Enhancing Aviation Security Through the Use of Signal Detection Theory

Nicholas Scurich, Ph.D.
University of California—Irvine

Richard S. John, Ph.D.
University of Southern California

William Burns, Ph.D.
Decision Research

The research described in this report was funded by the Department of Homeland Security (award # 2017-ST-061- QA0001). All views expressed in this document are solely the authors' and do not necessarily represent the views of any organization with which the authors are affiliated.
Correspondence concerning this article may be addressed to Nicholas Scurich, 4213 Social & Behavioral Sciences Gateway, Irvine, CA 92697-7085. Electronic correspondence should be addressed to nscurich@uci.edu

Table of Contents

Project Overview

Predictive algorithms derived from big databases have been used in homeland security threat assessment of individuals, such as whether or not to subject an aviation passenger to heightened passenger screening or whether or not to add a person to a "watch list" or "no fly list." Ultimately, the results of such predictive algorithms must be translated into decisions regarding whether to classify an individual as dangerous or non-dangerous. Regardless of how accurate the algorithm might be, the distributions of dangerous and non-dangerous individuals will overlap on the predictive index.

In order to use the predictive index for decision making, a threshold must be established in order to allow individuals to be classified as either dangerous or non-dangerous (Swets, 1992; Swets, Dawes & Monahan, 2000). Fundamentally, specifying this threshold involves a tradeoff between the possibility of misclassifying a non-dangerous individual as dangerous (i.e., a false positive error) and misclassifying a dangerous individual as non-dangerous (i.e., a false negative error). It is important to note that misclassification is inevitable when decisions are made under the condition of uncertainty and on the basis of limited and imperfect information, as is nearly always the case, and particularly so in security contexts.

We propose the use of a framework – Signal Detection Theory – to reify this tradeoff, and to demonstrate its applicability and develop a template for its use in the context of aviation security. Signal Detection Theory (SDT) was initially developed to use in conjunction with radar in World War II (Peterson, Birdsall, & Fox, 1954). It has since been used in a variety of fields, such as engineering, psychometrics, forensic science, and medicine (Egan, 1975; Green & Swets, 1966; Swets & Pickett, 1982; Wickens, 2002; Scurich & John, 2010, 2012). One component of the SDT model is the inevitable tradeoff between false positive and false negative classification errors. Specifying an acceptable tradeoff is an inherently subjective value judgment. Once specified, SDT provides the machinery to combine this tradeoff with other relevant components such as the base rate of the target event and the predictive efficiency of the algorithm.

There are two fundamental objectives of this project:

1. To elucidate the Signal Detection Theory framework, and demonstrate how it can be used coherently to make decisions regarding the provision of security measures to aviation passengers. The current decision-making

process used by aviation security officials makes implicit assumptions about a tolerable value tradeoff; these assumptions may neither optimize safety nor be acceptable to all stakeholders (see Scurich & John, 2014; Gigerenzer, 2004), and they may not be combined to yield the optimal solution. SDT provides an analytic framework to evaluate such decisions, and can used to derive an optimal security threshold for aviation security using any predictive index.

2. To construct a template software tool that will allow the SDT model to be applied in any context involving a predictive index designed to result in a binary classification. Such a tool will enable policy makers to explore the implications of setting different thresholds, conditional on different tradeoffs, base rates, and predictive efficiencies.

Primer on Signal Detection Theory

In May 2019 the authors of this report gave a training workshop to members of the Department of Homeland Security (DHS) and other federal stakeholders on the application SDT to security screening. What follows in this section is a "nuts and bolts" description of what we shared with this group on how to carry out an empirical-based application of SDT to security assessments.

The process begins with identifying the criterion to be predicted, in this context, riskiness of a passenger. In the case of binary classification, two populations of interest that might differ on the distribution of the criterion (e.g., low risk passengers versus moderate-to-high risk passengers) have to be selected. Membership in each of these population needs to be identified (e.g., FBI reports of suspicious activity or not) (see Basuchoudhary, & Razzolini, 2006). Next comes the calculation of sensitivity and specificity for each cutoff level (e.g., risk score), ROC curve and optimal cutoff point (Swets, 1986). How this is done will be illustrated in following sections.

Having selected the prediction criterion and corresponding populations, a sample needs to be drawn from each population. It makes sense to sample from these populations proportional to their membership size. For example, first sample from a FBI list of those in which no suspicious behavior has been reported and call then Low Risk Passenger (LR). Likewise, sample from a FBI list of those in which suspicious behavior has been reported and call them Moderate-to-High Risk Passenger (MHR). For all potential passengers assign

a risk score (e.g. 1-10) based on a set of predictor variables that are readily identifiable and assessable (e.g. frequent flyer, age). At this point, each passenger in this assessment should be a member of either the Low Risk or the Moderate-to-High Risk groups and should have a risk score.

The following table illustrations what a potential data set might look like.

Empirical Estimation of ROC Curve

(hypothetical data)

Risk Score (1-10-less risky) Low Risk (LR = no suspicious activity) Moderate Risk (MHR = suspicious activity)

Passenger	Risk Score	Riskiness
1	1	MHR
2	1	MHR
3	1	MHR
4	1	MHR
5	1	MHR
6	2	MHR
7	2	MHR
8	2	MHR
9	2	MHR
10	2	MHR
11	2	MHR
12	2	LR
13	3	MHR
14	3	MHR
15	3	MHR
16	3	MHR
17	3	MHR
18	3	MHR
19	3	MHR
20	3	LR
21	3	MHR
22	4	MHR
23	4	MHR
24	4	MHR
25	4	MHR
26	4	MHR
27	4	MHR
28	4	MHR
29	4	LR
30	4	LR
31	4	LR
32	5	MHR
33	5	MHR
34	5	MHR
35	5	MHR
36	5	MHR
37	5	MHR
38	5	MHR
39	5	MHR
40	5	MHR
41	5	LR
42	5	LR
43	5	MHR
44	5	MHR
45	6	MHR
46	6	MHR
47	6	MHR
48	6	MHR
49	6	MHR
50	6	MHR
51	6	MHR
52	6	MHR
53	6	MHR
54	6	MHR
55	6	MHR
56	6	MHR
57	6	LR
58	6	LR
59	6	LR
60	6	LR
61	6	LR
62	7	MHR
63	7	MHR
64	7	MHR
65	7	MHR
66	7	MHR
67	7	LR
68	7	LR
69	7	LR
70	7	LR
71	7	LR
72	7	LR
73	7	LR
74	7	LR
75	7	LR
76	7	LR
77	7	LR
78	7	LR
79	7	LR
80	7	LR
81	7	LR
82	8	MHR
83	8	MHR
84	8	MHR
85	8	MHR
86	8	MHR
87	8	LR
88	8	LR
89	8	LR
90	8	LR
91	8	LR
92	8	LR
93	8	LR
94	9	MHR
95	9	LR
96	9	LR
97	9	LR
98	10	LR
99	10	LR
100	10	LR

Using the (hypothetical) data in the table above, it is possible to calculate sensitivity (i.e., the true positive rate), specificity (i.e., the true negative rate), the false positive rate, and the false negative rate. Using the sum of sensitivity and specific allows for the determination of an optimal cutoff point, in this case a risk score of 7. The following table illustrates these calculations for each of the possible risk scores 1-10.

Tabulation of Sensitivity and Specificity

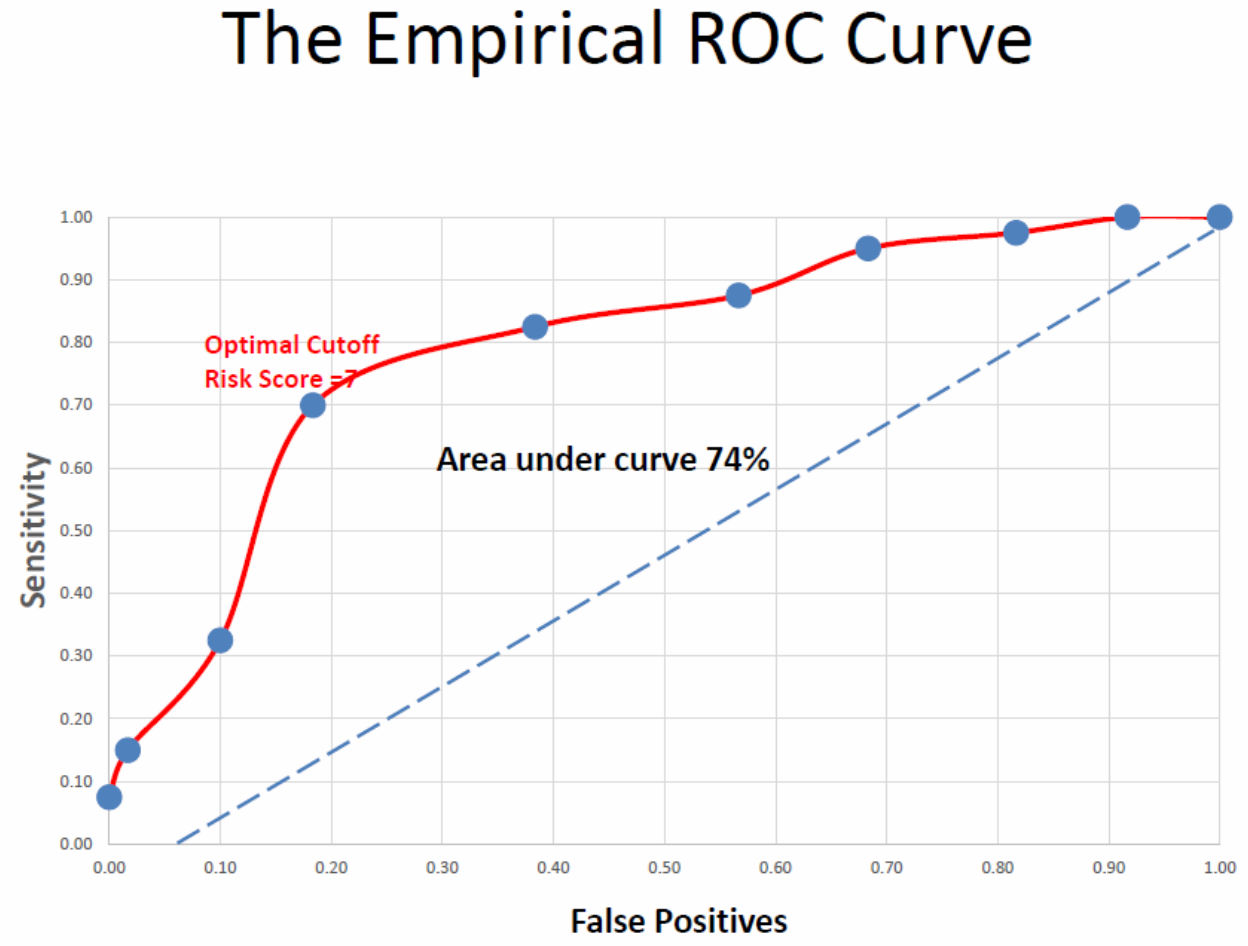
Count of Riskiness	Column Labels		
Row Labels	**LR**	**MHR**	**Grand Total**
1		5	5
2	1	6	7
3	1	8	9
4	3	7	10
5	2	11	13
6	5	12	17
7	15	5	20
8	7	5	12
9	3	1	4
10	3		3
Grand Total	40	60	100

Risk Score	Sensitivity	Specificity	False Positives	False Negatives	Sensitivity + Specificity
1	100.0%	0.0%	100.0%	0.0%	100.0%
2	100.0%	8.3%	91.7%	0.0%	108.3%
3	97.5%	18.3%	81.7%	2.5%	115.8%
4	95.0%	31.7%	68.3%	5.0%	126.7%
5	87.5%	43.3%	56.7%	12.5%	130.8%
6	82.5%	61.7%	38.3%	17.5%	144.2%
7	**70.0%**	**81.7%**	**18.3%**	**30.0%**	**151.7%**
8	32.5%	90.0%	10.0%	67.5%	122.5%
9	15.0%	98.3%	1.7%	85.0%	113.3%
10	7.5%	100.0%	0.0%	92.5%	107.5%

Pessimistic depiction of traveling public-hypothetical

67

Using the calculations in the above table, an empirical ROC curve can now be constructed. The false positive rate (horizontal axis) can be plotted against measures of sensitivity (vertical axis) to create a ROC curve. A decision threshold at any given point on the curve reflects a particular tradeoff between sensitivity and specificity. For example, using group 7 as a cut point (meaning any score greater than or equal to 7 results in an affirmative decision) results in a value of sensitivity of 70% and a false positive rate (1-specificty) of 18.3%. As a general matter, the point on the curve closest to the upper left most point on the graph represents the optimal cutoff point. Please see the following graph as an illustration.



There are a number of practical questions that can be addressed with the above approach and calculations: (1) How do we know which training approach works best?; (2) How can the skill (e.g., detection of a prohibited item during x-ray screening or a pat down) of a particular individual be assessed?; (3) How do we know whether an individual's or a team's performance is improving over time?; and (4) How can we assess the relative difficulty of judgment of one task versus another?

Answers to the above questions can be obtained by tracking the sensitivity, specificity and ROC curves across training approaches, individuals and tasks. Likewise, these measures can be tracked over time as well.

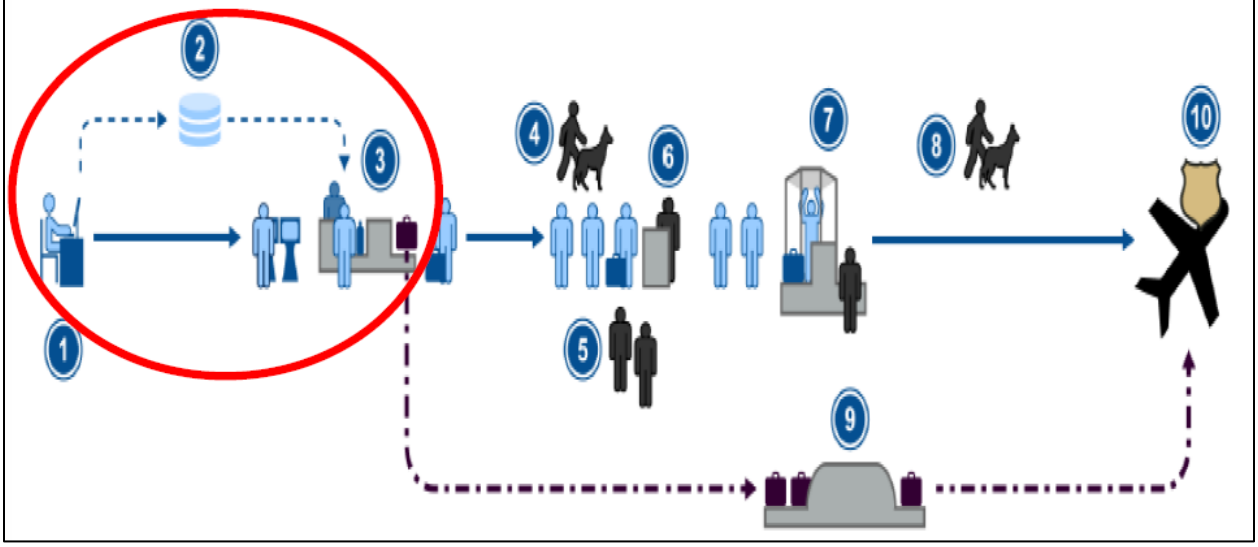
While collecting needed data and computing relevant metrics is straight forward this kind of assessment is not without its challenges. For example, finding validation data such as the FBI suspicious behavior list mentioned above might prove difficult. Certainly, this step requires thought and time (see Sandler, & Enders, 2007). Additionally, an assessment of the relative

costs of false positives and false negatives needs to be thought about carefully to determine optimal classification thresholds.

An Application to Aviation Security

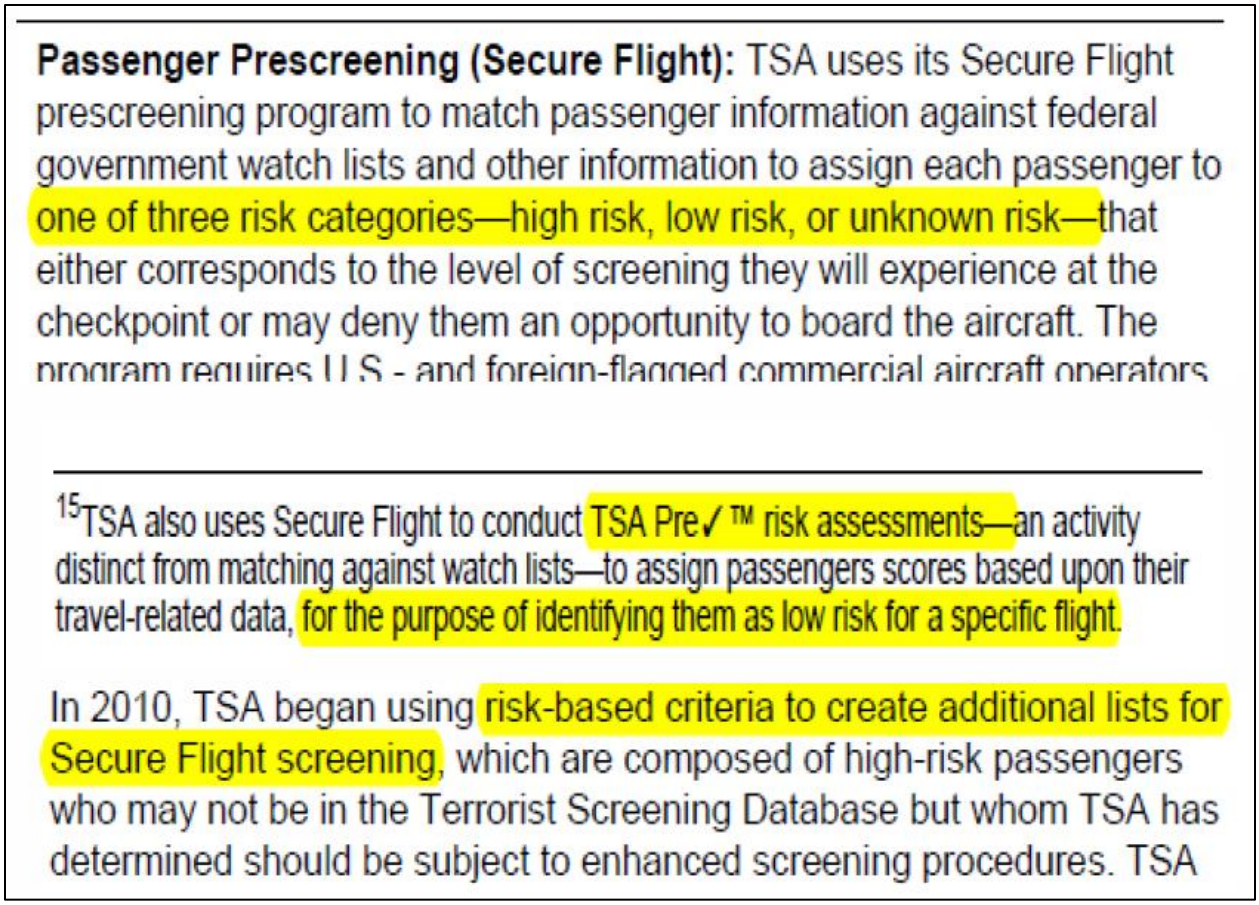
The TSA has a portfolio of counter measures that are designed to identify adversaries. Some of these measures are pictorially represented in the Figure below:

Figure 1: Illustrative Example of a Passenger Experience with Selected Transportation Security Administration (TSA) Countermeasures



Source: GAO-17-794 at 6.

The red circle includes frontend measures that are designed to identify potential adversaries right after they purchase their ticket. This involves searching passenger names against lists of known adversaries or persons of interest. It also involves a passenger prescreening procedure known as "Secure Flight" that is used to assess the risk of passengers. The following passage describes how Secure Flight is used:



Passenger Prescreening (Secure Flight): TSA uses its Secure Flight prescreening program to match passenger information against federal government watch lists and other information to assign each passenger to one of three risk categories—high risk, low risk, or unknown risk—that either corresponds to the level of screening they will experience at the checkpoint or may deny them an opportunity to board the aircraft. The program requires U.S.- and foreign-flagged commercial aircraft operators

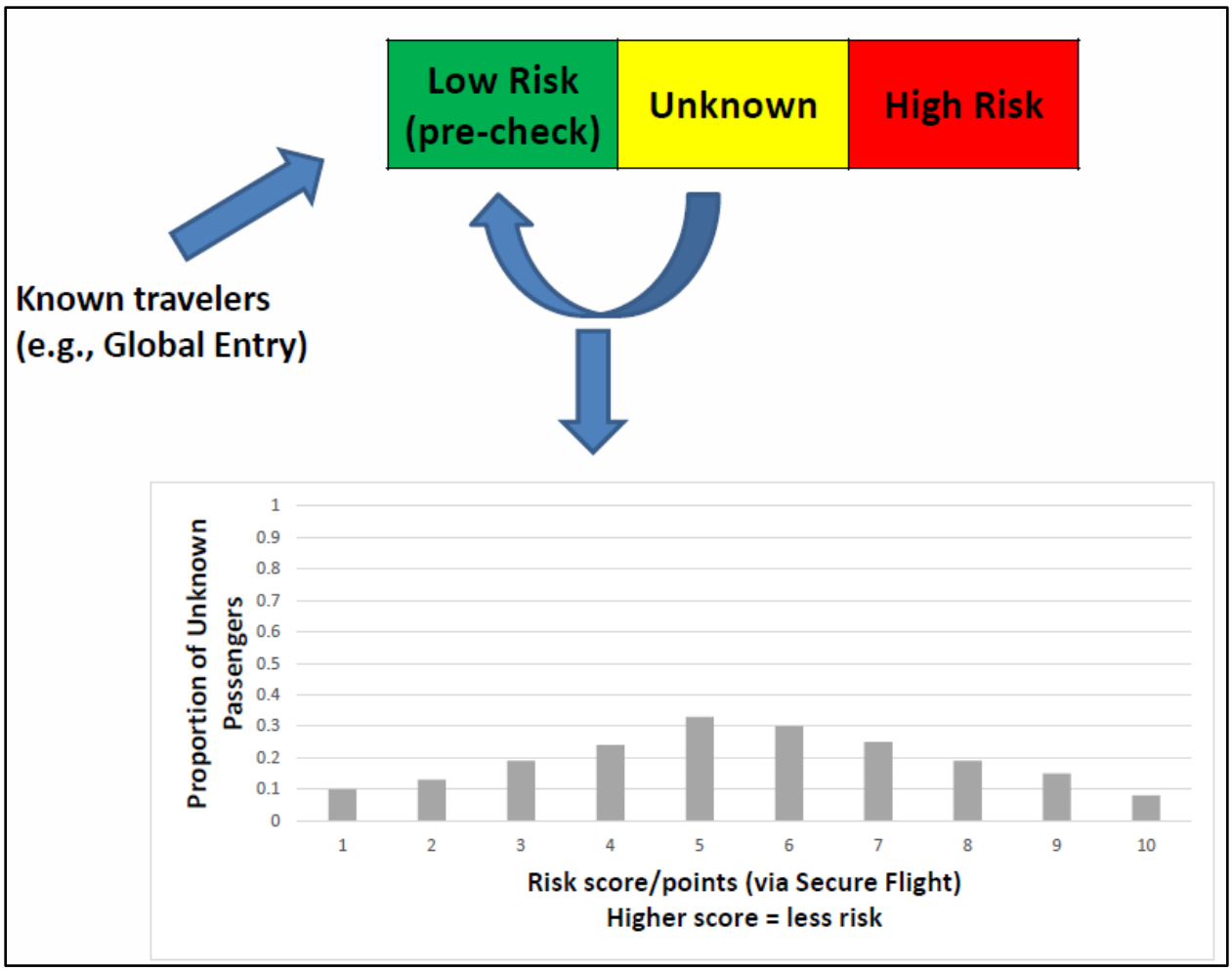
[15]TSA also uses Secure Flight to conduct TSA Pre✓™ risk assessments—an activity distinct from matching against watch lists—to assign passengers scores based upon their travel-related data, for the purpose of identifying them as low risk for a specific flight.

In 2010, TSA began using risk-based criteria to create additional lists for Secure Flight screening, which are composed of high-risk passengers who may not be in the Terrorist Screening Database but whom TSA has determined should be subject to enhanced screening procedures. TSA

Source: GAO-17-794 at 7-8.

Passengers will be subject to different levels of security measures depending on their risk as determined by Secure Flight. Passengers deemed "high risk" will experience heightened security measures. Passengers deemed "low risk" may receive expedited security screening, sometimes referred to as "Pre-Check," and passengers in the middle category will be subject to the standard security screening measures.

There has been much interest in identifying low risk passengers. Low risk passengers require fewer screening resources, and thus can increase operational efficiency and reduce lines and wait times at the normal security ques. There are roughly two ways to be designated as low risk: first, known travelers, such as those registered with TSA precheck or Global Entry, will typically – but not always - obtain a low risk designation; second, by accumulating points from Secure Flight such that some pre-determined amount of points is sufficient to designate the passenger as a sufficiently low risk to accord her expedited security screening. This process is pictorially represented below:

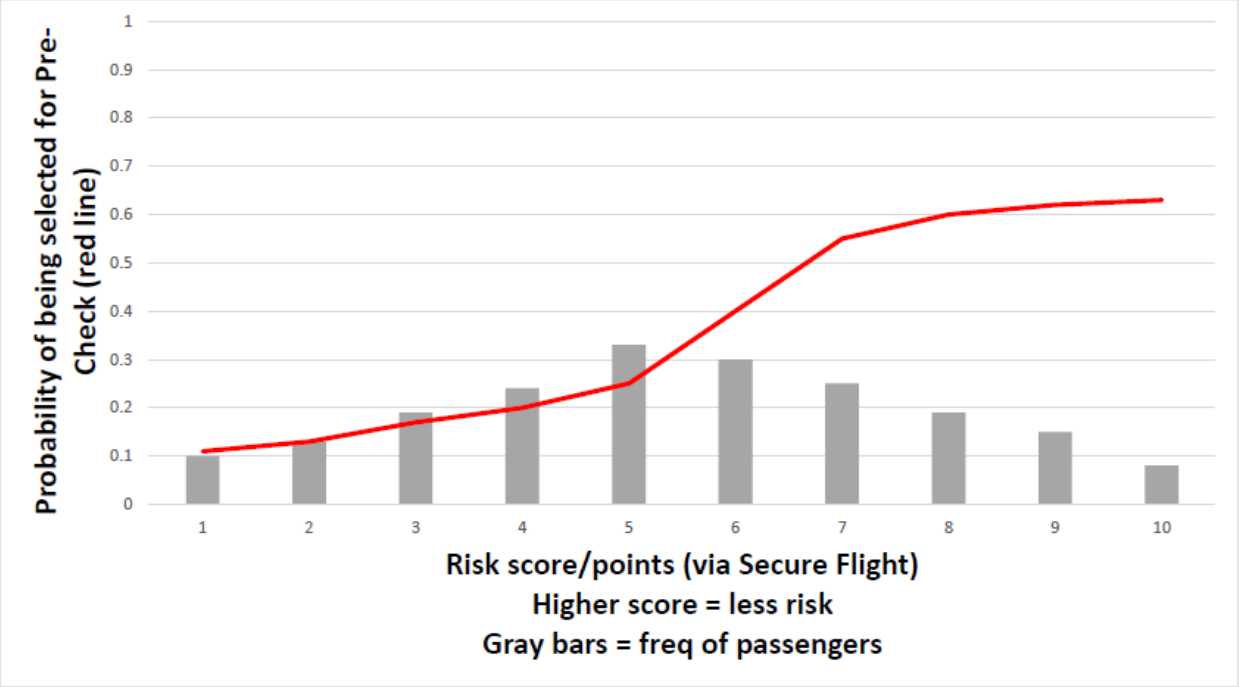


It is important to stress that this pictorial representation, including the values contained in the plot, is completely hypothetical. However, it does provide a rough illustration of the process by which an unknown passenger gets designated as "low risk."

Secure Flight is used to accord passengers with points. This project does not examine the process by which Secure Flight delegates points. That is sensitive information. Rather, the current project focuses on how to set a threshold such that a specific number of points is sufficient to designate the passenger as low risk and thus receive expediated screening. In other words, where to set the "low risk" threshold.

This would seem to be a simple application of SDT, in which the low risk threshold is set based upon an appropriate tradeoff between false positive and false negative errors. However, as we began working with Subject Matter Experts at the Transportation Security Authority (TSA), we learned that a different approach was used. Rather than a hard-cut point or

threshold, the TSA uses a probabilistic approach to select passengers for the low risk designation. This process is pictorially illustrated below:



It is important to stress that this pictorial representation, including the values contained in the plot, is also completely hypothetical.

As depicted above, passengers get a risk score from Secure Flight. These scores are depicted by the gray bars. There is then some probability that an individual with a given risk score will be selected for precheck. This probability is not equally distributed across risk scores. Passengers in lower risk groups (e.g., those with 10 points) have a higher probability of being selected for expedited screening than relatively higher risk passengers (e.g., those with 1 point).

The use of a probabilistic selection process rather than a hard-cut point or threshold is based on a deterrence rationale (Ridinger, John, McBride, & Scurich, 2016). The idea is to mitigate the possibility that adversaries would be able to detect and exploit observable patterns. For instance, (hypothetically) an adversary could observed that individuals with a risk score of 10 always get selected for expediated screening and thus they might try to exploit a passenger with a 10 risk score (setting aside the fact that passengers never learn their risk scores); if passengers with a risk

score of 10 sometimes receive expedited screening and sometimes do not, then it would be less appealing to attempt to exploit such passengers in an effort to gain expedited security screening.

The probabilistic selection model described above adds complexity to the application of SDT to this decision problem, but nevertheless, the same principles of SDT can be used to select low risk passengers. A description of this model follows.

The Use of Signal Detection Theory to Select Low Risk Aviation Passengers

We apply SDT to the problem of identifying low-risk passengers from the population of unknown passengers using information available at the time of screening. These unknown passengers have not been vetted and have not provided any information beyond that available in the Secure Flight database, e.g., sex, age, etc. Our application of SDT and analysis assumes a continuous risk score for each passenger, but does not address how this score is formulated from available data. For demonstration purposes, we assume that the risk scores are normally distributed with equal variance for both unknown and low-risk passengers, and that higher risk scores indicate lower risk. We demonstrate how SDT would be applied to identify an optimal threshold (cut-point) in which passengers with scores above that threshold are assigned to an expedited screening process that is quicker than general unknown passenger screening. The optimal threshold depends on three variables that characterize the specific passenger screening environment (Lynn & Barrett, 2014):

1. Diagnosticity of the risk score, expressed as the normalized difference in the distribution means, d' (Kadlec, 1999).
2. Base-rate of passengers who should not be selected for expedited screening, expressed as an odds ratio, p(~low-risk)/p(low-risk)
3. Relative costs of errors, expressed as the ratio of the cost of false positives (select for expedited screening when not low-risk) to the cost of false negatives (do not select for expedited screening when low-risk).

Calculation of the optimal Beta is given in Equation 1 (Swets & Pickett, 2002, p. 40):

Beta-optimal =cost(false-positive)/cost(false-negative) * p(non low-risk)/p(low-risk)

We provide an Excel spreadsheet that uses these three input variables to calculate the optimal threshold in standardized risk score units and in raw risk score values. This spreadsheet allows for a custom calculation of the optimal threshold that tailors the cut-off to the diagnosticity of the risk score utilized, the estimated base-rate of low-risk passengers, and the relative costs of misclassification of passengers. This flexibility is important since the risk score may change as more diagnostic data sources are identified and utilized. Likewise, it is important to allow the optimal beta to vary depending on the base-rate of passengers for whom expedited screening is appropriate. Flexibility to tailor the optimal cut-off to the base-rate of low-risk passengers is particularly important, since the proportion of low-risk passengers is expected to vary over time and by location.

We present an example calculation of error rates as a function of the diagnosticity of the risk score, d', assuming that the base-rate of low-risk passengers is 50% (odds ratio =1) and the cost of miss-classification is the same for false-positives and false-negatives. In this case, the optimal threshold (Beta = 1.0) is at the mid-point of the means of the two (equal variance) normal distributions, which is at the point in which the normal density functions intersect. This calculation generalizes to all cases in which the penalty ratio for mis-classification is the reciprocal of the odds ratio, which results in optimal Beta = 1.0. For example, optimal Beta = 1.0 if the penalty for classifying a non low-risk passenger as low-risk(expedited screening) is 3 times the penalty for classifying a low-risk passenger as not low-risk (standard unknown screening), and the proportion of low-risk passengers is 25% (odds ratio of 0.25/0.75 = 1/3).

Figure 1 plots error rates, necessarily equal for Optimal Beta = 1.0, as a function of d'. Error rates vary from 50% (random selection, non-diagnostic risk score) to about 30% (moderate diagnosticity, d'=1.0) to about 15% (high diagnosticity, d'=2.0). In order to reach error rates approaching 5%, a d' of over 3.0 is required, and a d' greater than 4.5 is required to achieve error rates of 1%. This analysis demonstrates the extreme sensitivity of error rates to the diagnosticity of the risk score, d', under one set of conditions (optimal Beta = 1.0).

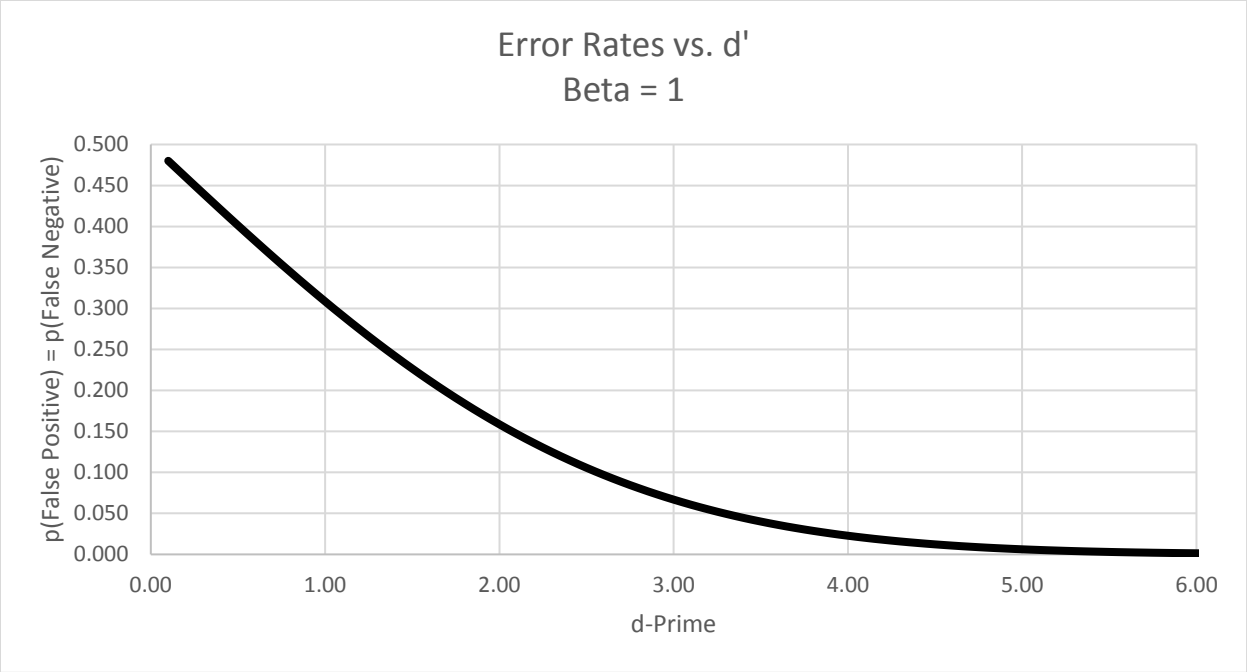


Figure 1. Error rates as a function of d’ for optimal Beta = 1.0.

At present, we do not have data to accurately estimate the diagnosticity of the risk scores currently in use to select passengers for expedited screening (see Dankiewicz, 2012). For comparison purposes, Table 1 presents reported d' values for a wide range of diagnostic tests used for various classification purposes. Typical d' values for commonly used diagnostic tests are between 1.0 and 2.0; d' values of 3.0 or greater are rarely reported. These values provide reference points for gauging the diagnositicy of current and future risk scores.

Table 1. Typical d' values reported in the literature for various diagnostic tests (from Arkes & Mellers, 2002).

Legal Domain	
Polygraphs	
	1.5-2.3 (Swets, 1996)
	2.7 (Raskin and Honts, 2000)
Eyewitness Facial ID	0.8 (Shapiro and Penrod, 1986)
Medical Domain	
Cervical Cancer testing	
	1.6 (Experts)
	1.8 (Algorithmic)
Prostate Cancer PSA testing	2.0
Other Domains	
Weather forecasting	
Rain in Chicago	1.5
Minimum Temperature Albuquerque	1.7
Tornados	1.0
Fog forecasts Canberra airport	0.8-1.2
Job Success for Navy personnel selection (Armed Forces Qualification Test)	0.6-0.8

Explicit consideration of the relative costs of false-positive and false-negative errors can be avoided by selecting a fixed level of either type of error. In the present context, it is natural to consider a fixed false-positive error rate (i.e., classifying as low-risk when not low-risk), and calculating the

implied level of false-negatives (i.e., classifying as not low-risk when low-risk), contingent on the diagnosticity of the risk score, d'. Figure 2 plots false-positive rate as a function of d' for 6 different values of false-negative rate, ranging from 1 in ten to 1 in a million by powers of 10. This analysis suggests that even for modest aspirations for false-positive rates (ranging from 10% to 1%), moderately diagnostic risk scores (d'=1.0) will result in relatively high false-negative rates, ranging from 60% for 10% false-positive rate to 90% for a 1% false positive rate. For highly diagnostic risk scores, d' = 2.0, false negative rates are cut to between 23% and 62% for false positive rates of 10% and 1%, respectively.

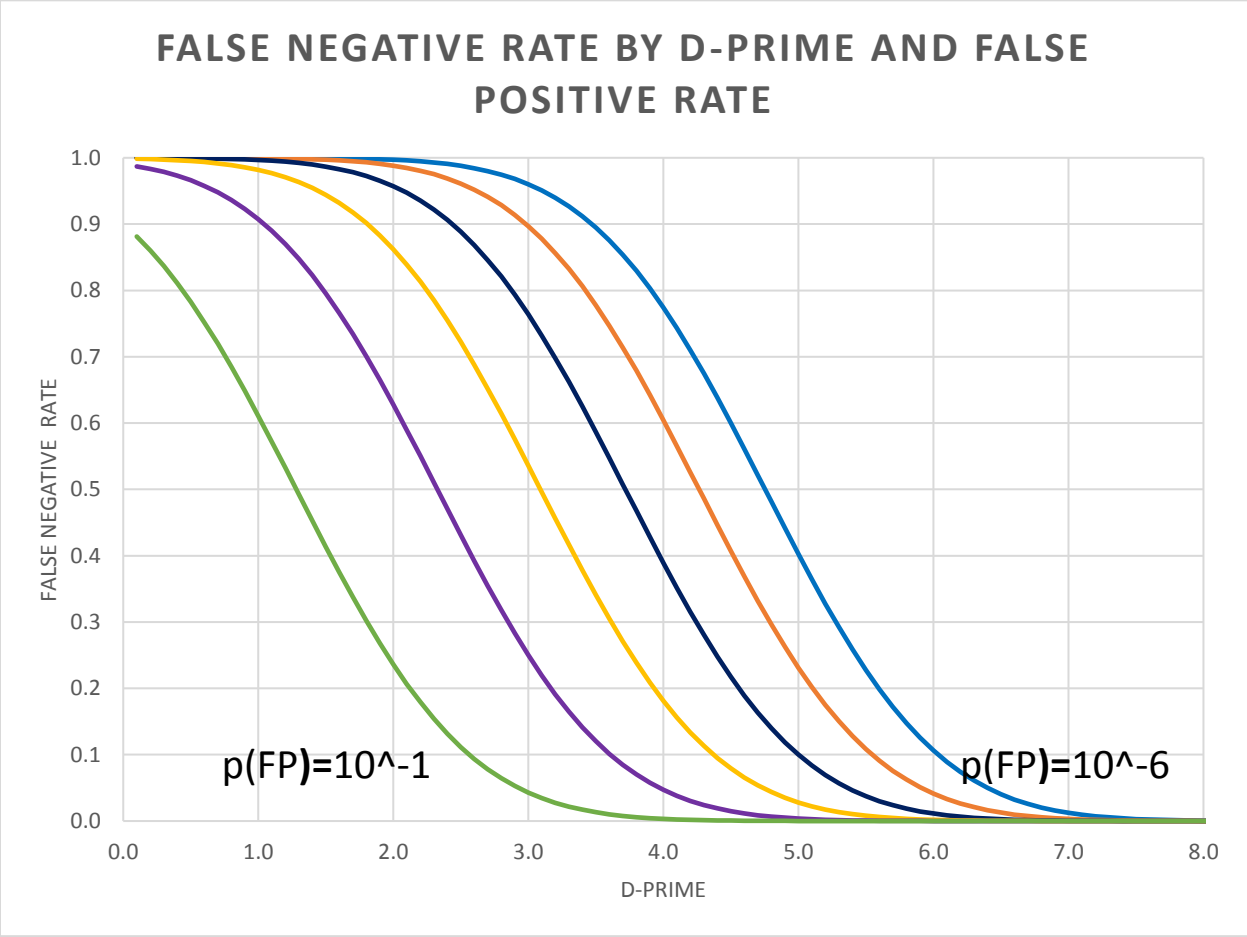


Figure 2. False-negative error rate by d' and fixed false-positive error rates, assuming use of an optimal cut-point for the threshold, Beta.

Figure 2 demonstrates that many low-risk passengers will be required to undergo the standard screening for unknown passengers in order to control the rate of non low-risk passengers selected for expedited screening. Even with extremely high d' values (above 3.0), fixing false-positive rates below 1% will likely result in high false positive rates, and very few passengers selected for expedited screening. The likelihood of misclassification is highly dependent on d'. These analyses highlight the importance of accurate estimation of the diagnosticity of the risk score, d'.

The sensitivity analyses presented in Figures 1 and 2 assume optimal selection of the cut-point, Beta. In the absence of an accurate estimate of d', selection of Beta is unlikely to be optimal, and misclassification rates will necessarily be greater than those plotted in Figures 1 and 2. Understandably, cut-points on the risk score may be determined without a formal estimate of the optimal threshold, and adjusted up or down based on operational variables, such as passenger volume and staffing availability. We conducted a sensitivity analysis to assess the expected cost of applying a threshold either less than or greater than the optimal Beta. In these analyses, we assumed that the optimal Beta is 1.0, which holds when the base-rate of low-risk passengers is 50% and false-positives and false-negatives are equally costly. In this special case of Optimal Beta = 1.0, the expected cost is the average of the false-positive and false-negative rates.

Figures 3-6 display the results of these sensitivity analyses for d' values of 0.50, 1.0, 2.0, and 3.0, assuming normal distributions with equal variance for both the low-risk and non low-risk unknown passengers. The optimal cut-point in each case is equal to half the value of d', since the non low-risk passenger distribution is centered at 0.0 and the low-risk distribution mean is equal to d'. In each case, the lowest expected cost corresponds to the optimal cut-point, and equals the same values plotted in Figure 1 under the same assumptions, i.e., 0.40, 0.31, 0.16, and 0.07 for d' = 0.5, 1.0, 2.0, and 3.0, respectively. As the cut-point choice deviates from the optimal Beta value, the expected cost approaches the worst case value of 0.50, corresponding to selecting all passengers as either low-risk or non low-risk, again assuming that they are equally likely and the cost of each false-positive and each false-negative is fixed at 1.0.

The values plotted are the expected (relative) cost per unknown passenger screened. It is important to note that the expected costs increase more steeply for more diagnostic risk score measures (d'=2.0 or 3.0) than for risk scores that that are less predictive (d'=0.5 or 1.0). That is, it is more important to determine the optimal cut-point for more valid risk scores than

for less valid risk scores. Clearly, misspecification of the cut-point for identifying low-risk passengers from the population of unknown passengers is potentially costly on a per passenger basis; these expected costs are of course multiplied by the millions of unknown passengers screened each year and the actual dollar cost per passenger, now represented as 1 unit for purposes of this analysis.

In some cases, there may be good reasons related to operational efficiency to deviate from the Optimal Beta. For example, small deviations may be preferred if the expected costs are small compared to the cost of maintaining the fixed Optimal Beta across passenger volumes that vary by hour of the day, by day of the week, and by week of the year. Explicit estimation of d' and optimal Beta allows for such trade-offs to be carefully considered and weighed against the expected costs of greater false-positives and/or false-negatives.

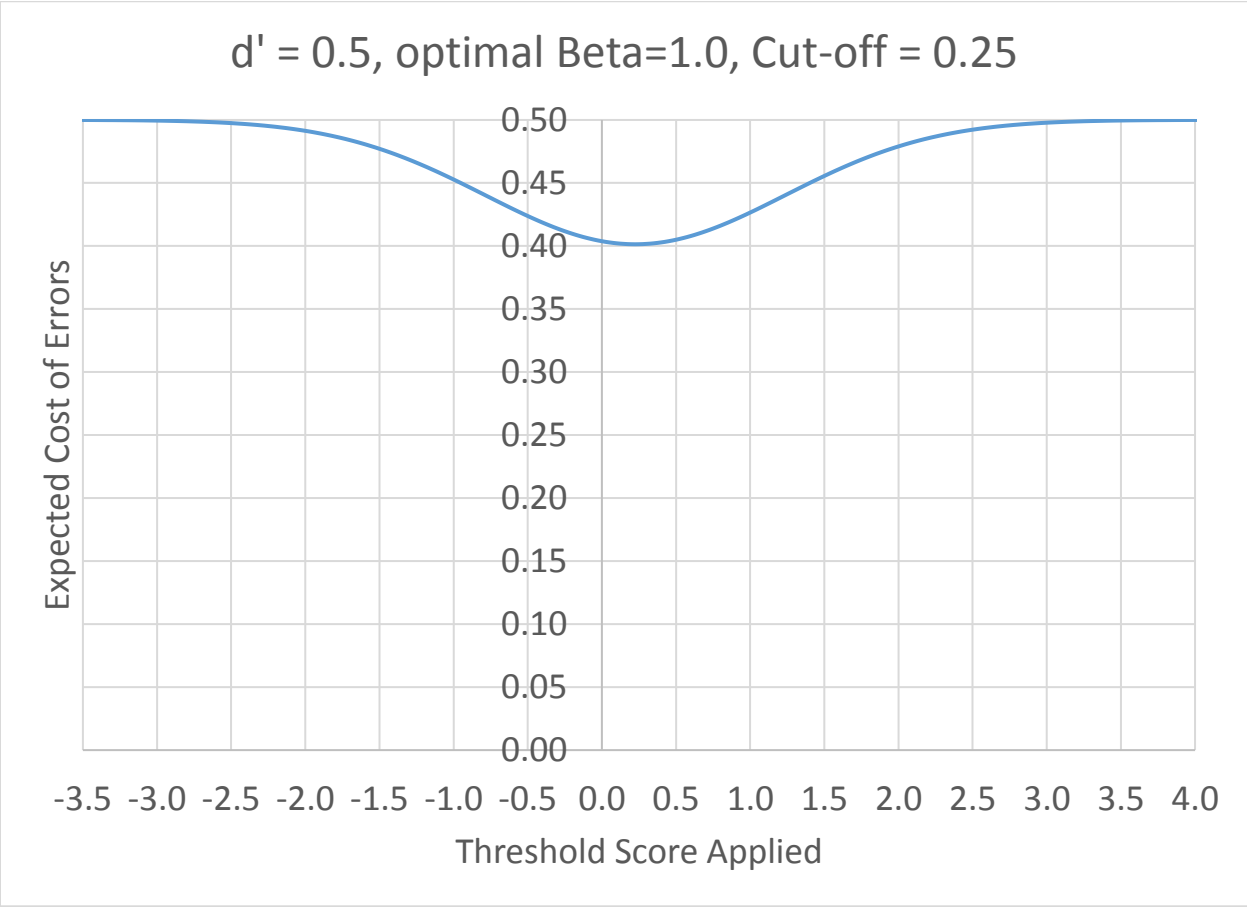


Figure 3. Expected cost per passenger of applying various cut-points (Beta) for d'=0.50, Optimal Beta = 1.0, Optimal threshold = 0.25 (normal distributions with equal variance).

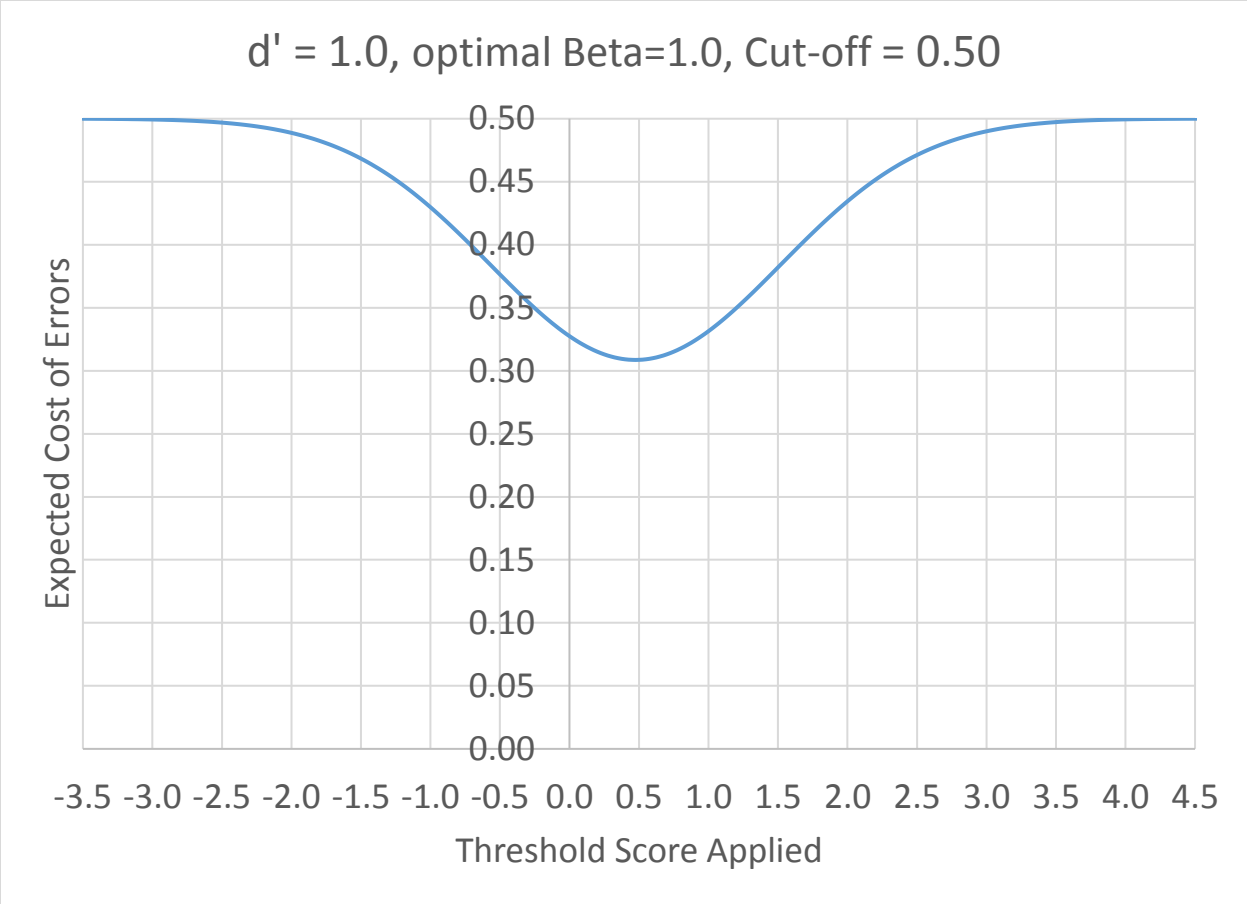


Figure 4. Expected cost per passenger of applying various cut-points (Beta) for d'=1.0, Optimal Beta = 1.0, Optimal threshold = 0.5 (normal distributions with equal variance).

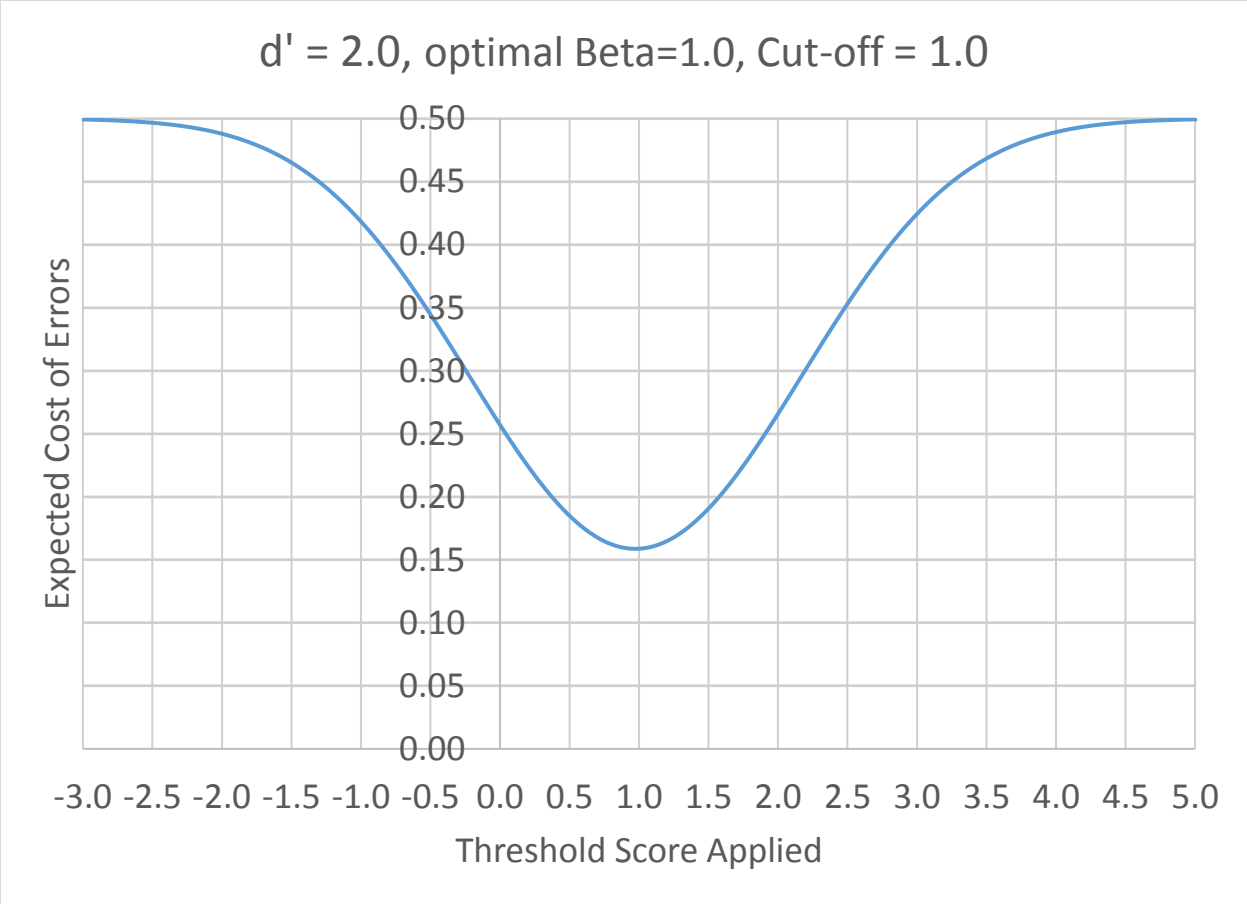


Figure 5. Expected cost per passenger of applying various cut-points (Beta) for d'=2.0, Optimal Beta = 1.0, Optimal threshold = 1.0 (normal distributions with equal variance).

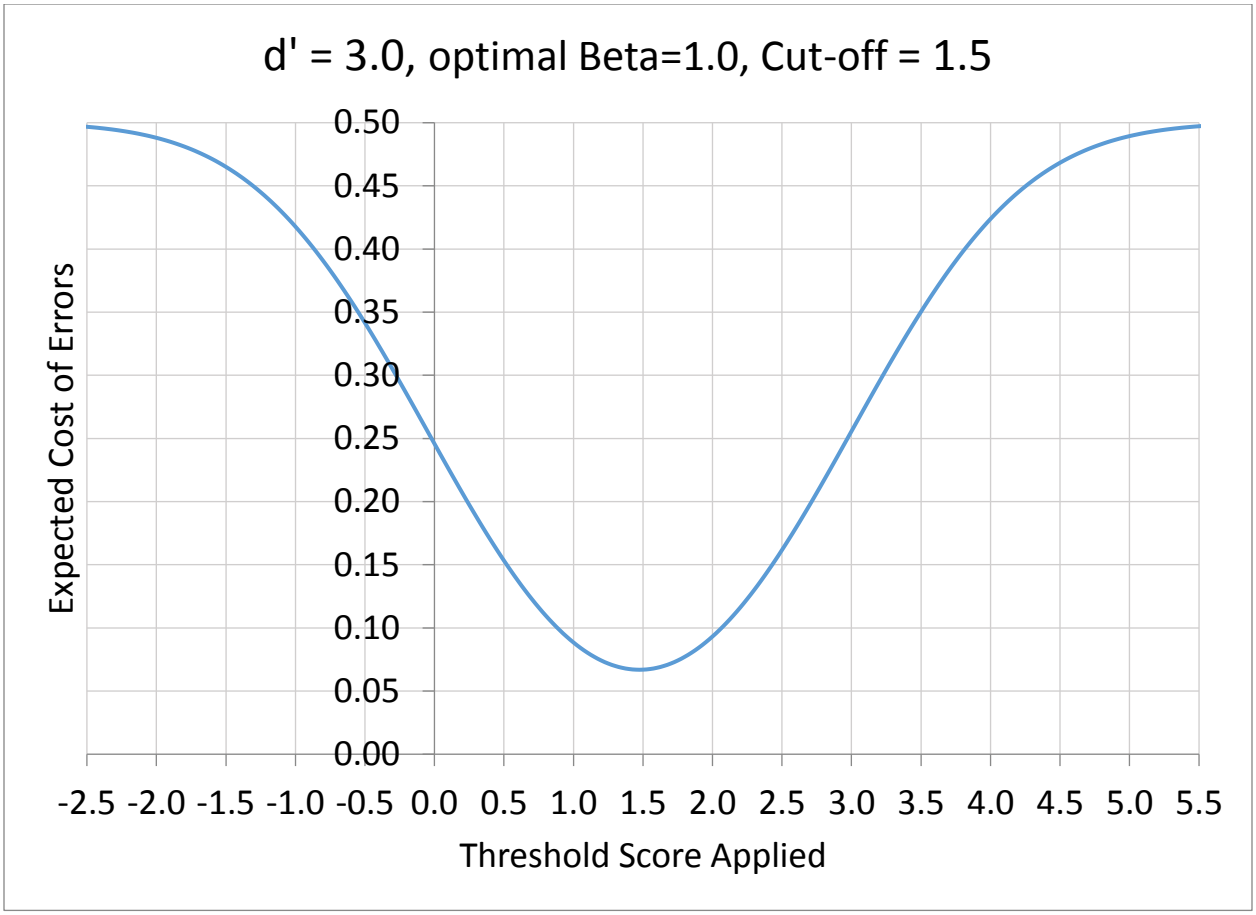


Figure 6. Expected cost per passenger of applying various cut-points (Beta) for d'=3.0, Optimal Beta = 1.0, Optimal threshold = 1.5 (normal distributions with equal variance).

Signal Detection Theory prescribes that the minimum cost of misclassification errors (false-positives and false-negatives) can only be achieved for a particular test or indicator (risk score) with a fixed cut-value for the threshold. That is, all those scoring at or above a fixed value are classified as low-risk and afforded expedited screening, and all those below a fixed value remain unknown and receive usual screening for unknown passengers. An alternative is to randomly assign passengers to expedited screening, where a passenger's probability of expedited screening is monotonically increasing in the risk score assigned to that passenger. Thus, no unknown passenger could be assured of expedited screening, regardless of risk score. This process has the potential advantage of reducing threat by deterring potential attackers who desire or even require the certainty of receiving expedited screening. However, using a random assignment process based on the risk score does result in greater expected costs (false-positives and false-negatives).

We conducted a sensitivity analysis to assess the magnitude of additional expected costs from randomization. The curves in Figure 7 display seven different randomization strategies for the case of optimal Beta = 1.0, centered at the optimal cut-point (midpoint between distribution means). The optimal strategy is to classify all passengers above the midpoint as low-risk and assign to expedited screening, and classify all passengers below the midpoint as non low-risk and assign to usual unknown passenger screening. Note that the optimal threshold value is the point where all of the probability curves intersect in Figure 7, representing a 50-50 probability of assigning to expedited screening at the cut-point for all 7 randomization strategies.

These seven curves all represent various levels of randomization (based on a logistic function) in which the probability of classification as low-risk and assignment to expedited screening is monotonically increasing with risk score. The steepest curve represents a strategy that most closely approximates the strict threshold prescribed by SDT, but does randomize for risk scores very close to the threshold value, the midpoint between the means of the 2 distributions means. As the curves become less steep, greater randomization is utilized; however, the probability of expedited screening is always monotonically increasing with risk score. The steepest logistic curves most nearly approximate the strict threshold approach prescribed by SDT, and result in expected cost closest to the minimum when using the optimal Beta threshold and applying a non-randomization strategy. Flatter curves that deviate from the strict threshold approach would be expected to result in relatively greater expected costs, although they would be less predictable to an adaptive adversary who would prefer assurance of expedited screening based on an anticipated risk score.

Although data regarding randomization based on risk scores was not available for this study, we believe that these seven strategies encompass the range of practical randomization implementation. The most realistic strategies are the three middle curves in Figure 7. That is, the two steepest curves involve very little randomization and thus would not achieve deterrence objectives, while the two flattest curves randomize to such an extent that the risk score plays a greatly attenuated role in determining whether a passenger receives expedited screening. Any practical use of randomization is likely represented by one of the three middle curves in Figure 7.

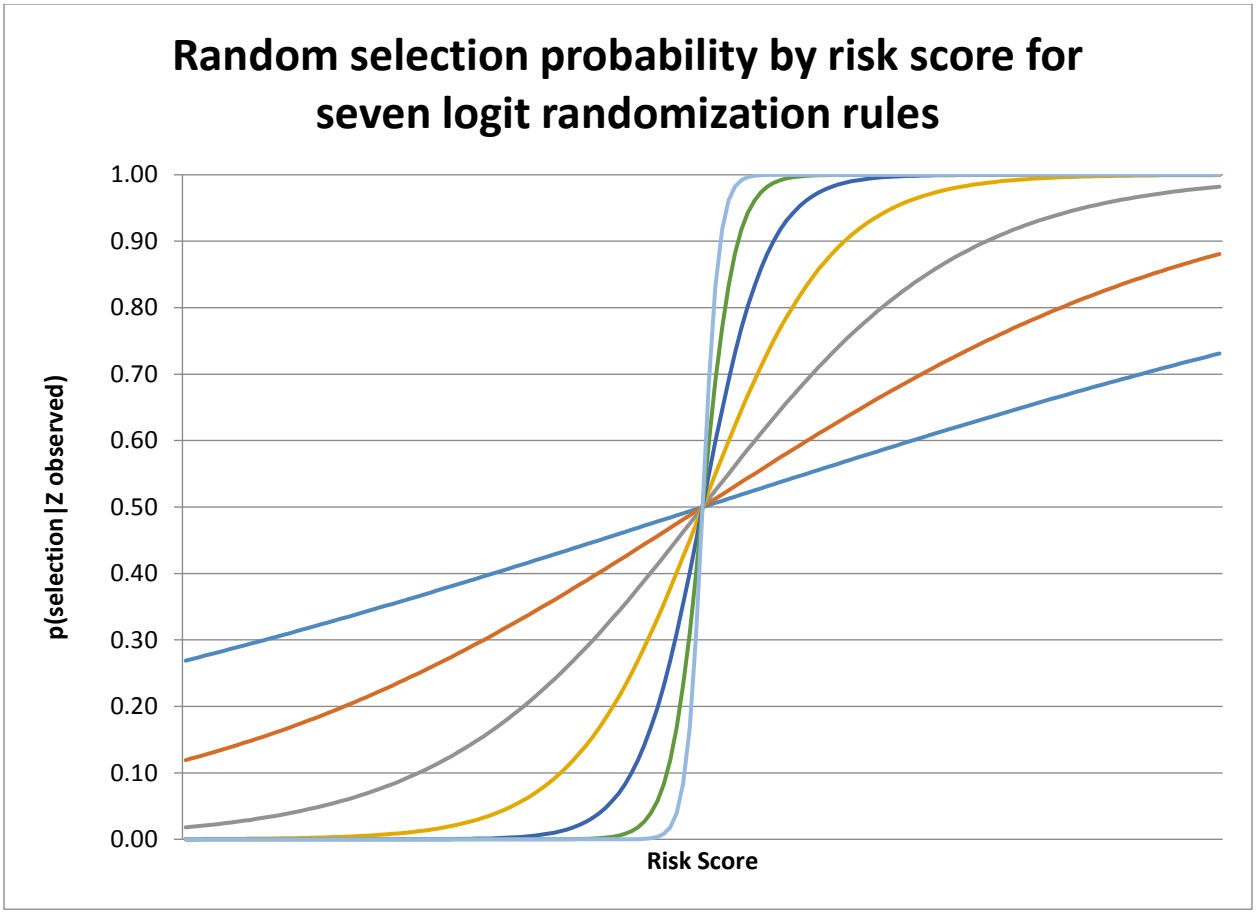


Figure 7. Probability of selection for expedited screening conditional on risk score for seven (logit) randomization strategies centered at optimal cut-point, Beta=1.

We conducted a sensitivity analysis for these seven hypothetical randomization strategies to determine the magnitude of expected costs (increased false-positives and false-negatives) as a function of the extent of randomization utilized. Figures 8-11 summarize expected costs for the case of optimal Beta = 1.0, varying risk score diagnosticity, d'=0.50, 1.0, 2.0, and 3.0. The horizontal lines in all 4 figures represent the minimum possible expected cost for applying the strict threshold at the optimal cut-point with no randomization. Regardless of diagnosticity, d', the additional expected costs for the two steepest randomization curves (logistic parameters, 8 and 16 in Figure 7) are quite low, as the expected costs are quite close to applying a strict threshold (horizontal line). As the logistic parameter decreases and the randomization curve from Figure 7 becomes flatter, expected cost increase and approach the maximum expected cost resulting from complete random assignment, 0.50. In the case of moderate diagnosticity (d'=1, Figure 9), the flattest randomization curve (logistic

parameter = 0.25, Figure 7), expected cost increase over 50%, from 0.31 to about 0.47, which is very close to the maximum cost of 0.50 resulting from completely random selection for expedited screening. The middle three curves (logistic parameters 1, 2, and 4, Figure 7) result in modest increases in expected cost, ranging from about 10% increase (logistic parameter = 4) to about 30% increase (logistic parameter = 1).

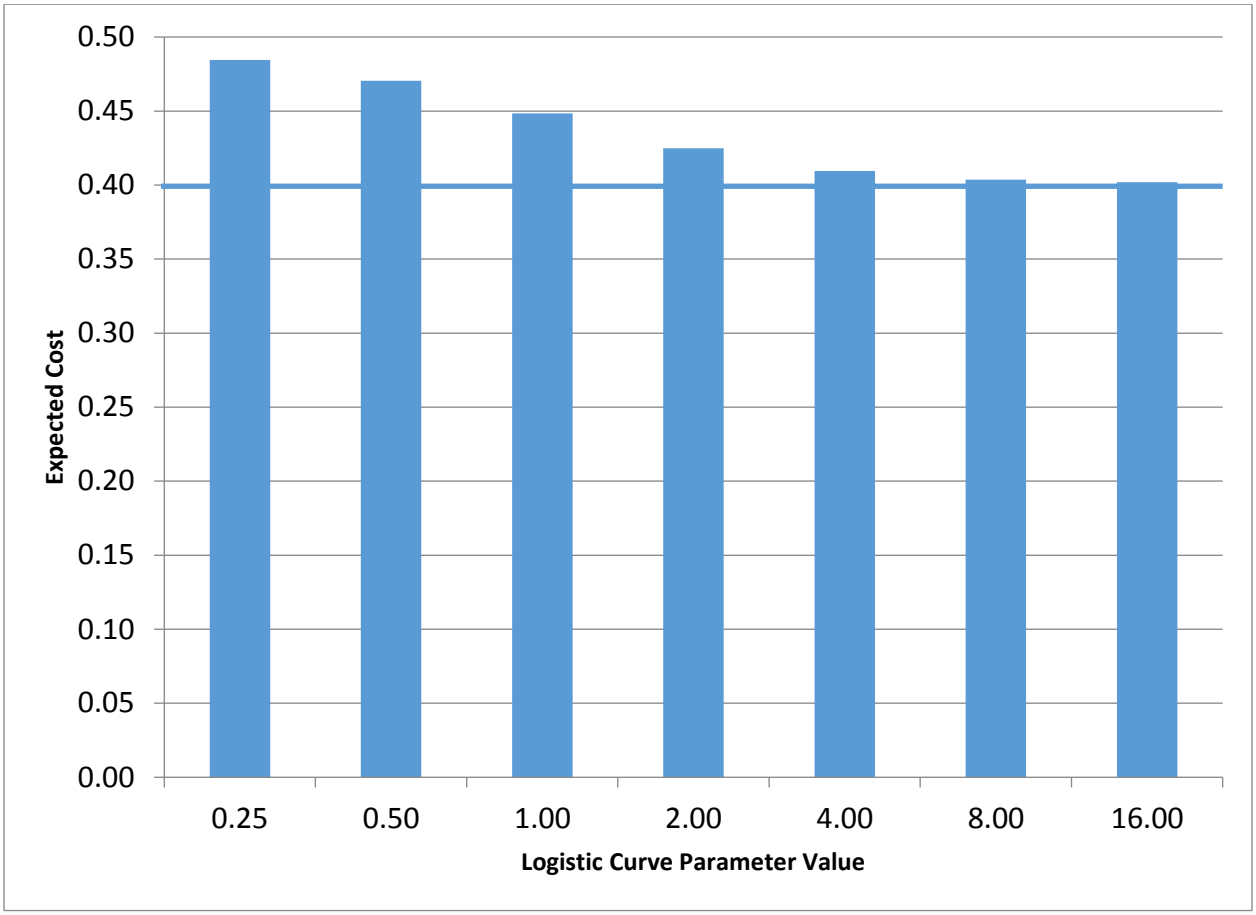


Figure 8. Expected costs (false-positives and false-negatives) by randomization, d'=0.50, Optimal Beta = 1.0, Optimal Threshold Z = 0.25. Lower values of the logistic curve parameter result in more randomization (completely random = 0); higher values of the logistic curve parameter result in less randomization, i.e., approaching SDT strict threshold, 0.40, represented by horizontal line.

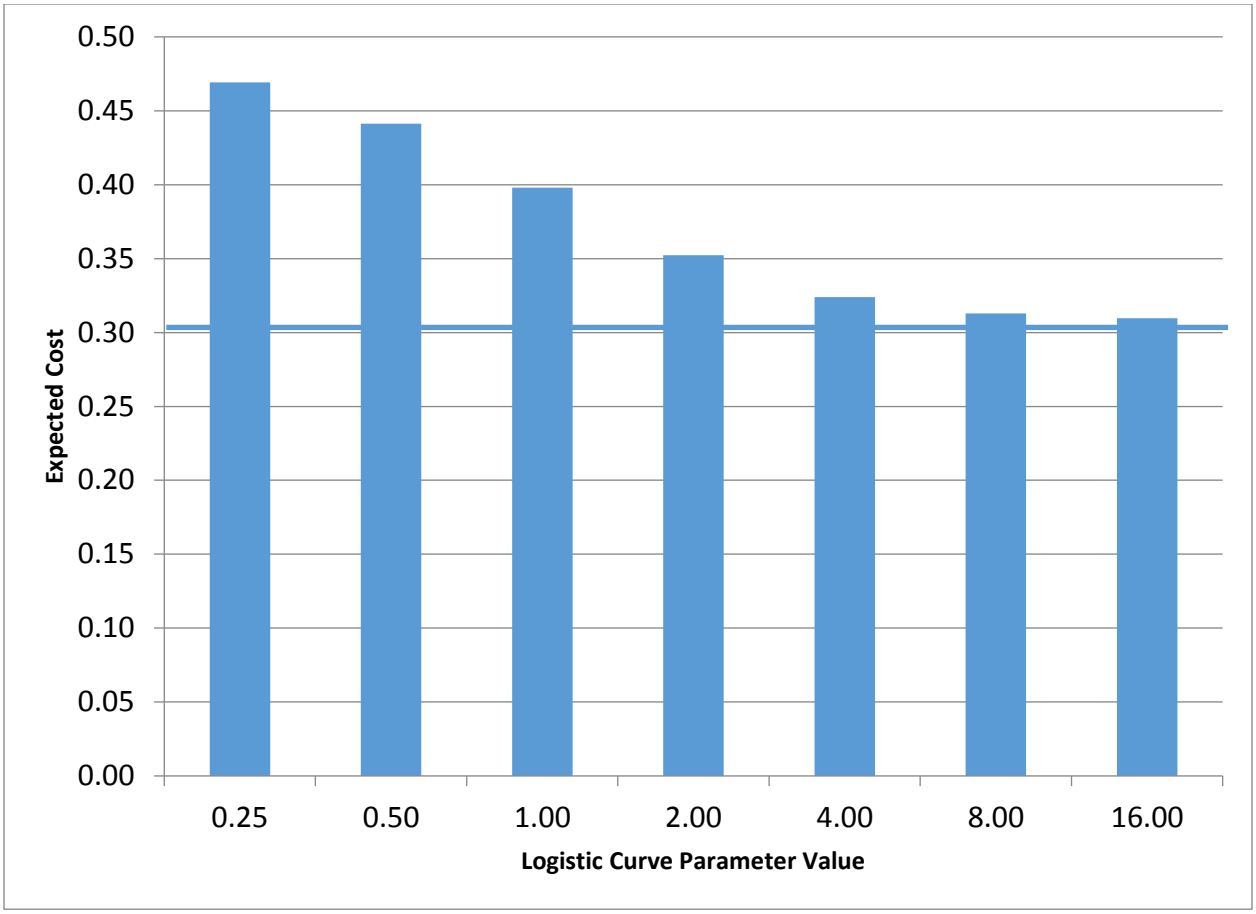


Figure 9. Expected costs (false-positives and false-negatives) by randomization, d'=1.0, Optimal Beta = 1.0, Optimal Threshold Z = 0.50. Lower values of the logistic curve parameter result in more randomization (completely random = 0); higher values of the logistic curve parameter result in less randomization, i.e., approaching SDT strict threshold, 0.31, represented by horizontal line.

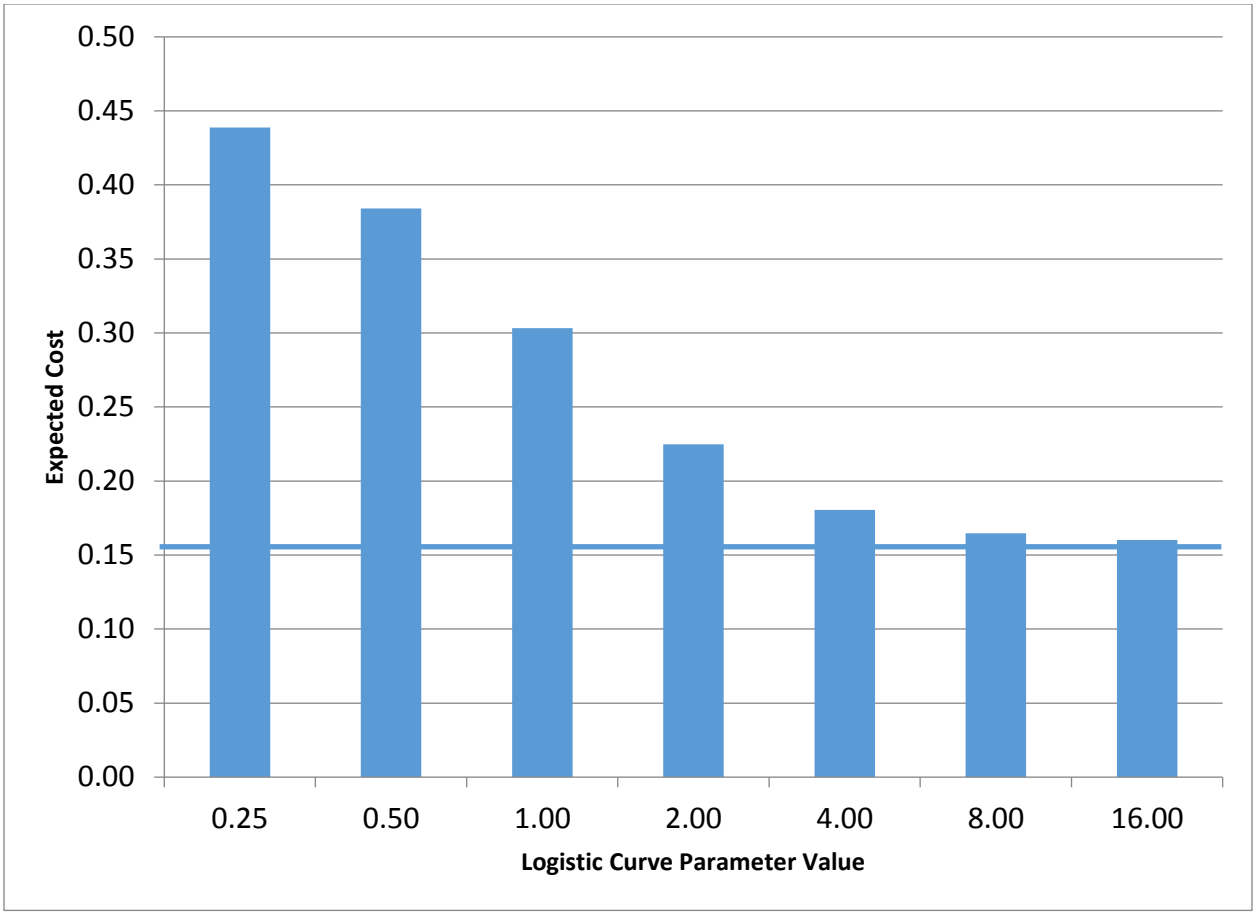


Figure 10. Expected costs (false-positives and false-negatives) by randomization, d'=2.0, Optimal Beta = 1.0, Optimal Threshold Z = 1.00. Lower values of the logistic curve parameter result in more randomization (completely random = 0); higher values of the logistic curve parameter result in less randomization, i.e., approaching SDT strict threshold, 0.16, represented by horizontal line.

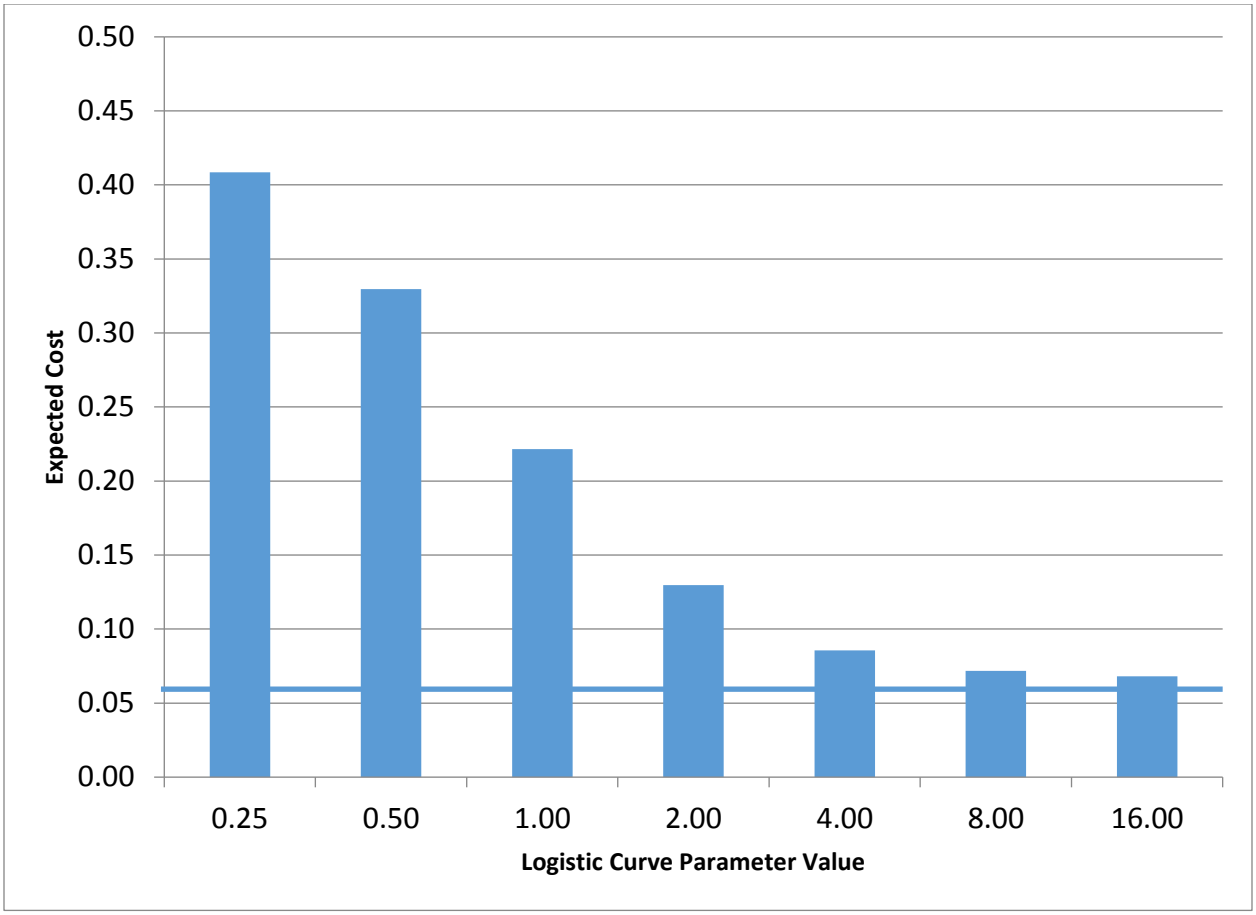


Figure 11. Expected costs (false-positives and false-negatives) by randomization, d'=3.0, Optimal Beta = 1.0, Optimal Threshold Z = 1.50. Lower values of the logistic curve parameter result in more randomization (completely random = 0); higher values of the logistic curve parameter result in less randomization, i.e., approaching SDT strict threshold, 0.07, represented by horizontal line.

In the case of lower diagnosticity (d'=0.50, Figure 8), expected costs increase with increased randomness (lower logistic parameter values), but not as steeply as for d' = 1.0 (Figure 9). In contrast, for cases with high diagnosticity (d'=2, Figure 10 and d'=3, Figure 11), expected costs increase dramatically with increased randomness (lower logistic parameter values). For d'=2 (Figure 10), a moderate degree of randomness (logistic parameter = 1) nearly doubles the expected cost of misclassification, from 0.16 (no randomness) to 0.30. For a high level of diagnosticity, d'=3 (Figure 11), a moderate degree of randomness (logistic parameter = 2) nearly doubles expected cost from 0.07 to 0.13, and the expected loss triples for slightly more randomness (logistic parameter = 1), to 0.22. Randomization has substantial impact on expected costs when diagnosticity is high.

The sensitivity analyses presented in Figures 8-11 make clear that randomization can result in significant expected costs from misclassification errors (false-positives and false-negatives), particularly when the risk score is more diagnostic. It is clear that any randomization strategy should be chosen with care, based on a consideration of minimizing miss-classification errors (expected costs from the SDT model) and minimizing threat through increased deterrence resulting from randomization. This trade-off should be carefully considered after estimation of the diagnosticity of the risk score (d'), the expected costs from randomization, and the anticipated benefit of reduced threat through deterrence following randomization.

Implications of Results and Future Directions

We have demonstrated the value of applying an SDT model to the problem of classifying unknown passengers for expedited screening. The efficacy of using a risk score for classification and selection purposes depends on accurate assessment of its diagnosticity (d'), as well as an accurate assessment of base-rate of low-risk passengers in the population of unknown passengers and the relative costs of false-positives (designate low-risk when not low-risk) and false-negatives (designate non low-risk when low-risk). Heuristic rules that do not attempt to estimate d' and determine an optimal threshold value for classification are likely to result in suboptimal performance, greatly increasing the misclassification errors during screening. As indicated in Figures 3-6, costs from misclassification errors can be dramatic, particularly for more diagnostic risk score indices. Furthermore, while randomization may be justified for the purpose of increasing deterrence, there is a price to pay with respect to increased costs due to misclassification. Any use of randomization should involve a careful trade-off analysis of the expected increase in miss-classification errors and their cost against the anticipated benefits of threat reduction due to randomization.

The next step in this research is to demonstrate use of the SDT model at a particular airport. Such a project would require data that would allow us to parameterize the SDT model and estimate various parameters, including diagnosticity of the risk scores (d'), proportion of low risk passengers (e.g., those who would be granted expedited screening), and the relative costs of passenger misclassification (false-negatives and false positives). Results from such a study would allow us to compare the current screening strategy with both the optimal SDT strategy (no randomization) and with strategies utilizing various levels of randomization. This would allow for potential improvements in applying risk scores and tailoring them to specific locations and times.

References

Arkes, H. & Mellers, B. (2002). Do juries meet our expectations? *Law and Human Behavior, 26*(6), 625-639.

Basuchoudhary, A., & Razzolini, L. (2006). Hiding in plain sight–using signals to detect terrorists. Public Choice, 128(1-2), 245-255.

Dankiewicz, M. (2012). Methods of detecting potential terrorists at airports. Security Dimensions. *International and National Studies, (07)*, 33-46.

Egan, J. (1975). *Signal Detection Theory and ROC analysis*. New York: Academic Press.

Gigerenzer, G. (2004). Dread risk, September 11, and fatal traffic accidents. *Psychological Science, 15*(4), 286-287.

Green, D.M., & Swets J.A., (1966). *Signal Detection Theory and psychophysics.* New York: John Willey

Kadlec, H. (1999). Statistical properties of d' and β estimates of signal detection theory. *Psychological Methods, 4*(1), 22.

Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" signal detection theory. *Psychological Science, 25*(9), 1663-1673.

Peterson, W.G., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 171-212.

Peterson, J. L., Saks, M. J., & Phillips, V. L. (2001). The application of signal detection theory to decision-making in forensic science. *Journal of Forensic Science*, *46*(2), 294-308.

Ridinger, G., John, R.S., McBride, M., & Scurich., N. (2016). Attacker deterrence and perceived risk in a Stackelberg security game. *Risk Analysis, 36*(8), 1666-1681.

Sandler, T., & Enders, W. (2007). Applying analytical methods to study terrorism. *International Studies Perspectives, 8*(3), 287-302.

Scurich, N., & John, R.S. (2010). The normative threshold for psychiatric civil commitment. *Jurimetrics Journal, 50*(4), 425-452.

Scurich, N. & John, R.S. (2012). Constraints on restraints: A Signal Detection analysis of the use of mechanical restraints on adult psychiatric inpatients. *Southern California Review of Law and Social Justice, 21*(1), 75-107.

Scurich, N. & John, R.S. (2014). Perceptions of randomized security schedules. *Risk Analysis, 34*(4), 765-770.

Swets, J. & Pickett, R. (1982). *Evaluation of diagnostic systems: Methods from Signal Detection Theory*. New York: Academic Press.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin, 99*(1), 100-117.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*(4), 522-529.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1-26.

Wickens, T. (2002). *Elementary Signal Detection Theory*. Oxford: Oxford University Press.