2014

The Paradox of Misaligned Profiling: Theory and Experimental Evidence

Charles A. Holt

Andrew Kydd

Laura Razzolini

Roman M. Sheremeta
Chapman University

The Paradox of Misaligned Profiling: Theory and Experimental Evidence

The Paradox of Misaligned Profiling: Theory and Experimental Evidence

Charles A. Holt, University of Virginia,
Andrew Kydd, University of Wisconsin
Laura Razzolini, Virginia Commonwealth University
Roman Sheremeta, Case Western Reserve University and the Economic Science Institute

May 20, 2014

Abstract

This paper implements an experimental test of a game-theoretic model of equilibrium profiling. Attackers choose a demographic "type" from which to recruit, and defenders choose which demographic types to search. Some types are more reliable than others in the sense of having a higher probability of carrying out a successful attack if they get past the security checkpoint. In a Nash equilibrium, defenders tend to profile by searching the more reliable attacker types more frequently, whereas the attackers tend to send less reliable types. Data from laboratory experiments with financially motivated human subjects are consistent with the qualitative patterns predicted by theory. However, we also find several interesting behavioral deviations from the theory.

JEL Classifications: C72, C91, J16
Keywords: terrorism, profiling, game theory, laboratory experiment

---

Corresponding author: Laura Razzolini, E-mail: lrazzolini@vcu.edu
We thank two anonymous referees and the Editor of this journal for their valuable suggestions. We have benefitted from the helpful comments of Rachel Croson, Catherine Eckel, seminar participants at the University of Virginia, the University of Texas at Dallas, SUNY-Buffalo and participants at the North American Economic Science Association Conference in Tucson. We also wish to thank Michael Patashnik and Emily Snow for research assistance. Research support provided by the University of Texas at Dallas, the Virginia Commonwealth Presidential Research Incentive Program and NSF grant NSF/NSCC-0904695 to Razzolini is gratefully acknowledged. Holt's work on the project was funded in part by NSF/NSCC grants 0904795 and 0904798 and BCS-0905044, and the UCS CREATE National Center for Risk and Economic Analysis of Terrorism Events.

1. Introduction

An important problem facing security personnel is to identify terrorists within large groups of mostly innocent people. This problem arises at checkpoints of all kinds, such as roadblocks, permanent checkpoints between different regions or countries and airport security counters. In such settings, large numbers of people pass through a screening process designed to detect and detain terrorists. Typically, the volume of traffic in comparison with the number of potential terrorists is so large that it is neither economically nor politically sensible to screen everyone with the intensity required to detect a terrorist. The security personnel, therefore, face the difficult task of deciding whom to take out of line and subject to greater scrutiny when there are many innocent people and few terrorists. For instance, this issue was faced at U.S. military checkpoints that attempted to secure the Green Zone in Baghdad during the U.S. occupation, and it is also faced daily at busy airports, especially during periods of high security alert.

The problem is compounded by the fact that a terrorist group may be strategic in the sense of being able to respond quickly to any targeted screening program. If the government screens one category that is closely associated with the terrorist group (for instance, young men from a certain province in the Iraq case), then the terrorist group faces a strong incentive to begin recruiting and sending people outside that category, for example, women. The government would then rationally respond by searching women, at least to some extent. The equilibrium outcome of this interaction is unclear. Some argue for a completely random search process, but that would give the terrorists an incentive to send only their core supporters, since they would be no more likely to be searched than anyone else, and these core supporters *would* presumably be more likely to carry out a successful attack if not searched. There is a need for careful analysis to determine what kind of search strategy is rational, implementable and efficient in an environment with threats that evolve in response to securities measures.

Profiling occurs when a certain characteristic or signal, such as race or ethnicity, is used to decide who to subject to a more intrusive investigation. Racial profiling, for instance, is based on the belief that certain crimes are committed disproportionately by the members of a particular race. The scholarly debate in the social science literature is mostly focused on whether profiling occurs and whether it is effective, rather than on normative and constitutional issues. Profiling is not *per se* illegal in most countries, and legal scholars have discussed the pros and cons of various types of profiling and the circumstances under which it might be justified (Barnes and

Gross, 2002; Ellmann, 2003).[1] Economists have studied racial profiling issues from the perspective of whether profiling is a rational use of limited enforcement assets. A recent literature pioneered by Knowles, Persico and Todd (2001) has characterized the Nash equilibrium of a simultaneous move game between the police and a specific group of the population, such as motorists/drivers, who may commit a crime. The police objective is to minimize crime when deciding which vehicles to search, while motorists choose whether to carry contraband or not. The equilibrium involves unequal investigation rates across different demographic groups, even if police officers are unbiased, as long as the members of one group incur higher costs of carrying contraband than those of another group. In this sense, profiling can result from a type of rational experienced-based or "statistical" discrimination. Antonovics and Knights (2009) point out that if statistical discrimination alone is used to explain differences in the rates at which vehicles of drivers of different races are searched, then these search decisions should be independent of police officers' *own* race. They test this prediction using data from the Boston Police Department and find that officers are more likely to conduct a search if the race of the officer differs from the race of the driver.

The debate on profiling significantly changed after the terrorist attacks of September 11th, 2001. Screening procedures have included different versions of the Computer Assisted Passenger Prescreening System (CAPPS) and the Secure Flight Passenger Screening Program, a computerized tool to select passengers for screening, and more recently the full body image scanning. Passengers with elevated ratings according to these mechanisms are selected for additional searches and for baggage inspection, while some other passengers are still searched at random. If previously the discussion was about whether demographic profiling was happening, after 9/11 researchers and politicians have focused on the conditions under which such profiling is acceptable, either constitutionally or as a policy matter. All screening and profiling mechanisms have encountered criticisms and often legal actions. Computerized searching mechanisms are often accused of inducing racial or religious profiling and discrimination. In

[1] In *Protecting Liberty in an Age of Terror*, Heymann and Kayyem (2005) discuss the tension between avoiding past abuses of profiling and the need to confront high-stakes threats. They suggest more reliance on nationality-based profiling, as opposed to pure racial profiling. Barak-Erez (2007) offers an effects-based consideration: if profiling really is necessary, then it should be used more often in situations in which it is less likely to have a long-lasting effect on the lives of those being profiled. Harcourt (2007) provides a thoughtful discussion of the unintended consequences of profiling, such as the tendency for "false positive" searches to induce more terrorism.

fact, there is a debate among experts about whether profiling strategies are more effective than pure random searches.

Several years ago, on November 22, 2010, National Public Radio held an Oxford-style debate at New York University with the two teams arguing the motion "Should U.S. Airports Use Racial and Religious Profiling?" Advocates of the motion supported the use of profiling specifically concentrating on young fundamentalist Muslim males from the Middle East, as the majority of recent terrorist attacks have been associated with this type of individual. Opponents argued that profiling just invites terrorist groups to recruit agents who do not fit the profile. Bin Laden himself, in his hand-written journals, "exhorted followers to explore ways to recruit non-Muslims … – particularly African Americans and Latinos" (the *New York Times*, 5/12/2011).

This paper contributes to this debate by providing theoretical analysis and experimental validation to guide policy makers to improve the effectiveness of targeted and/or profiled screening. Our research investigates the conditions under which profiling is a rational and efficient counterterrorism policy. Although our work is related to the previous economic literature, we take a somewhat different approach by assuming that the terrorist group rationally chooses individuals with certain characteristics to carry out its attacks, rather than viewing terrorism as the result of decentralized individual choices. In the model, the terrorist group (attacker) decides which demographic "type" to send through a security checkpoint. The security officials (defender) decide which type or category to subject to an extensive search. Some types (for instance, young males with military and ideological training) are more "reliable" than others (women and children) in the sense of having a higher probability of mounting a successful attack once they pass undetected through a security checkpoint. If the attackers were not selective, sending a mix of types with equal probabilities, then defenders would use limited resources to search the most reliable types who would cause the greatest damage if they passed security. Consequently, attackers would respond by sending less reliable types more often. In turn, defenders would respond by defending less reliable types more often. In equilibrium, attackers and defenders should not have any additional incentive to change their strategies. We show that, in a mixed strategy Nash equilibrium, there is a tendency to use low-reliability attack strategies and high-reliability defense strategies. Thus, attack and defense strategic patterns are seemingly "misaligned," even though both players are rational and there are no surprises in terms of observed behavior.

Our profiling game is similar in spirit to a 2-person "hide-and-seek" game in which the Hider decides where to place an object and the Seeker decides which of *n* locations to search. The Hider wins if the locations do not match, and the standard zero-sum version of this game has (Hider, Seeker) payoffs of (1, 0) if the location decisions do not match and (0, 1) otherwise.[2] Our profiling game is similar in that the defender desires to match by searching the demographic type selected by the attacker, who in turn desires to select a type that does not match. However, our profiling game is different in that the different demographic "types" are assumed to have different success or "reliability" probabilities, an asymmetry that introduces asymmetries in the attack and defense probabilities across attacker reliability types. As a result, in some cases, only a subset of reliability categories are actually used in equilibrium, and within this subset the defense probabilities and the terrorists' attack probabilities are inversely related, with the government searching less reliable types less frequently and the terrorists sending them more frequently.

With only two locations, hide-and-seek games are sometimes referred to as "matching pennies" games, and such games have been used to analyze strategic play in professional sports contests, e.g. whether to pass or lob in tennis or to which side of the goal to direct a penalty kick or defense (Chiappori, Levitt and Groseclose, 2002). Matching pennies games have also been used to model predator-prey interactions, where the predator desires the match. These games are structured to be realistic for the biological applications considered, and payoffs can be stochastic. For example, the Avrahami, Güth and Kareev (2005) "parasite game" is a hide-and-seek game preceded by a move by "nature" that randomly determines which of two locations will have a food resource. The "producer" selects a location to look for food and harvests if it is present. The "parasite" then selects a location to search, in order to steal the food if it has been harvested by the producer. Our profiling model also has stochastic payoffs, but the focus is on a monotonic array of reliability parameters for a set of possible attacker types, and on the monotonic arrays of attack and defense probabilities that can extend to *n*-categories or types.

[2] These payoffs are used for the 4-location version of the hide-and-seek game examined by Rubinstein, Tversky and Heller (1996). In the experiment the locations were arrayed along a line and the mixed-strategy Nash equilibrium for this symmetric game involved equal hide and seek probabilities of 0.25 for each location. The results of the experiment showed that subjects select endpoint locations somewhat less frequently than the middle locations. Crawford and Iriberri (2007) reconsider the Rubinstein, Tversky and Heller data and conclude that a model of levels of strategic thinking ("level-k") can explain the observed patterns when subjects have psychological aversions and attractions to endpoints. We will discuss the level-k analysis further in the results section below.

We use an experiment to assess the extent to which individual decisions are consistent with theoretical predictions of misaligned profiling. The experiments are motivated, in part, by the somewhat counter-intuitive nature of equilibrium patterns of the randomized strategies. In particular, the theory produces a paradox of misaligned profiling: in equilibrium the high reliability categories are searched more intensively, even though they are used less intensively by the terrorist organization. Field experiments with "professional" terrorists and security officials to test these predictions would be expensive and controversial, if possible at all, and the results would surely be confidential. Instead, we rely on laboratory experiments, which provide the ability to replicate and control the environment, even though the laboratory environment is admittedly highly simplified. The results of the experiment reveal behavioral patterns that are consistent with predicted patterns. However, we also find several interesting behavioral deviations, with defenders tending to search more reliable attacker types more often than predicted.

2. Theoretical Model and Predictions

To address the problem of profiling consider a simple two-player model of screening at security checkpoints. The first player represents a government agency that is attempting to discover terrorists, e.g., military officers at a checkpoint, Transportation Security Agency officials at an airport security counter, or other agencies dealing with homeland defense, such as the Coast Guard, the Border Patrol, the Customs Service, or the Immigration and Custom Enforcement department of the Department of Homeland Security. Their objective is to identify any terrorists attempting to penetrate their checkpoint and, thereby, block an attack. The second player represents a terrorist group, which is assumed to be centrally directed, strategically rational, and motivated by a desire to penetrate the defenses and commit an attack. Experts agree that there is usually a strategy behind terrorists' actions. Whatever form it takes, terrorism is typically not random, or blind; it is a deliberate use of violence against civilians for political or religious reasons. Therefore, following the spirit of most game theoretic literature on terrorism, we model terrorists as rational actors; for an excellent survey, see Sandler and Arce (2007).[3] In

[3] Our models of strategic terrorism can be modified in a straightforward manner to include the possibility of exogenous, decentralized sources of terrorism of the emotional or "home-grown" variety.

what follows, we will refer interchangeably to the terrorist as the "attacker" and to the government agency as the "defender."

The main motivation for our research can be illustrated with a simple model taken from Kydd (2011). The population of individuals passing through the checkpoint is divided into $n \geq 2$ different observed categories or types. A successful attack by a person from any category will result in a gain of G for the terrorist and a loss of L for the government security agency. In what follows, we will refer interchangeably to terrorists as "Attackers" and to the government agency as the "Defender." Let the n categories be indexed by i, where the lower the value of i, the higher the chances an undetected person would succeed in carrying out the attack. The probability that a person from category i will succeed if undetected is denoted by r_i, with $r_1 > r_2 > r_3 > \ldots > r_n$. We will refer to r_i as the "reliability" of category i. Each category could be identified and determined by different criteria, such as age, gender, and country or region of origin, religion, or any other observable personal characteristic. For instance, the *New York Times* on June 27, 2011 reported that in a remote area in central Afghanistan "insurgents tricked an 8-year-old girl … into carrying a bomb wrapped in cloth that they detonated remotely when she was close to the police vehicle." This is an example of an attack using a person from a less reliable type or category who is less likely to be searched.[4]

We assume that the defender selects one category to search, and that the search is fully effective in detecting the attacker.[5] Hence, the attacker of type i who is searched would fail, and one who is not searched would carry out an attack with probability of success r_i. Let d_i denote the probability of defending against type i, and let a_i denote the probability of attacking with type i, with $0 < d_i, a_i \leq 1$. Then, a person from category i would succeed only if the defender searches another type and the attack turns out to be successful, an outcome that occurs with probability $(1-d_i)r_i$. The attacker, therefore, faces a tradeoff between sending high reliability people and sending others who are less likely to be searched. If the attacker uses type i with probability a_i, then the defender's probability of a loss from defending against type i is $(1-a_i)$ times the average reliability across all other types.

[4] To be clear, there is heterogeneity within a category, and the "reliability" of a category represents the average effectiveness of the best people in that category who can be successfully identified, recruited and trained, i.e., the "tail" of the distribution in terms of physical and emotional fitness, willingness to risk extreme injury or death, unwavering loyalty to the cause, and the ability to mask emotions and improvise to defeat unanticipated challenges.

[5] This assumption, used in Kydd (2011) and Basuchoudhary and Razzolini (2006), can be generalized to derive comparative statics predictions for an improvement in search technology (see Holt, Kydd, Razzolini and Sheremeta, 2011).

Before deriving the equilibrium, it is useful to provide some intuition. The equilibrium involves randomization, since a deterministic attack via one type would lead to a sure defense there, and a deterministic defense against one type would lead to a sure attack via another. In equilibrium with randomization, the expected payoffs for all decisions used must be equal, otherwise the player would prefer decisions with higher expected payoffs. The main result reveals a *paradox of misaligned profiling:* in equilibrium the high reliability types are searched *more* intensively, even though they are used *less* intensively by the terrorist organization. This is a paradox in the sense that the equilibrium pattern makes the defense strategy appear to be misguided, and hence, ineffective, which is not the case. Second, there will be a cutoff point beyond which the defender will not bother searching at all, so types with a low reliability score will be ignored and not searched. Third, the attack probabilities will actually increase as the category's reliability declines, up to a cutoff point, after which the probabilities will decline to zero. That is, a set of low reliability categories will not be recruited by the attacker nor searched by the defender. Within the set of higher reliability categories, the defense probabilities and the terrorists' attack probabilities will be inversely related, with the government searching less reliable categories less frequently and the terrorists sending them more frequently.

The key structural feature of our model is the difference in attack success reliabilities for different attacker types. The increasing probabilities of success of increasingly reliable attacker types introduce a stochastic element into the payoff functions, which differentiates our model from standard hide and seek games in which the person doing the hiding will win with probability 1 if the seeker looks elsewhere. In our model, the monotonic ranking of attacker type reliabilities generates misaligned, monotonic attack and defense profiles, with defense probabilities increasing in attacker type reliability, and with attack probabilities inversely related to reliability, at least for the range of attacker types that are actually used with positive probability in equilibrium. The *n*-dimensional nature of the attack and defense decisions also differentiates our model from other games with stochastic payoff elements, such as penalty kicks games in soccer where the kick can go to one side or the other, or a two-player food search game with two locations.[6] In particular, the attack and defense probability predictions for the *n*-

[6] See Chiappori, Levitt and Groseclose (2002) for a discussion of the penalty kicks in soccer and evidence that patterns of penalty kicks for two European soccer leagues correspond to qualitative predictions of a mixed-strategy Nash equilibrium. The Avrahami, Güth and Kareev (2005) parasite game discussed earlier is another game with two locations and with probabilistic outcomes. If the producer and parasite search the same location, then the parasite

dimensional profiling model we consider are shown to be monotonic in the ranked reliability parameters and to be inversely related or misaligned.

The intuition behind misaligned profiling is surprisingly simple. An attacker choosing between n different reliability types would have an expected payoff of $G(1-d_i)r_i$ from selecting type i. It must be the case that these attacker expected payoffs are equal for all types that are used with positive probability in equilibrium; that is, $G(1-d_i)r_i = \pi_a$, where π_a is the attacker payoff in equilibrium. Since the right side of the equation is a constant with respect to i, the inverse relationship between d_i and r_i on the left makes it clear that that attacker types with higher reliability are defended with higher probabilities. An analogous argument can be constructed from equating defender expected payoffs and removing terms in sums that cancel out, to show that attack success probabilities are equalized: $a_i r_i = a_j r_j$ for all types i, j used with positive probability, and hence, attack probabilities and associated type reliabilities are inversely related in equilibrium.

These results can be illustrated for the special case of a two-category zero-sum game, in which a successful attack results in payoffs of 1 for the attacker and -1 for the defender. To be willing to randomize, the attacker's expected payoff for sending either type must be equal, i.e., the product of the probabilities of not being searched and of succeeding are the same for both categories: $(1-d_1)r_1 = (1-d_2)r_2$. Since $1-d_2 = d_1$, we obtain a single equation, $(1-d_1)r_1 = d_1 r_2$, which can be solved for defense probability against type 1:

$$\text{Defense probability against type 1: } d_1 = r_1/(r_1+r_2).$$

Since type 1 is more reliable, $r_1 > r_2$, it follows that $d_1 > d_2$, or the defender searches the more reliable type 1 more often, which is intuitive.[7] Conversely, for any given attack probabilities a_1 and a_2, a defense against type 1 will result in a loss with probability $a_2 r_2$ whereas a defense against type 2 will result in a loss with probability $a_1 r_1$. Since $a_2 = 1-a_1$, the equality of expected defender payoffs results in an equation, $(1-a_1)r_2 = a_1 r_1$, which can be solved for the attack probability via type 1:

$$\text{Attack probability via type 1: } a_1 = r_2/(r_1+r_2).$$

---

takes the food resource from the producer, and if not, the producer keeps the food. This game has a non-constant sum structure, since sites that are more likely to contain food offer a higher payoff sum for producer and parasite. Our profiling game is, on the other hand, a constant sum game.

[7] The use of expected payoffs implicitly assumes risk neutrality, but a reformulation in terms of expected utility yields the same monotonicity, since the attacker's expected utility would still be decreasing in d_i and increasing in r_i, so equating expected payoffs would still imply that high reliability types are searched more often.

Hence the counter-intuitive result that $a_1 < a_2$; that is, in equilibrium, the attacker uses the *more* reliable type 1 *less* often, since $r_2 < r_1$.[8, 9]

3. Experimental Design and Procedures

Our objective is to evaluate the extent to which observed behavior is consistent with theoretical predictions. To this end, we design two treatments with ($r_1 = 0.67$, $r_2 = 0.33$) and ($r_1 = 0.80$, $r_2 = 0.20$). For simplicity, we selected reliability parameters that sum to 1 in both treatments: $r_1+r_2 = 1$. In the 67/33 treatment, type 1 is twice as reliable as type 2. Theoretical prediction is that the defender searches type 1 with probability $d_1 = 2/3$ and the attacker uses type 1 with probability $a_1 = 1/3$. Conversely, the probability of defense against type 2 is $d_2 = 1/3$ and the probability of attack via type 2 is $a_2 = 2/3$. The best-response functions for this game are shown in Figure 1.[10] For example, if the probability of an attack via type 1 is low (left side), the probability of a defense against type 1 is low (bottom left part of the figure). Conversely, if the probability of defense against type 1 is high (top), then the attacker will use type 1 with probability 0 (upper left side). The intersection of the best response lines determines the equilibrium, with a 2/3 probability of a defense against type 1, and a 1/3 probability of an attack via type 1.

In the 80/20 treatment, we increase the reliability of type 1 so that type 1 is four times more reliable than type 2, i.e., $r_1 = 0.80$ and $r_2 = 0.20$. The prediction of the theory is that since type 1 is more reliable, the probability of defense against type 1 should *increase* to $d_1 = 4/5$, while the probability of attack via type 1 should *decrease* to $a_1 = 1/5$. Correspondingly, the probability of defense against type 2 should decrease to $d_2 = 1/5$, while the probability of attack via type 2 should increase to $a_2 = 4/5$.

---

[8] Given the theoretical predictions about the defense and attack probabilities, the expected payoff of the attacker is $r_1r_2/(r_1+r_2)$ and the expected payoff of the defender is $1-r_1r_2/(r_1+r_2)$. If $r_1 = 2/3$ and $r_2 = 1/3$, for example, the attacker and defender expected payoffs are 2/9 and 7/9. The equilibrium would not be altered if there were a third attack type with $r_3 = 1/6$, since a unilateral deviation to use this type would only provide a 1/6 expected payoff even though it is undefended, which is less than the 2/9 payoff the attacker can expect using types 1 or 2 in equilibrium.

[9] The general model with n categories is solved similarly by equating expected payoffs and rearranging terms. Attack probabilities can be expressed as ratios of the reciprocals of the reliability parameters $a_i = (1/r_i)/\Sigma(1/r_i)$ for all i that are used in equilibrium, and defense probabilities are $d_i = [1-(s-\delta)a_i]/\delta$, where s is the number of categories used in equilibrium, and δ is the probability of a successful defense, which is set equal to 1 in the experiment.

[10] The best-response functions should actually have sharp corners, since perfectly rational players are assumed to respond sharply to small differences in expected payoffs. The lines in the figure are plotted with a slight amount of curvature to make it easier to visually separate the two best-response lines.

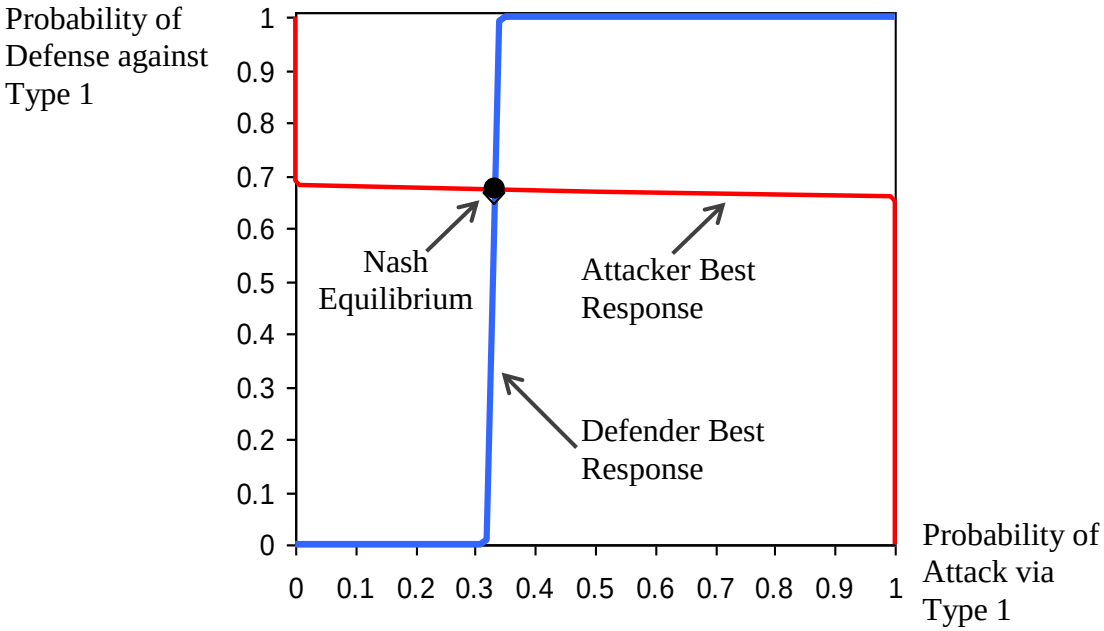

Figure 1: Best Responses and Equilibrium in the 67/33 Treatment

Subjects for the experiment were recruited from student populations at the University of Virginia, with the promise that they will "participate in a research experiment" and will receive a fixed payment of $6 plus additional cash earnings, which will depend on their own and others' decisions. When subjects arrived in the lab, they were seated in visually separated cubicles with networked computers. The software kept track of total earnings, and subjects were paid in cash at the end of each session, after they signed receipt forms. A total of 144 subjects participated in the experiment with 72 subjects (36 pairs) participating in one treatment and 72 subjects (36 pairs) participating in the other treatment. Instructions for the experiment are included in the Appendix. The screen displays listed the most reliable category on the left side for half of the subject pairs, and on the right side for the other half. The experiment was run in a series of sessions consisting of 12 to 18 subjects each, with fixed pair matching for 50 rounds. Subjects in each pair were given the role of attacker or defender and stayed in that role in all rounds of the experiment.

In each round, the attacker chose a category corresponding to a type of terrorist agent (type 1 or type 2), and the defender chose a profiling strategy, or type of person to search. If the selected categories matched then the attack failed. If the selected categories did not match, then the attack success probability was determined by the reliability of the attacker's category choice. A successful attack resulted in a fixed payoff of 1 dollar to the attacker and a loss of 1 dollar for the defender. The payoffs were added to private incomes of $1 for the defender and $0.60 for the

attacker in each round. These outside incomes were selected to equalize final payoffs and were private information (defenders did not know attacker incomes, and vice versa). On average subjects earned $26 and the experiment lasted for about 30 minutes.[11]

4. Results

Table 1 reports for the two treatments the predicted and average observed probabilities of defense and attack for type 1, respectively from the first round, the second half and all 50 rounds of play.

Table 1: Experimental Data and Predictions

	Treatment			
	67/33		80/20	
	d_1	a_1	d_1	a_1
Predicted probabilities	0.67	0.33	0.80	0.20
1st round average data	0.81	0.36	0.92	0.19
2nd half average data	0.78	0.30	0.88	0.20
All rounds average data	0.79	0.31	0.88	0.22

We begin by analyzing the data from the 67/33 treatment. Figure 2 displays the average defense and attack probabilities, while Figure 3 displays the average attack and defense probabilities for each of the 36 fixed pairs. As predicted by the theory, there is strong "misaligned profiling." Specifically, the defenders search more reliable type 1 with higher probability than less reliable type 2 (0.79 versus 0.21; Wilcoxon signed-rank test, p-value < 0.01, n = 36).[12] Conversely, the attackers employ more reliable type 1 with *lower* probability than less reliable type 2 (0.31 versus 0.69; Wilcoxon signed-rank test, p-value < 0.01, n = 36).

Relative to theoretical point predictions, we find that defenders tend to defend against the more reliable type 1 more than predicted (0.79 versus 0.67; Wilcoxon signed-rank test, p-value < 0.01, n = 36). The behavior of attackers is not significantly different from theoretical predictions for category 1 (0.31 versus 0.33; Wilcoxon signed-rank test, p-value = 0.42, n = 36).

[11] The data for all sessions are available on the web at: http://www.people.virginia.edu/~cah2k/profiling_data.htm. The 10 sessions are numbered adn24-adn28, adn30, adnp1-adnp4. For each session, the data table provides links to a graph of data averages for each round, a color-coded verbal/numerical summary of each attacker-defender interaction, and a presentation of all data in column form, which was used to create an Excel data file for each session (also included in the table).

[12] In conducting statistical tests, we treat the average over all 50 rounds of the experiment by the same subject as one observation. The results also hold if we analyze only the first round of the experiment and are available upon request.

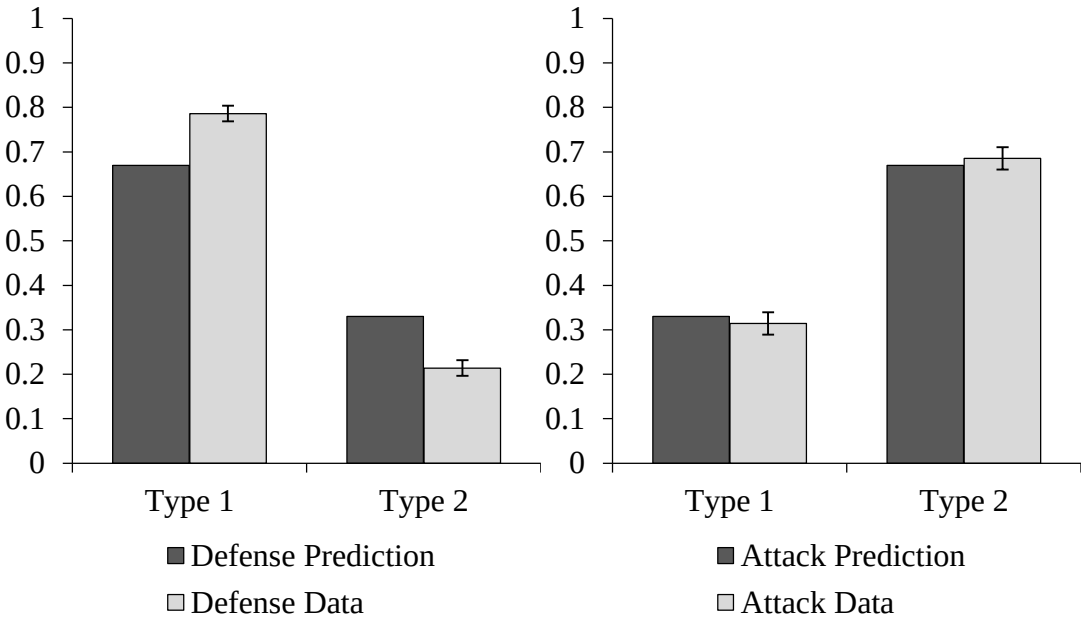

Figure 2: Average Data for All Rounds and Theoretical Predictions in the 67/33 Treatment

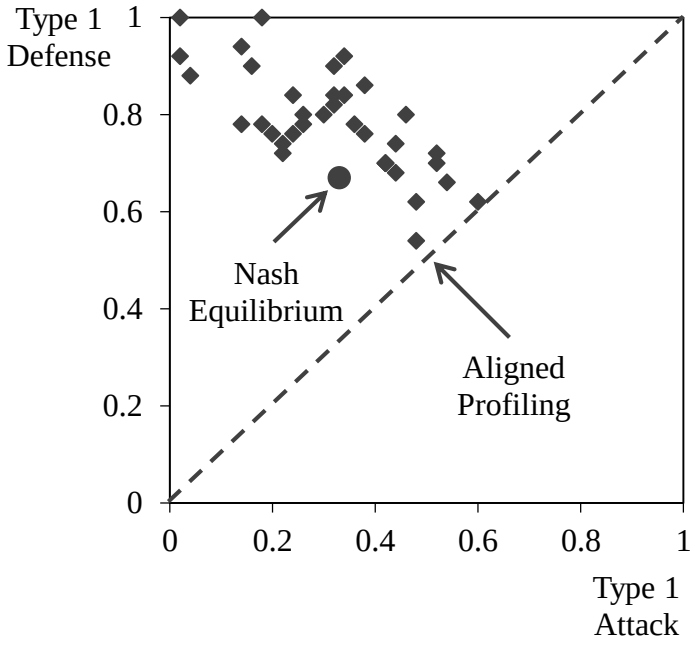

Figure 3: Attack and Defense Probabilities for 36 Fixed Pairs in the 67/33 Treatment

One may argue that subjects in a role of defender search type 1 more often because this option is presented to the left from type 2. In a related literature on multi-battle contests, called Colonel Blotto games, where attackers and defenders allocate resources on multiple battlefields, it is documented that subjects often exhibit allocation bias towards left battlefields (Chowdhury,

Kovenock and Sheremeta, 2013).[13] Nevertheless, it is unlikely that allocation bias can explain our data, since in half of the sessions type 1 option was presented to the left from type 2 and in the other half it was presented to the right. Moreover, the probability that the defender searches type 1 is virtually the same disregarding whether type 1 is located to the left or to the right of type 2 (0.79 versus 0.79; Wilcoxon rank-sum test, p-value = 0.72, $n_1 = n_2 = 18$).

The pattern of data that we observe in the 67/33 treatment is also observed in the 80/20 treatment with $r_1 = 0.80$ and $r_2 = 0.20$. Figures 4 and 5, displaying the average defense and attack probabilities for the 80/20 treatment, show even stronger "misaligned profiling." Specifically, the defenders search more reliable type 1 with much higher probability than less reliable type 2 (0.88 versus 0.12; Wilcoxon signed-rank test, p-value < 0.01, n = 36). Conversely, the attackers use more reliable type 1 with much *lower* probability than less reliable type 2 (0.22 versus 0.78; Wilcoxon signed-rank test, p-value < 0.01, n = 36). As before, the defenders tend to defend against the more reliable attacker type 1 more than predicted (0.88 versus 0.80; Wilcoxon signed-rank test, p-value < 0.01, n = 36). This behavior cannot be explained by the allocation bias, since the probability that the defender searches type 1 is very similar in sessions where type 1 is located to the left versus sessions where type 1 is located to the right of type 2 (0.86 versus 0.90; Wilcoxon rank-sum test, p-value = 0.19, $n_1 = n_2 = 18$).

A comparative static prediction of the theory is that increasing reliability of type 1 should increase the defense probability against this type. This is exactly what we observe in the experiment. When the reliability of type 1 increased from $r_1 = 0.67$ to $r_1 = 0.80$, the defense probability against type 1 increased from 0.79 to 0.88 (Wilcoxon rank-sum test, p-value < 0.01, $n_1 = n_2 = 36$). The prediction for the attacker is the opposite, increasing reliability of type 1 should *decrease* the probability of using type 1. Again, we observe this in the experiment. When the reliability of type 1 increased from $r_1 = 0.67$ to $r_1 = 0.80$, the probability of using type 1 by the attacker decreased from 0.31 to 0.22 (Wilcoxon rank-sum test, p-value = 0.01, $n_1 = n_2 = 36$).

[13] For a comprehensive review of this literature see Dechenaux, Kovenock and Sheremeta (2014). The two most relevant studies that investigate the behavior of attackers and defenders are Kovenock, Roberson and Sheremeta (2010) and Deck and Sheremeta (2012). Both studies examine behavior in attacker-defender games, where the defender needs to win all targets, while the attacker needs to win only one target to secure the prize.

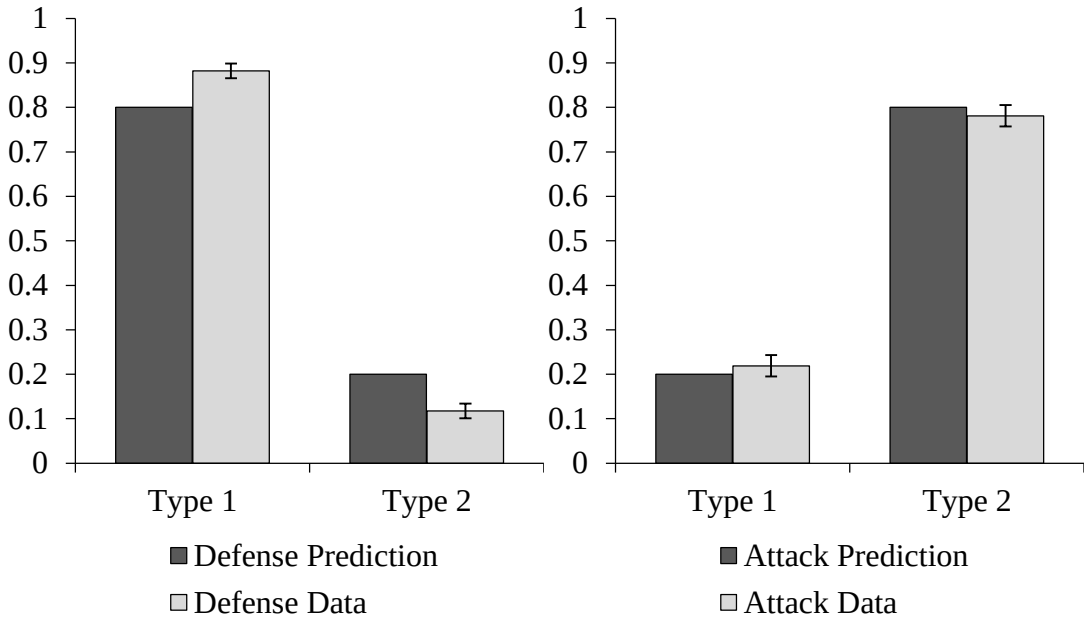

Figure 4: Average Data for All Rounds and Theoretical Predictions in the 80/20 Treatment

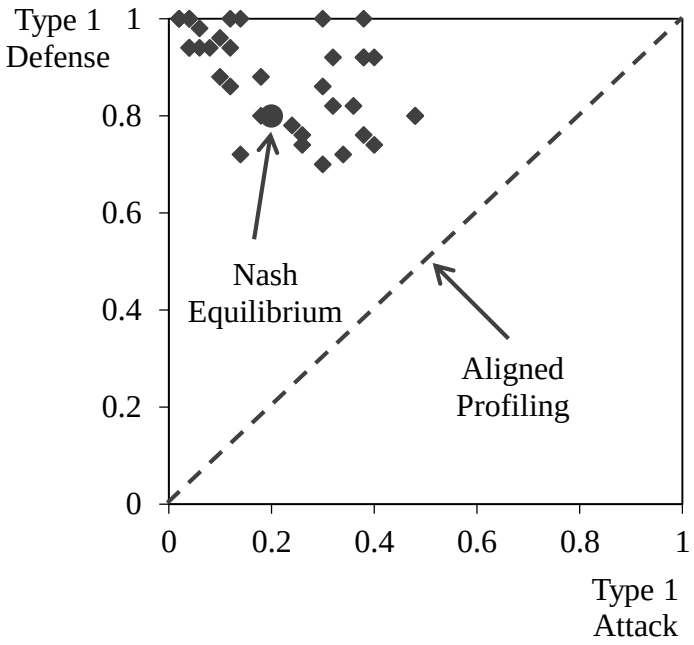

Figure 5: Attack and Defense Probabilities for 36 Fixed Pairs in the 80/20 Treatment

Although the data averages are close to Nash predictions for the two treatments, there is a clear tendency for defenders to defend against the more reliable type more often than predicted (the attackers' behavior, on the other hand, is quite close to the Nash predictions). These deviations from Nash predictions are not large in magnitude, but it is notable how the deviations

are largely concentrated among defenders in both treatments. These deviations are in an intuitive direction for defenders, who might find it natural to over-defend against high reliability attack types, whereas the Nash prediction for attackers is less intuitive, i.e. attack more often with the less reliable attacker type.[14] Even if this strategy is less intuitive, it does show up in the very first round of play, where it seems that attackers expect that defenses are more likely to be targeted to more reliable attack types. One way to model such strategic thinking about what the other person might do is to use a "level-*k*" analysis (Stahl and Wilson, 1994, 1995).[15] After the initial round of play in the experiment, however, a level-*k* analysis is less appropriate since subjects probably learn more from their own experience than from introspection and levels of strategic thinking. Moreover, a level-*k* analysis of best responses to level *k*-1 is the same for both of our treatments, so this approach cannot explain the salient treatment effect in the data, an effect that is predicted by the Nash equilibrium.

Another approach that we considered is the quantal response equilibrium (QRE) introduced by McKelvey and Palfrey (1995), which in this case adds curvature to the sharp-line best response function in Figures 1 and 2 and would tend to pull attack probabilities closer to 0.5. Such a smoothing, however, could not explain the observation that attack probabilities are close to Nash predictions, while the probability of defense of the high reliability type is even more extreme than predicted. It is worth noting that the QRE has been useful in explaining "own-payoff effects" in 2x2 matrix games (McKelvey and Palfrey, 1995; Goeree and Holt, 2001; Goeree, Holt and Palfrey, 2003, 2005) with asymmetries that differ from those that arise in our profiling model.[16] One asymmetry in the profiling game used in our experiment is that an attacker may feel some extra regret when a low-reliability attack type is used and is not defended against, but fails anyway. Several subjects mentioned this type of reaction in an informal

[14] We thank a referee for pointing this out.

[15] Level-*k* thinking would imply that level-0 attackers randomly choose between types 1 and 2, and level-0 defenders randomly search type 1 and type 2, with equal frequency. In response to level-0 behavior, level-1 attackers should use type 1 and level-1 defenders search type 1. Similarly, in response to level-1 behavior, level-2 attackers should use type 2 and level-2 defenders should search type 1. Further switching occurs at higher levels, so predictions would have to be based on estimated proportions of people at each level (see Crawford, Costa-Gomes and Iriberri, 2013). But note that these predictions (and analogous predictions for higher levels) are the same for the $r = 0.67$ and $r = 0.80$ treatments, in contrast with observed data patterns and Nash predictions.

[16] An example of the type of payoff asymmetry for which QRE has been useful in the past would be adding an extra cost for the attacker for training a low-reliability type. In a simple game with 2 reliability types, such change in the attacker's payoff function would affect the defense probabilities, while it would not affect the Nash attack probability associated with either reliability type, since those probabilities are constructed so to keep the defender indifferent between defense options. As noted previously, such "own-payoff effects" are observed in the data from laboratory experiments and are well explained by QRE models.

debriefing conversation after one of the sessions. Another possible approach to explaining deviations from Nash predictions could involve probability weighting, e.g. a tendency to overweight the relatively low probability (0.33 in equilibrium) of an attack with a reliable type may cause the defender to "over-defend" against that type.[17] While some of these approaches seem plausible, we believe that any *ex post* explanation would be tenuous at best without additional research.

6. Conclusions

This paper implements an experimental test of a game-theoretic model of equilibrium profiling. Attackers choose a demographic "type" from which to recruit, and defenders choose which demographic types to search. Some types are more reliable than others in the sense of having a higher probability of carrying out a successful attack if they get past the security checkpoint. Using a controlled laboratory experiment with financially motivated human subjects, we find strong support for game-theoretic model of equilibrium profiling. Consistent with theoretical predictions, the defenders search more reliable types with *higher* probability, while the attackers employ more reliable types with *lower* probability than less reliable types. However, we also find small (but systematic) deviations with defenders searching more reliable types more often than predicted. These deviations, however, do not obscure the clear pattern of misaligned profiling and the sharply different effects of attacker-type reliability parameters on predicted and observed attack and defense propensities.

There are several important implications of our findings. The Department of Homeland Security (DHS) in October 2010 issued the following statement: "As a precaution, DHS has taken a number of steps to enhance security. [….] Passengers should continue to expect an unpredictable mix of security layers that include explosives trace detection, advanced imaging technology, canine teams and pat downs, among others." As our theoretical model predicts and experimental results confirm, a security agency such as the DHS or the Transportation Security Agency must respond optimally to terrorist organizations' actions and pre-empt any terrorist attack by identifying terrorists within large groups of mostly innocent people. In this context, profiling is rational and the government should actually screen individuals according to their

[17] Probability weighting has been used to explain overbidding phenomena in a verity of conflict games (Sheremeta, 2013).

potential to be reliable recruits for the terroristic organization. The security agency should search more often the individuals belonging to the most reliable categories with an apparently unfair profiling practice. The intense search directed toward high reliability individuals should induce the terrorists to send less reliable categories more often. Our findings should be interpreted with caution, however, since our experiment does not shed light on the indirect effects of profiling, e.g., the possible increases in a propensity for terrorist activity among groups being profiled (Harcourt, 2007).

There are many possible avenues for future research. The general *n*-type version of the model presented in section 2 also predicts misaligned profiling for the types that are reliable enough (sufficiently high r_i values) to be used in equilibrium, with a minimum reliability cutoff that separates types that are used from those that are not. These predictions with *n* possible attacker types could be compared with data. In particular, it would be interesting to see whether high-reliability attack types are "over-defended" relative to the theoretical predictions in this richer setting, as observed in the simple model with only two attacker types. In addition, it would be interesting to investigate both theoretically and experimentally a non-constant sum version of the profiling game, where each player can choose to attack (or defend) more categories at once and the cost of attacking (or defending) is increasing in the number of categories attacked (or defended). Another avenue would be to introduce multiple attackers and defenders, with defenders trying to defend against multiple independent (or dependent) attackers. Also, it is important to examine the problem of profiling in the setting of incomplete information, i.e., when defenders and attackers know only the distribution of the reliability of each type. In such case, players would need time and experience to learn how reliable different types are. These are all very interesting questions and we leave them for future research.

References

Antonovics, K., and Knight, B.G. (2009). A New Look at Racial Profiling: Evidence from the Boston Police Department. *Review of Economics and Statistics*, 91, 163-177.

Avrahami, J., Güth, W., and Kareev, Y. (2005). Games of Competition in a Stochastic Environment. *Theory and Decision*, 59, 255-294.

Barak-Erez, D. (2007). Terrorism and Profiling: Shifting the Focus from Criteria to Effects. *Columbia Law Review*, 29, 1-9.

Barnes, K., and Gross, S. (2002). Road Work: Racial Profiling and Drug Interdiction on the Highway. *Michigan Law Review*, 101, 653-754.

Basuchoudhary, A., and Razzolini, L. (2006). Hiding in Plain Sight – Using Signals to Detect Terrorists. *Public Choice*, 128, 245-255.

Chiappori, P. A., Levitt, S., and Groseclose, T. (2002). Testing Mixed-Strategy when Players are Heterogeneous: The Case of Penalty Kicks in Soccer. *American Economic Review*, 92, 1138-1151.

Chowdhury, S.M., Kovenock, D., and Sheremeta, R.M. (2013). An Experimental Investigation of Colonel Blotto Games. *Economic Theory*, 52, 833-861.

Crawford, V.P., and Iriberri, N. (2007). Fatal Attraction: Salience, Naivete, and Sophistication in Experimental Hide-and-Seek Games. *American Economic Review*, 97, 1731-1750.

Crawford, V.P., Costa-Gomes, M.A., and Iriberri, N. (2013). Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications. *Journal of Economic Literature*, 51, 5-62.

Dechenaux, E., Kovenock, D., and Sheremeta, R.M. (2014). A Survey of Experimental Research on Contests, All-Pay Auctions and Tournaments. *Experimental Economics*, forthcoming.

Deck, C., and Sheremeta, R.M. (2012). Fight or Flight? Defending Against Sequential Attacks in the Game of Siege. *Journal of Conflict Resolution*, 56, 1069-1088.

Ellmann, S. (2003). *Racial Profiling and Terrorism*. New York Law School Review, 46, 675-730.

Goeree, J.K., and Holt, C.A. (2001). Ten Little Treasures of Game Theory and Ten Intuitive Contradictions. *American Economic Review*, 91, 1402-1422.

Goeree, J.K., Holt, C.A. and Palfrey, T.R. (2003). Risk Averse Behavior in Asymmetric Matching Pennies Games. *Games and Economic Behavior*, 45, 97-113.

Goeree, J.K., Holt, C.A. and Palfrey, T.R. (2005). Regular Quantal Response Equilibrium. *Experimental Economics*, 8, 347-367.

Harcourt, B.E. (2007). Muslim Profiles Post-9/11: Is Racial Profiling and Effective Counter Terrorist Measure and Does it Violate the Right to be Free from Discrimination? In B. J. Gould and L. Lazarus, Eds., *Security and Human Rights*. Oxford: Hart Publishing, 73-98.

Heymann, P.B., and Kayyem, J.N. (2005). *Protecting Liberty in an Age of Terror*. Cambridge, MA: MIT Press.

Holt, C.A., Kydd, A., Razzolini, L., and Sheremeta, R.M. (2011). Theory and Experiments on Profiling and Terrorism. Draft NSF Proposal.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, 263-291.

Knowles, J., Persico, N., and Todd, P. (2001). Racial Bias in Motor Vehicle Searches: Theory and Evidence. *Journal of Political Economy*, 109, 203-229.

Kovenock, D., Roberson, B., and Sheremeta, R.M. (2010). The Attack and Defense of Weakest-Link Networks. Chapman University, Working Paper.

Kydd, A. (2011). Terrorism and Profiling. *Terrorism and Political Violence*, 23, 458-73.

McKelvey, R., and Palfrey, T. (1995). Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, 10, 6-38.

Rubinstein, A., Tversky, A., and Heller, D. (1996). Naive Strategies in Competitive Games. In Albers, W., W. Guth, P. Hammerstein, B. Moldovanu and E. van Damme, Eds., *Understanding Strategic Interaction - Essays in Honor of Reinhard Selten*, Springer-Verlag, 394-402.

Sandler, T., and Arce, D.G. (2007). Terrorism: a game-theoretic approach. In T. Sandler and K. Hartley, Eds., *Handbook of Defense Economics*: *Defense in a Globalized World*, vol. 2, Amsterdam: North-Holland.

Sheremeta, R.M. (2013). Overbidding and Heterogeneous Behavior in Contest Experiments. *Journal of Economic Surveys*, 27, 491-514.

Stahl, D., and Wilson, P. (1994). Experimental Evidence on Players' Models of Other Players. *Journal of Economic Behavior and Organization*, 25, 309-327.

Stahl, D., and Wilson, P. (1995). On Players' Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior*, 10, 218-254.

Appendix - Instructions

- **Rounds and Matching:** The experiment consists of a number of rounds. Note: You will be matched with the same person in all rounds.
- **Interdependence:** Your earnings are determined by the decisions that you and the other person make.
- **Roles:** In each pair of people, one person will be given the role of "attacker" and the other will be given the role of "defender." You role will be (attacker or defender) in all rounds.
- **Locations:** There are 2 locations that will be designated as: L1 and L2. An attack can be targeted to either of these locations, and a defense can be augmented at either of these locations.
- **Attack Success Probabilities:** An attack will always fail at a site that is defended. An attack at an undefended site may or may not succeed, and the probability of attack success at undefended sites will depend on the site, as explained later.

Instructions (page 2)

Location:	**L1**	**L2**
Attacker Gain from Successful Attack:	**$1.00**	**$1.00**
Defender Loss from Successful Attack:	**$1.00**	**$1.00**
Position Your Asset	○ **L1**	○ **L2**

- **Attacker Gains**: If an attack is successful at a location, the attacker earns an amount of money shown in the Attacker Gain row of the table, for that location.
- **Defender Losses**: A successful attack at any location results in a loss to the defender, as shown in the Defender Loss row of the table, for that location.
- **Available Assets**: Each attacker has 1 asset to allocate (1 attack), and each defender has 1 asset to allocate (1 defense).

Instructions (page 3)

- **Attack Outcomes:** If a site is defended, an attack at that site will fail. The chances of a successful attack at an undefended site depend on the site, as shown in the table below, which will be reproduced for you when you submit your decision.

Location:	**L1**	**L2**
Probability of Attack Success at a Defended Site:	**0**	**0**
Probability of Attack Success at an Undefended Site:	**0.67**	**0.33**

- **Random Outcome Determination**: Consider an attack on site L2. If this site is defended, the attack will fail. If this site is undefended, the probability of attack success

attack at that site is 0.33. You can think of this process as spinning a Roulette wheel with stops labeled 1, 2, ... 100 and the outcome is a success if the wheel stops on a number that is less than or equal to 33, so a probability of 0.33 corresponds to 33 chances out of 100 of attack success.

- **View Failed Attacks:** After all decisions are made and confirmed, the defender will always be able to see where an attack occurred, even if it fails.
- **Visibility of Defense Assets:** The attacker will NOT be able to see where a particular defense asset is located prior to making an attack decision.
- **Cause of Failed Attack:** At the end of each round the attacker's results table will indicate whether a site was defended or not. Thus if an attack does fail, the attacker will be able to see whether it failed because the site was defended or because the attack at an undefended site failed due to random causes.
- **Private Incomes**: In addition to the earnings, losses, and costs that result from asset allocations and attack outcomes, each person will receive a fixed income in each round. Attackers and defenders may have different private incomes, which are not public information. As a (Attacker or Defender), your income will be: **$*.**** per round.

Instructions (summary page)

- There will be one or more rounds in this part of the experiment, and the final round will not be announced in advance.
- You will be matched with the **same** person in all rounds.
- In each group, there will be **1 attacker** and **1 defender**.
- Your role is that of ********
- Defenders each have 1 asset to allocate across the 2 sites, and attackers each have 1 asset to allocate to one of the 2 sites.
- If a site is defended, an attack at that site will fail. The chances of a successful attack at an undefended site depend on the site, as shown in the table below.
- A successful attack at a site reduces the earnings for the defender, as indicated by the Defender Loss for that site.
- A successful attack at a site increases the earnings for the attacker, as indicated by the Attacker Gain for that site.
- The defender will always be able to see where an attack occurred in previous rounds, even if it fails.
- The attacker will not be able to see where a defense asset is located (before the attack decision is made).
- In addition, your payoff will be raised by an amount **$*.**** in each round, which is your private income, not observed by the others with a different role.
- **Special Earnings Announcement:** Your cash earnings will be **50%** of your total earnings at the end of the experiment.