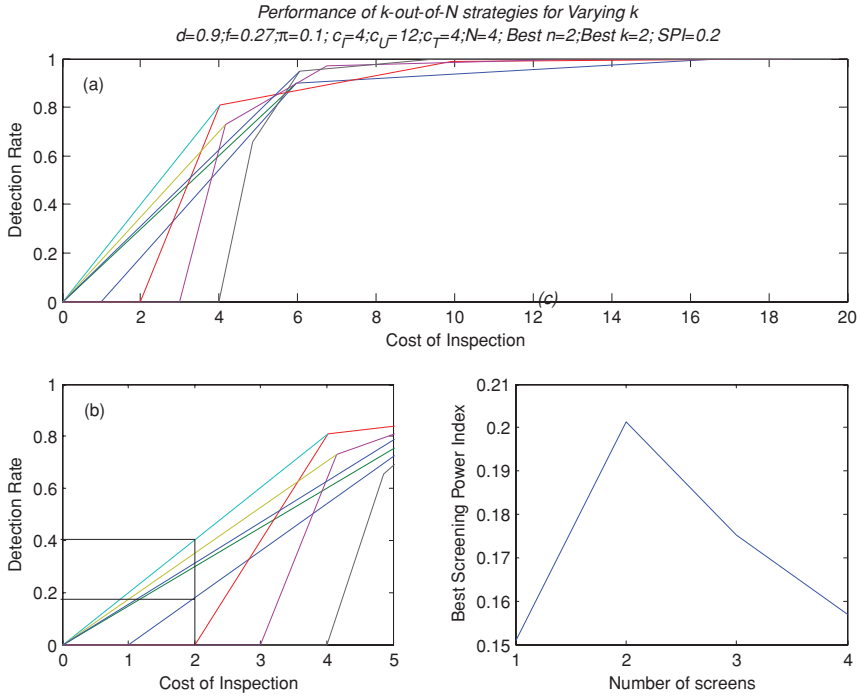

Fig. 4. We consider the situation where 4 independent tests, with the same operating characteristic, are available. The operator may choose to use 1,2, 3, or all 4 of the tests. For each of these choices (which corresponds to a curve in (a)), there is a specific optimal value of k, the number of positives needed to trigger an unpacking operation. (b) This is selected by comparing the tangent lines from the origin, shown in an expanded view. (c) Changing the x-axis, the dependence of the best achievable TPI is shown as a function of the number of tests deployed. In this case the line guides the eye. The TPI has a clear maximum at $N = 2$. Compared to using exactly two tests, using fewer tests would decrease detection by almost 30% and increasing it would provide no gain, but approximately a 15% decrease in TPI performance.

This is an increase in detection from 28% to 38% (that is, a nearly 35% greater detection rate) at absolutely no increase in cost!

2.2. Multiple Tests May be Reduced to an Optimal Subset: k-Out-of-*N* Tests

Suppose that there are N different document checks available, which each have their own performance characteristics (f_i, d_i) $i = 1, \ldots N$. Each of the tests may flag a container for unpacking. The terminal operator should use all of this information to decide which ones to unpack. For simplicity of presentation, we examine the case in which all the tests are *stochastically independent*, and all have *the same performance* characteristic. In this case the probability that k of the N tests will flag a container depends only on the number k and not on *which particular tests* flagged that container. Since for independent tests the conditional probabilities can be computed as products, we have, for the overall test *k-out-of-N:*

$$f^{k-out-of-N} = \binom{N}{k} f^k (1-f)^{N-k}$$

$$d^{k-out-of-N} = \binom{N}{k} d^k (1-d)^{N-k}. \qquad (5)$$

This is computed for the example values $C_T = 1; c_U = 12; c_I = 4; f = 0.27; d = 0.9; \pi = 0.1;$ and is shown in Fig. 4.

What does this have to do with practical port security? It is known that at present some 6% to 9% of containers are tested, and this is usually described by saying "all the highly suspicious containers are checked." We conjecture that what happens, in practice, is that the threshold (in this example, the value of k) is set so that the resulting operations are possible within the available budget.

We have shown here, by example, how much improvement may be possible with a mixed strategy. We also have shown that, even though a number N of inexpensive tests may be available, it may be optimal to pay for only some smaller number of them, in order to achieve the optimal TPI. In this example (see Fig. 4 caption for details) it is best to use just two of the four available tests.

2.3 Parameter Uncertainties Will Affect the Decision Process

The technical exposition of the TPI is complete. However, the real world may be less cooperative than we have assumed in our model. To begin with, there will be some uncertainty in the underlying performance characteristic, $(f(t), d(t))$. This can be explored using a variety of methods (some of which may simply be expert judgments) to come up with a band of *plausible ROC curves.* In addition, there is sure to be uncertainty about the value of the *a priori*

probability π. This can also be factored into the analysis, and the result will finally be some *family of cost-performance* curves corresponding to each method of testing. Each such curve will have an optimal point, and the resulting TPI values will span some range.

Now, consider any two competing tests. In some cases the range of (uncertain) TPI values for one will always be higher than the range of TPI values for the other. In such cases, the choice is clear. In some cases the ranges will overlap. There are no universally accepted rules for making the decision in this case. There is some preference for the range that reaches higher, but that choice would require careful scrutiny of the specific assumptions about performance, costs, and the *a priori* probability that combine to make it reach these highest values. In a competitive economic or political environment, advocates can be expected to argue vigorously about these "edge cases."

Because there is a rigorous theory underlying the computation of the TPI, however, it will be difficult to "game the decision." The underlying ROC arises from detailed technical or empirical data, and is hard to manipulate. Similarly, although *proposed* costs for military and security systems exhibit a degree of fantasy, insofar as there are any *real* performance data (f, d), there must have been real systems built and operated, so that the costs should be known.

Finally, of course, it may be that the ranges of TPI factors for the two systems essentially coincide. In this case, the choice of systems may be determined by other principles. The first is the *deception principle*, used above. Insofar as one can conceal which of the two tests is in use, it becomes more difficult for an opponent to use countermeasures. The second is the *robustness principle*, which favors using more than one approach, in anticipation of the possibility that the opponent may discover ways to reduce the effectiveness of any given system. From the political point of view one also notes the *distributional principle*, which favors having some part of the overall effort "*made in each congressional district.*"

3. CONCLUSIONS

3.1 Optimal Strategies Balance Testing and Expensive Unpacking to Maximize Detection with Inadequate Budgets

In summary, for any test, or combinations of tests, there is a corresponding operating characteristic. This may be combined with known cost information, and estimated probabilities, to determine a curve representing detection as a function of budget. This entire cost-detection performance curve may be summarized by a single number, the Testing Power Index, which corresponds to an "optimal" operating point on the curve. Finally, the available budget is balanced according to the best random strategy, by either testing or ignoring containers. The result can be substantial increases in the detection of contraband, with essentially no increase in costs for testing or inspection. We can think of these mixed strategies as shifting a certain portion of the budget from testing to unpacking, resulting in a higher overall detection rate.

3.2 To Protect Multiple Ports Extensions May be Needed

To protect the entire nation, one would need the *a priori* probability that containers entering via different ports contain contraband. These prior probabilities, π_P, for the port p, will affect the best detection strategy. If our assessment of these prior probabilities could be kept confidential, and could not reasonably be worked out by an opponent, then the optimal strategy concentrates our resources first on the port with the highest *a priori* risk. This is analogous to the problem of assigning risk reduction funds to cities or to states.

On the other hand, if this ranking could reasonably be estimated by an opponent, then the opponent would concentrate efforts precisely where we have estimated them to be least likely. The resulting situation can be treated in first approximation as a "zero-sum" game. That is, we seek to maximize detection, and the opponent seeks to avoid detection. These extensions of the problem will be discussed elsewhere.

3.3 Limitations of This Analysis

We close by noting a few limitations of this analysis. On the technical side, we have assumed (when considering multiple tests) that the tests are stochastically independent. While this is a widely used assumption (as in the literature cited above) it is not likely to be true. Thus analyses making this assumption should be regarded as "first-order approximations" to be improved upon as better algorithms and better information become available.

In framing the problem, we have also assumed that the impact of our detection policy is linear. However, the cost, to an opponent, is roughly proportional to $1/(1-d)$, since that is the expected

number of attempts required to achieve one success. As the detection rate approaches 1, this grows very rapidly. It is quite possible that there is a threshold effect, and that at some quite achievable point, the cost becomes prohibitive for the opponent and *he or she does not even try.* If such a threshold exists, then increasing d above that threshold level is merely wasteful. A related, but different suggestion is that increasing the risks of detection in some settings (where individuals are being searched) may affect the "upstream" terrorist recruiting efforts, as groups will seek out sympathizers with "harmless profiles." This has been used to argue for the deterrent effect of more detailed inspection (equivalent to our unpacking) of airline travelers who are not flagged as suspicious.[14] An analogous argument can be made with respect to reliance on document checks as tests that might release a container from the chain of inspections.

The TPI is clearly limited to the linear regime in the cost-detection curve, and if budgets were large enough (or technologies sufficiently inexpensive) one can imagine that the full analysis of the problem would move past the point that we have called "optimal" and into the region of decreasing returns to scale. This problem can be addressed with modern techniques of operations research,[2] but we believe that the present situation is one of inadequate resources and therefore in the linear region.

Finally, we have said nothing about the initial decision about whether to implement a testing policy at all. Our discussion is framed entirely in the setting where it is "agreed" that some degree of detection (and associated deterrence) is "needed." This decision is a complex one, influenced by political, psychological, and economic factors. It is conceivable that some of these factors would be accounted for by sham procedures that simply provide reassurance. All of these issues remain valid and pressing. But we hope that the introduction of the TPI will provide policymakers and managers with objective information on which to base their deliberations.

3.4. Deceptive Detection

The technical nature of this discussion may obscure one key point about the use of randomness, and we return to it briefly. The random mixing of a "test and unpack flagged items" and a "release everything" strategy is optimal because of the linear nature of the payoff function (fraction of contraband detected) and of the cost. But it has the added effect, when truly random, of making it impossible for the opponent to know "where the gaps are." Basically, no matter how the opponent schedules his or her attacks, or efforts to sneak through, the chance that the attacker will be detected remains the same. In effect, every testing procedure, or scanning station, becomes a mix of reality and decoy. Since the opponent cannot know whether he or she will be tested or not, the attacker cannot avoid the risk of detection, and the associated costs.

3.5 Separation of Technology and Policy

To sum up, this analysis has been guided by a principle with broad applicability. We have separated the problem into a technical component, which involves reasoning *from* ROC and cost data, *to* a rule for comparing tests. This enables technical experts to present a decisionmaker or policymaker with a range of consistent alternatives. In addition, the analysis is done in such a way that *no better detection rate* can be achieved for the given technology of testing and unpacking, at each budget level. Similarly, if the policy is set in terms of a detection level, we can be sure that, with any given technology, there is no cheaper way to achieve the detection rate. This is, we believe, the proper way to separate the technical and policy issues.

4. DISCUSSION

While this article was under review and revision we have learned of several conceptually related initiatives that we mention here. First, the ARMOR program, based at the CREATE Center,[15] has been used to introduce random patrolling in a large space (such as an airport) that cannot possibly be completely controlled. If we think of "portions of space" as being analogous to "items to be inspected" there is an analogy. The spatial patrol is limited in that a patrol (e.g., a dog) cannot move instantly to a distant location, while in the model given here the decision about whether to test or skip must be made as each item (or batch of items) is presented. Second, work at the RAND Corporation has identified the fact that even partially effective testing can have a significant deterrent effect because the attacker must deal with the higher cost of achieving a given level of "success."[16]

ACKNOWLEDGMENTS

The support of the Office of Naval Research, Contract Number N0014-07-1-0299, of the NSF, under Contract Number 05-18543, the National

Science Foundation and DHS under Award CBET-0735910 and DHS Contract Number N000-14-07-1-0150, under ONR N00014-07-1-0299, and DHS 2008-DN-077-ARI003-02 are all most gratefully acknowledged. Thanks to Fred Roberts, Alex Kogan, and Sheldon Jacobson for stimulating conversations. We also gratefully acknowledge the comments of two referees who have greatly improved the clarity of this presentation.

REFERENCES

1. Martonosi SE, Ortiz DS, Willis HH. Evaluating the Viability of 100 Per Cent Container Inspection at America's Ports. Rand Technical Report, 2006.
2. Boros E, Fedzhora L, Kantor P, Saeger K, Stroud P. A large-scale linear programming model for finding optimal container inspection strategies. Naval Research Logistics, 2009; 56(5):404–420.
3. Stroud PD, Saeger KJ. Enumeration of increasing Boolean expressions and alternative digraph implementations for diagnostic applications. Proceedings of Computer, Communication and Control Technologies, 2003; IV:328–333.
4. Madigan D, Mittal S, Roberts F. Sequential decision making algorithms for port of entry inspection: Overcoming computational challenges, Proceedings IEEE Intelligence and Security Informatics, 2007; ISI2007:1–7.
5. McLay LA, Jacobson SH, Kobza JE. Multilevel passenger screening strategies for aviation security systems. Naval Research Logistics, 2006; 53:183–197.
6. McLay L, Jacobson S, Kobza J. Making Skies Safer. ORMS Today, 2005; 32:24.
7. Virta J, Jacobson SH, Kobza JE. Analyzing the cost of screening selectee and non-selectee baggage. Risk Analysis, 2003; 23:897–908.
8. Jacobson SH, Karnani T, Kobza JE. Assessing the impact of deterrence on aviation checked baggage screening strategies. International Journal of Risk Assessment and Management, 2005; 5:1–15.
9. Egan JP. Signal Detection Theory and ROC Analysis. New York: Academic Press, 1975.
10. Swets J. The relative operating characteristic in psychology. Science, 1973; 182:990–1000.
11. Zhou K. http://splweb.Bwh.Harvard.Edu:8000/pages/ppl/zou/roc.Html.
12. Maschler M. A price leadership method for solving the inspector's non-constant-sum game. Naval Research Logistics Quarterly, 1966; 13:11–33.
13. Etheridge B. Etheridge Calls for 100 Percent Container Inspection at U.S. Ports. US Congress Press Release, May 4, 2006, US Federal News Service.
14. Martonosi SE, Barnett A. How effective is security screening of airline passengers? Interfaces 2006; 36(6):545–552.
15. Pita J, Manish J, Ordóñez F, Tambe M, Kraus S, Magori-Cohen R. Effective solutions for real-world Stackelberg games: When agents must deal with human uncertainties. In Proceedings of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009), Sierra C, Castelfranchi C, Decker KS, Sichman JS (eds). May 10–15, 2009, Budapest, Hungary.
16. Morral AR, Jackson BA. Understanding the Role of Deterrence in Counterterrorism Security. RAND Corporation. Santa Monica, CA, 2009.