

STRATEGIC IMPROVEMENTS
TO THE BEHAVIORAL
ASSESSMENT LEVEL II
EVALUATION

REPORT FROM THE
INDEPENDENT REVIEW PANEL
FOR ASSESSOR EVALUATION

MARCH 2012

PREPARED FOR THE
U.S. NAVAL RESEARCH LABORATORY
WASHINGTON, DC

SPONSORED BY THE
DEPARTMENT OF HOMELAND SECURITY
SCIENCE AND TECHNOLOGY DIRECTORATE
HUMAN FACTORS DIVISION
WASHINGTON, DC

EXECUTIVE SUMMARY

OVERVIEW

The Department of Homeland Security Science and Technology (DHS S&T) Directorate coordinated an independent effort to review and recommend enhancements to a study being planned by the Transportation Security Administration (TSA) to assess the effectiveness and efficiency of an enhanced behavior detection protocol known as Behavioral Assessment Level II (BA2), or "Assessor." DHS S&T assembled an independent panel of experts representing multiple organizations and disciplines to review and comment on TSA's research methodology, data collection efforts, analysis plan, and preliminary results. The review panel was charged with two goals: (1) review the emerging operational assessment process and provide recommendations to strengthen it; and (2) review the evaluation protocol and data collected from the Boston and Detroit screening pilot studies and make recommendations related to lessons learned and to the analyses that can be meaningfully conducted to support effectiveness and efficiency questions.

CONCLUSION

At a high-level, there was unanimous concern among review panel members that the limitations presented by the design of TSA's proof-of-concept studies were too severe to recommend using the data produced by the studies to assess the detection capability of the Assessor protocol. In general, the review panel felt that the design of the proof-of-concept studies produced data that were too confounded to allow meaningful interpretation of the effectiveness of the protocol. Panel members were more optimistic, however, that the proof-of-concept studies could provide information that would be meaningful for assessing the operational impact (i.e., efficiency data) of Assessor.

To address members' concerns, the panel recommended two new field experiments to deal with the confounding variables problem (Study 1) and to mitigate the risk posed by the extremely low incidence of high-risk outcomes (Study 2). Recommendations for Study 1 specified a detailed new design for an operational study of Assessor. Study 2 outlined a supplemental red team approach, and recommended that another group of subject matter experts be convened sometime in the near future to address complexities presented by the second experiment.

Detailed recommendations are made throughout the report in relation to specific design concerns. Each recommendation is embedded in a discussion of the nature of the concern and its potential impact on the conclusions that can be drawn from the current study. Specifically, recommendations address modifications to experimental and control groups, procedures for the random assignment of both

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

passengers and TSOs, concomitant operations of SPOT, the configuration of checkpoints, conceptualization of outcome measures, the utilization of particular metrics and statistical techniques, and methods to monitor protocol compliance, among others. The panel also raised concerns about the quality of the data being used to evaluate the protocols, and followed with a recommendation that modifications be made to future data collection instruments and to the computer systems being used to solicit and store the data.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

TABLE OF CONTENTS

1 CHARGE TO THE PANEL	1
2 INDEPENDENT PANEL REVIEW PROCESS.....	1
3 LIMITATIONS	2
4 BACKGROUND	2
4.1 SCREENING PASSENGERS BY OBSERVATION TECHNIQUE (SPOT)	2
4.2 BEHAVIORAL ASSESSMENT LEVEL II (ASSESSOR)	3
5 ASSESSOR EVALUATION: PROOF-OF-CONCEPT STUDIES.....	3
5.1 STANDARD TREATMENT (BASELINE).....	5
5.2 EXPERIMENTAL TREATMENTS (PROOF-OF-CONCEPT)	6
Treatment Design.....	6
Variations in Treatment Implementation	7
5.3 ASSESSING OPERATIONAL IMPACT.....	8
Throughput and Staffing.....	9
Efficiency	9
5.4 ASSESSING DETECTION CAPABILITY.....	12
Confounding Influences.....	12
Effectiveness Measures.....	13
5.5 RECOMMENDATIONS FOR USE OF CURRENT DATA	19
Assessing Detection Capability with the Current Design.....	19
Assessing Operational Impact with the Current Design.....	20
6 ASSESSOR EVALUATION: TWO FIELD EXPERIMENTS	20
6.1 STUDY 1—NEW DESIGN FOR AN OPERATIONAL STUDY OF ASSESSOR	20
Review Panel Assumptions.....	21
Treatment and Control Group Selection	23
Randomization Procedures	25
Location and Implementation Procedures	29
Compliance Monitoring.....	30
Outcome Measures.....	31
Metrics and Analytical Considerations.....	32
Random Selection Processes	38
Statistical Tests and Confidence Interval Estimation	40
Duration of Study.....	40
Benefits and Limitations of Study 1.....	43
6.2 STUDY 2—A FIELD EXPERIMENT UTILIZING RED TEAMS	44
Purpose of Red Teaming	44
Confederate Selection	46
Study Protocol.....	47

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

<i>Location and Implementation Procedures</i>	48
<i>Duration of Study</i>	48
<i>Statistical Analysis</i>	49
<i>Benefits and Limitations of Study 2</i>	49
6.3 A SUMMARY OF REVIEW PANEL RECOMMENDATIONS	50
<i>Study 1 Recommendations</i>	50
<i>Study 2 Recommendations</i>	52
7 REFERENCES	52
APPENDIX A: ACRONYMS AND ABBREVIATIONS	A-1
APPENDIX B: POTENTIAL AIRPORTS FOR A FUTURE STUDY	B-1
APPENDIX C: INDEPENDENT REVIEW PANEL MEMBERS	C-1
ACKNOWLEDGEMENTS	C-4

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

1 Charge to the Panel

The Department of Homeland Security Science and Technology (DHS S&T) Directorate was charged with coordinating an independent effort to review and recommend enhancements to a study being planned by the Transportation Security Administration (TSA) to assess the effectiveness and efficiency of an enhanced behavior detection protocol known as Behavioral Assessment Level II (BA2), or "Assessor". DHS S&T assembled a panel of experts from multiple organizations and disciplines to review and comment on TSA's research methodology, data collection efforts, analysis plan, and preliminary results. The review panel was charged with two goals: (1) review the emerging operational assessment process and provide recommendations to strengthen it; and (2) review the evaluation protocol and data collected from the Boston and Detroit screening pilot studies and make recommendations as to lessons learned and to what analysis can be meaningfully conducted to support effectiveness and efficiency questions.

2 Independent Panel Review Process

The review panel was comprised of subject matter experts representing the following substantive areas: behavioral research, transportation security, law enforcement, security-based targeting applications, job and occupational analysis, and program evaluation (see Appendix B for a list of members). Two face-to-face meetings and one conference call were conducted in support of the effort. At the first panel meeting on November 7, 2011, TSA presented an overview of the agency's risk-based screening efforts to the panel. The presentation included the Assessor and Screening Passengers by Observation Technique (SPOT) protocols, along with their research objectives, metrics, and preliminary results. Following the first meeting, additional documents were disseminated to the committee with the expectation that members would review and comment on the materials.

Panel members met face-to-face a second time on November 22, 2011 in conjunction with TSA staff to discuss the evaluation and to consider alternative study designs. A draft report was compiled from these discussions, and returned to the panel for comments. Comments were received from all panel members by January 18, 2012. This information was reviewed and incorporated into a second draft of the report. The second draft was transmitted to the panel on February 4, 2012 with a request that panel members conduct a second review and to confirm that the report attended to panel members' comments and other concerns. A final conference call was held on February 15, 2012 to discuss members' review of the second draft and to address any issues still outstanding. Recommendations were finalized on February 22, 2012.

The report from the Independent Review Panel for Assessor Evaluation synthesizes the members' comments and suggestions for strengthening the design of future evaluation efforts. The report is

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

organized around common themes expressed by panel members in the areas charged. In cases where members' opinions were unanimously in favor of a particular option, a recommendation is presented without alternative. When opinions from panel members diverged, a primary recommendation is paired with an alternative approach. The issues identified during the debate are discussed.

3 Limitations

The review panel report represents a summary of issues identified by panel members that confound the results of the Behavioral Assessment Level II evaluation currently being considered by TSA. Recommendations are made for improving the design and implementation of future studies. The report is not intended to be a definitive, in-depth review of all of TSA's behavior detection programs, but rather it represents a compilation of design considerations for the evaluation of a particular behavior detection protocol (Assessor) as implemented within the TSA operational arena. While a specific proposal is put forward for an Assessor Evaluation pilot, its design is specific to a particular context, and will not be appropriate for all behavior detection evaluation studies.

4 Background

4.1 SCREENING PASSENGERS BY OBSERVATION TECHNIQUE (SPOT)

In fiscal year 2007, the TSA began a broad-scale implementation of a behavior-based air passenger screening program known as Screening Passengers by Observation Technique (SPOT). The program employs Behavior Detection Officers (BDO) to observe passengers moving through airport security checkpoints, and to compare their behaviors against a set of known risk indicators. SPOT indicators were derived from behavior analysis programs employed by law enforcement and security personnel in the U.S. and abroad. Passengers who are observed by BDOs to exhibit a threshold number of high-risk behavioral indicators are referred to secondary screening for additional scrutiny and questioning. The purpose of further screening is to investigate the observed behavior in more depth in an effort to either corroborate or dismiss the concern. If a BDO is unable to dismiss the behavior, the case is escalated and the passenger is referred to a Law Enforcement Officer (LEO).

During fiscal year 2010, the TSA partnered with DHS S&T to conduct a large, year-long evaluation of the SPOT program with a goal of formally assessing screening accuracy. The SPOT Referral Report Validation Study was conducted in 43 airports between December 1, 2009 and October 31, 2010, and included over 70,000 random referrals and approximately 23,000 Operational SPOT referrals. Findings from the study indicated that the SPOT behavior detection method identified more high-risk passengers and inconvenienced fewer innocent travelers than would be achieved by flagging passengers at random. Specifically, the SPOT protocol was shown to be 4 to 9 times more likely to identify passengers engaged

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

in illicit activity than a random protocol. This seemingly strong result must be tempered, however, by noting that the true positive rates estimated by the study ranged between 0.09% to 1.13% across the four outcomes investigated, suggesting that under the current implementation, over 98% of illicit actors escape detection altogether. This result is likely explained by several factors, including the rareness of the event, and low instrument sensitivity in the context of sparse staffing levels.

4.2 BEHAVIORAL ASSESSMENT LEVEL II (ASSESSOR)

In June 2011, the TSA implemented an enhanced behavior detection capability known as Behavioral Assessment Level II (Assessor) at select airports. The enhanced protocol utilizes SPOT program BDOs for its implementation, and includes interaction of BDOs with 100% of passengers at the Travel Document Checker (TDC) station. Assessors ask approximately 3-6 questions of each passenger near the initial document check and checkpoint queue. Assessors receive SPOT training (or refresher training) and undergo Aviation Passenger Assessor (APA) training, which includes a five-day instructional course taught by a security contractor followed by an on-the-job-training certification process. The on-the-job-training procedure requires an additional 8-32 hours of focused practical application under the direct supervision of mentors to achieve certification. If the Assessor-certified BDO observes a suspicious sign, more time is spent engaging the passenger in an effort to resolve concerns. If concerns are resolved, the passenger is allowed to continue through the standard screening protocol; if the BDO is unable to resolve concerns then the passenger will be required to undergo more intrusive screening procedures and, like SPOT, may be referred to a LEO.

A major difference between the SPOT program and the Assessor model is that SPOT observes behaviors from a distance. Assessor engages passengers in face-to-face conversation, using direct questioning techniques and document review to uncover suspicious indicators. The behavioral indicators deemed "suspicious" by each technique are generally different, although some overlap exists.

5 Assessor Evaluation: Proof-of-Concept Studies

In order to ensure that the added demands of Assessor are worth the investment involved, TSA is engaged in an evaluation of the efficiency and efficacy of the Assessor program. The Assessor program was recently piloted as a Proof-of-Concept (PoC) at two airport checkpoints: Terminal A checkpoint at Boston-Logan International Airport (BOS) and Red-1 checkpoint at Detroit Metropolitan Wayne County Airport (DTW). Data to evaluate the assessment were collected at Boston-Logan between June 6 and December 15, 2011, and at Detroit between July 18 and December 15, 2011.

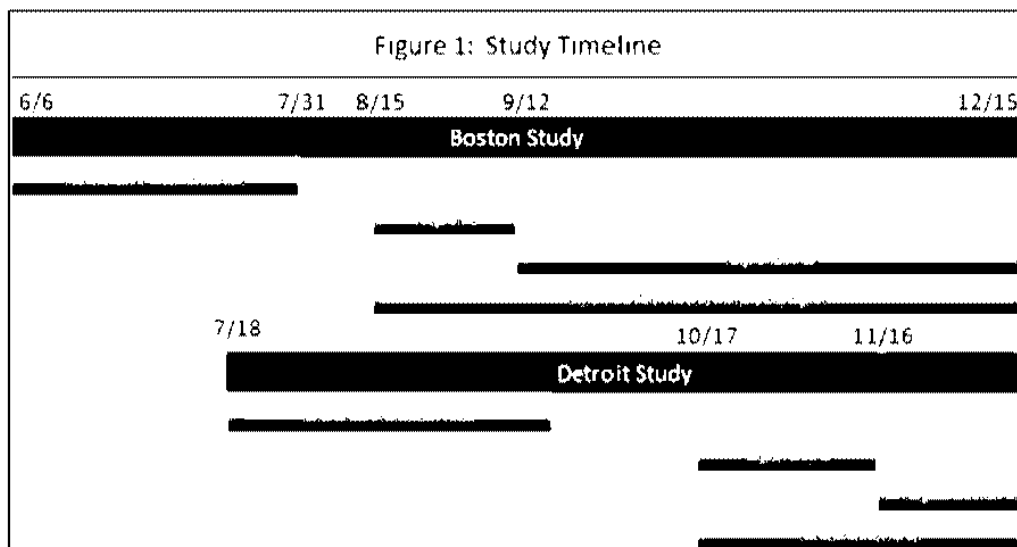
A detailed discussion of the PoC studies designs, measures, and results are presented in the remainder of this section. The PoC studies are discussed in detail here in order to provide context and information relevant to the Panel's recommendations for a stronger evaluation study design (see section 6). *The*

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

detailed discussion and presentation of results that follows is not intended to imply endorsement of the PoC studies by the independent review panel. Rather, the review panel noted a number of limitations in the PoC studies. These limitations are mentioned throughout the remainder of this section.

The design of the Proof-of-Concept is probably best described as a quasi-experimental, non-equivalent group study. Quasi-experimental designs are often used when it is difficult or impractical to randomize the assignment of subjects to treatment and control groups. Random assignment is a desired study characteristic because it minimizes the differences between the treatment and control groups. Without



randomization, confounding variables are less likely to be evenly distributed between treatment and control groups, making causation difficult to establish.

Groups in the Assessor PoC are judged to be non-equivalent because subjects were allocated to treatment and control groups through a non-random time-based procedure. Figure 1 above provides a graphical depiction of the time-based group allocation method used by the PoC studies. For example, the Assessor PoC implemented at Boston-Logan International Airport assigned passengers to control and treatment groups based on the time of transit through a single checkpoint (Checkpoint A). The group that moved through the checkpoint between June 6 and July 31 is referred to as the "Baseline" group. The "Baseline" group was subjected to current standard screening procedures and serves as the control group for the study. The group referred to as the "Proof-of-Concept group" transited the same checkpoint (Checkpoint A) at a separate point in time (Aug 15 through December 15). The "PoC group" served as the experimental group and received standard and Assessor-based screening.

Several changes were made to the Detroit-based study protocol as a consequence of preliminary results observed in Boston prior to the implementation in Detroit. These differences, as well as the differences

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

generated by the various "Options" tested (noted in Figure 1), are discussed in more detail in Section 5.2.

5.1 STANDARD TREATMENT (BASELINE)

The standard operating procedures (SOP) implemented at checkpoints include the following:

- TDC SOP: Passengers' travel documents are examined by a Transportation Security Officer (TSO) serving at the TDC position. If discrepancies cannot be resolved, a supervisor is notified¹ and *Selectee Screening* is conducted. Checkpoint SOP LEO notification guidelines are followed.
- SPOT SOP: Passengers are observed for checklist behaviors. A passenger exhibiting greater than a threshold number of behaviors is sent to secondary screening where *SPOT Referral Screening* is conducted. SPOT SOP LEO notification guidelines are followed. The SPOT SOP requires LEO notification when additional items are uncovered during the bag search and/or if signs of deception are observed.
- Secondary Security Screening Selection (SSSS, or Quad-S) SOP: Passengers who have not passed pre-screening checks are referred to secondary where *Selectee Screening* is conducted.

All SPOT Referrals and TDC/SSSS Referrals are subjected to the following identical screening procedures:

- Advanced Imaging Technology (AIT) screening or Walk-through Metal Detector (WTMD) (b)(3) 49 U.S.C. § (b)(3) 49 U.S.C. §
- (b)(3) 49 U.S.C. § Pat-Down
- Explosive Trace Detection (ETD) of accessible property
- Physical search of accessible property

In addition to the procedures described above, passengers who are referred based on SPOT scores receive additional scrutiny by BDOs. BDOs continue to observe individuals during the referral screening procedures for indications of stress, fear, and deception. SPOT BDOs also engage in a "casual conversation"² with passengers in an effort to elicit information that may serve to resolve the case. If SPOT scores reach the SPOT SOP thresholds for LEO notification, the passenger will become a mandatory LEO referral. SSSS referrals are screened by TSOs, not BDOs.

¹ Conflicting information was received from TSA with regard to the referral protocols utilized by TSOs serving as the TDC. One version stated that a supervisor must be notified before a TSO serving as a TDC could refer a passenger for additional screening (or to a Law Enforcement Officer). A second version stated that TSOs had latitude to make referrals independent of supervisor approval.

² "Casual conversation" is a specific sub-protocol of the overall SPOT protocol.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need-to-know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

5.2 EXPERIMENTAL TREATMENTS (PROOF-OF-CONCEPT)

Treatment Design

The primary experimental treatment consists of modifying operational screening procedures such that 100% Behavior Assessor interaction with passengers is added at the TDC position. Behavior Assessors are responsible for deploying the new behavior detection technique. BDOs who participated in the operational assessment were required to complete the following training courses in preparation for serving in the Behavior Assessor role:

- Aviation Passenger Assessor (APA) training
- On-the-job (OJT) training
- Screening of Passengers by Observation Techniques (SPOT) refresher training
- Travel Document Check (TDC) training

The highest performing participants in the APA training course were designated as Assessors during the PoC and were used to constitute the PoC group (treatment group).

In terms of implementation details, the Assessor protocol excludes TSA employees, air crew, and airport employees from questioning. Instead, the standard treatment was applied to these individuals. In addition, SPOT and Quad-S processes continued without change during the PoC.

As shown in Figure 1, four different screening options were tested in the PoC. The two options tested in Boston assessed optimal placement and function of the BDO at the checkpoint:

- **Option 1:** Assessors interact with each passenger in proximity to but independent of the TSO conducting travel document checks. The BDO asks the passenger 3-6 questions after the TSO completes the travel document check.
- **Option 2:** Assessors observe and engage each passenger while also performing TDC duties. A separate TSO serving as the TDC does not operate in this scenario.

Preliminary assessments at Boston suggested that passengers reacted more favorably to Option 2, possibly because of its similarity to standard TDC operations. Discussions with the Boston Assessor team indicated that passengers were confused and annoyed by having to wait in two different lines and show their documentation twice, once to the TDC and once to the Assessor. As a result of these considerations, Option 1 was discontinued in Boston and was not used in Detroit. Instead, Option 2 was used during the entire Detroit PoC.

During the first 15 days of the Boston pilot study, Assessors referred suspicious passengers to other Transportation Security Officers (TSOs) for *Selectee Screening*. Beginning September 1, Assessors began to conduct *Selectee Screening* on their own referrals, presumably utilizing a combination of SPOT- and Assessor-based indicators to resolve the case. Assessor-conducted screening occurred in Boston

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need-to-know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

through the end of the Proof-of-Concept (mid-December). Detroit employed a more structured approach (see Options 3 and 4 below), employing both sets of tactics during different timeframes to formally assess whether *Selectee Screening* or *SPOT Referral Screening* procedures produced more optimal results.

Back-end screening procedures were manipulated in Detroit in the following manner:

- **Option 3:** Assessor-based referrals are screened using *Selectee Screening* procedures. Selectees receive Advanced Imaging Technology (AIT) screening (if available), Walk Through Metal Detector (WTMD) screening, Property X-Ray, (b)(3):49 U.S.C. § 11 Pat-Down, Explosive Trace Detection (ETD) (b)(3):49 U.S.C. § 11 ETD search, and physical search of accessible property.
- **Option 4:** Assessor-based referrals are screened using *SPOT Referral Screening* procedures whereby BDOs engage in a non-custodial interview (casual conversation) with the passenger, and a comprehensive search of the referred individual's property is conducted. During SPOT Referral screening procedures, BDOs closely observe individuals and mentally calculate any additional SPOT point accrual. BDOs also make notes about the case and formally document the points and other notes in the SPOT Referral database. Otherwise, SPOT referrals receive AIT screening, WTMD screening, X-Ray of property, (b)(3):49 U.S.C. § 11 Pat-Down, and ETD and physical search of property.

Comparisons between the experimental and control groups are of primary interest. Specifically, within each airport the PoC group was compared to the Baseline data. There is also interest in comparing results between Options 1 and 2 (especially for the efficiency measures), and between Options 3 and 4 (especially for the effectiveness measures). However, notwithstanding problems of interpretation, the especially low base rates of illicit activity observed in Detroit made this comparison unachievable.

Variations in Treatment Implementation

In addition to some of implementation differences described above, TSA's assessment team documented several additional variations in how the protocols were implemented across the two PoC sites. Variations between the sites were noted in terms of: (1) the type of personnel utilized as Assessors, (2) procedures used for implementation and secondary screening, (3) the days and times Assessor shifts were scheduled, and (4) the length of the shift scheduled. Table 1 provides more detail about implementation variance.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15, and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Table 1. Comparison of Assessor Experimental Designs

Variable	BOSTON	DETROIT
ASSESSOR BACKGROUND	• BDOs	• BDOs, TSIs, and FAMs
TREATMENT OPTIONS	• Assessor conducted according to Option 1 (8/15/11-9/11/11) • Assessor conducted according to Option 2 (9/12/11-12/15/11)	• Assessor conducted according to Option 2 for the entire PoC period (10/17/11-12/15/11)
SECONDARY SCREENING OPTIONS	• Checkpoint Selectee Screening (8/15/11 - 8/30/11) • SPOT Referral Screening with a BDO partner (9/1/11-12/15/11)	BDOs • Checkpoint Selectee Screening (10/17/11-11/15/11) • SPOT Referral Screening with a BDO partner (11/16/11-12/15/11) TSIs and FAMs • Checkpoint Selectee Screening as only option
WEEKLY SCHEDULE	• PoC conducted several hours per day per shift	• PoC conducted every day from 0400-1900
SHIFT LENGTH	• Entire AM or PM Assessor shift (usually 3-4 hours)	• Entire AM or PM Assessor shift (7 or 8 hours)

Other personnel, in addition to BDOs, functioned as Assessors in Detroit during the PoC. These individuals included Federal Air Marshalls (FAMs) and Transportation Security Inspectors (TSIs). FAMs and TSIs were trained in the Assessor program alongside the BDOs. Unlike their BDO counterparts, though, FAMs and TSIs had not received SPOT training prior to Assessor training. Boston used only BDOs, who all have SPOT training and behavior detection experience.

One exception not addressed in Table 1 relates to the configuration of the checkpoint utilized in Detroit. Boston implemented the Assessor protocol at one checkpoint feeding one sterile area. The checkpoint chosen for use in the Detroit Assessor PoC (Red-1) represented just one of four checkpoints feeding a common sterile area. As a result, passengers were allowed to bypass the PoC checkpoint and to enter the sterile area through an alternate route (Red-2, Red-3, or Red-6). This problem severely limits the interpretation of results from Detroit, and is discussed in Section 5.4. A final difference between the two sites is that TSA conducted a "Pre-check" pilot project at one of the three other checkpoints (Red-2) during the Proof-of-Concept timeframe.

5.3 ASSESSING OPERATIONAL IMPACT

TSA's assessment team collected data to construct throughput and efficiency metrics to address the question of whether the Assessor implementation was feasible given the structure of checkpoint

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

operations at Boston and Detroit. Several aspects of checkpoint operations were measured during the Baseline and PoC assessments, including throughput and staffing measures, numbers of selectees, average processing times, average wait times, and passenger refusals/complaints. In this section, we report the data as provided to the review panel by TSA, and add interpretations as required. The results of the PoC studies serve as context for design considerations.

Throughput and Staffing

Information about passenger volume and the staffing provided for each of the PoC checkpoints is provided in Table 2. These data show that the average passenger throughput in Detroit on a daily basis is about half the throughput in Boston. Thus, Boston employs a greater number of lanes (8 versus 5), BDOs (66 versus 37), and Assessor personnel (66 versus 47).

Table 2. Key Measures of Throughput and Staffing for Boston and Detroit

THROUGHPUT AND STAFFING	Boston	Detroit
<i>Checkpoint</i>	Terminal A (8 lanes)	Red 1 (5 lanes)
<i>Year of SPOT Program Implementation</i>	2004	2006
<i>Baseline Period</i>	6/6/11-7/31/11	7/18/11-9/16/11
<i>Proof-of-Concept Period</i>	8/15/11-12/15/11	10/17/11-12/15/11
<i>Total Available Certified Assessors</i>	66 BDO	37 BDO, 5 TSI, 5 FAM
<i>Passenger Throughput</i>		
Total Throughput (PoC)	360,281	323,383
Total Throughput (Baseline)	597,550	327,640
Total Spoken to by Assessor	313,293 (87%)	287,805 (89%)

Efficiency

Key operational measures for the PoC are provided in Table 3. These data demonstrate that by the middle of December, the Assessor PoC had screened over 600,000 passengers. Assessors referred 689 passengers to selectee screening (0.11%), resulting in 142 referrals (20.6%) to Law Enforcement Officers (LEO). The number of passenger refusals and formal complaints was very low (0.03%).

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1620. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Parts 15 and 1620, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1620.~~

Table 3. Key Operational Measures for Boston and Detroit during Proof-of-Concept

PROOF-OF-CONCEPT OPERATIONS	Boston	Detroit	Total
Total Passengers Spoken to by an Assessor	313,293	287,805	601,098
Passengers Receiving Additional Questioning at Podium	4,496	2,461	6,957
Passenger Refusals	95	48	143
Total Formal Complaints	3	8	11
Total Assessor Referrals (Selectees)	506	183	689
Total LEO Referrals (from Assessor Selectees)	137	5	142
Percent of Assessor Selectees Referred to LEO	27.1%	2.7%	20.6%

Table 4 shows that passenger processing time was reduced from 63 seconds to 44 seconds when the Assessor also handles TDC duties (Boston), suggesting that this configuration is more efficient than Option 1. On average, the Assessor program adds an additional 25 seconds of processing time to each passenger (44 vs. 19 seconds), increasing average processing times by approximately 140%.

Table 4. Passenger Processing Times during Proof-of-Concept

PROCESS TIME MEASURES (seconds)	Boston	Detroit
Baseline - TDC Cycle Time	18.6	17.0
Option 1 - TDC Cycle Time	22.2	
Option 1 - Assessor Cycle Time	40.5	
Option 1 - TDC + Assessor (22.2+40.5)	62.7	
Option 2 - Assessor as TDC	44.1	44.0

Given the increase in processing times, it is perhaps surprising that passengers did not experience significant increases in wait times. Table 5 provides evidence that the overall average wait time was essentially unchanged between the Baseline period and the Option 2 period. During non-peak hours, passenger wait times actually declined on average by about 30 seconds; during morning peak hours, wait times were unchanged. During afternoon peak periods, wait times increased by approximately a minute on average. This seemingly counterintuitive result was true even though the number of passengers processed per staff hour was similar between the groups (see Table 5, far right column). Assessors who served as TDCs processed on average only 2 fewer passengers per hour than TDCs did during the Baseline period.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Table 5. Average Passenger Wait Times in Boston during Peak and Non-Peak Hours

WAIT TIME MEASURES (minutes)	Overall	Non-Peak	AM Peak (0500-0700)	PM Peak (0530-1730)	Passengers per Staff Hour
Baseline	0:10:40	0:10:53	0:09:12	0:11:44	19.3
Option 1 - TDC + Assessor	0:10:25	0:09:24	0:09:12	0:17:25	18.1
Option 2 - Assessor as TDC	0:10:38	0:10:21	0:09:18	0:12:55	17.1

A multivariate analysis of one month of data from the Boston PoC³ suggested that TDC officers proceeded more slowly when more TDC positions were open. This result implies that officers may slow down or speed up depending upon queue conditions—implying that cycle times may be endogenous to queue wait times, not the reverse. Another explanation may relate to the fact that other bottlenecks in the process drive wait times. For example, the speed at which passengers can progress through the X-Ray and WTMD stations may at times be limited by volume, creating situations where TDCs lengthen their cycle times because the passenger cannot yet be sent to X-Ray/WTMD. Consequently, the degree to which cycle times can be used to predict staffing requirements is unclear.

Table 6. Passenger Processing Times during Proof-of-Concept For Selected Risk-based Traveler Characteristics

RISK GROUP	Boston	Detroit	
		Average Processing Time per Passenger (seconds)	Average Transit Time (seconds)
Military Personnel (N=58)		31.8	3.7
Passengers Over 65 (N=371)		41.8	4.3
Passengers Traveling with Children (N=643)		24.5	4.2
All Other Travelers (N=8,717)		41.1	4.1

The Detroit Proof-of-Concept gathered data on average processing and transit times for several different categories of passengers as presented in Table 6. The first three categories represent low-risk travelers. Assessors were instructed to expedite the screening of low-risk passengers. Passengers traveling with children were processed the quickest, followed by uniformed military personnel. Both

³ See The Screening of Passengers by Observation Techniques Program's Behavioral Assessment Level I Proof-of-Concept Baseline Report, dated August 8, 2011.

~~Control the Security Information (SSI).~~

~~WARNING: This report contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this report may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

figures suggest that Assessor standard operating procedures are being followed. Passengers over 65 were processed no more quickly than general travelers—a finding that is inconsistent with the SOP.

The final report noted that slightly longer transit times (the time required to move from the front of the queue to the podium) were observed for passengers over 65 (far right column of Table 5). However, the difference in transit time was not large enough to account for the increased cycle time observed. It is possible that operational pressures or other factors may influence

(b)(3) 49 U.S.C. § 114(r)

5.4 ASSESSING DETECTION CAPABILITY

Measures of effectiveness in the PoC studies are necessarily proxy measures. Few actual terrorism-related incidents are available for analysis. The original measures reported for the PoC included: (1) Serious prohibited and illegal items (SPII), (2) Artful concealments, (3) LEO arrests, and (4) High-risk outcomes representing membership in any of the previous three categories. These measures are used to estimate the effectiveness of the new Assessor procedure as compared to standard checkpoint procedures and SPOT referrals.

Two different proportions are evaluated for each type of illicit activity: (1) a checkpoint metric: the number of illicit actors detected per number of passengers transiting the checkpoint, and (2) a referral metric: the number of illicit actors detected per number of referrals to secondary. PoC results are compared to Baseline results through the use of the checkpoint metric. Assessor results are compared to SPOT results via the referral metric. The four original measures described above were later supplemented with additional information on referrals (referrals to secondary and subsequent referrals to a LEO) and the types of illicit activities observed.

Confounding Influences

The quasi-experimental design associated with the PoC studies allocates subjects to experimental and control groups through a non-random, time-based procedure. Lack of an appropriate control group has made it difficult to be certain that observed differences between the groups are due to the Assessor implementation alone. The data collected to address the effectiveness of the Assessor protocol may be affected by a number of uncontrolled and unobserved confounding influences. Systematic bias has been hypothesized to impact results in a number of ways. Examples include:

1. Experimental and control groups are comprised of different passenger populations, collected during different months, different seasons, and possibly for different flight destinations.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

2. Inherent differences exist in Assessor and SPOT protocols such that opportunities exist for Assessors to resolve suspicions before referral, while SPOT must first refer passengers to receive a similar opportunity. The outcome of this practice is that comparisons to SPOT may show Assessor to be proportionally higher than expected or justified.
3. The structure of the Proof-of-Concept likely created differences between the Assessor and SPOT groups in terms of resource and capability inequalities. More qualified BDOs were assigned to Assessor. Consequently, it is not clear that higher rates of detection are due to the Assessor protocol rather than to officer capabilities.
4. Month-to-month variation in detection rates under the Assessor program may fall within the standard range of variation observed for checkpoints without Assessor involvement. Extremely low base rates given the current design and implementation period make an analysis of the issue difficult, if not impossible.
5. Given the checkpoint configuration, the Detroit implementation allows high-risk passengers to "shop" for lower risk checkpoints—rendering the Detroit results un-interpretable.

The effectiveness data are, therefore, too problematic for panel members to be confident in the meaningfulness of the results. While the data are reported below, **the Independent Review Panel recommends that interpretations of effectiveness data be made cautiously (observed differences might not be due to Assessor), and that future designs attempt to resolve confounding factors (Recommendation #1).**

Effectiveness Measures

From Table 7 the reader can see that approximately 2 high-risk passengers per 100,000 were detected during the Baseline time period compared to 15 high-risk passengers during the Boston PoC period. These data suggest that the high-risk passenger detection rate is nearly 8 times greater during the PoC time period than the Baseline time period. As a point of information, a Fisher-Irwin "Exact" test of significance was applied to evaluate whether the difference between the two groups in detecting high-risk passengers could have been generated by chance.⁴ The results of this test are presented in columns 4 and 7 of Table 7.

⁴ Fleiss, Levin, & Paik (2003) argue that the Fisher-Irwin Exact test is a better test to use than the chi-squared test when expected cell sizes are small (typically < 5). By this criteria, expected cell sizes for Baseline results across all outcomes are small. In addition, the Fisher-Irwin Exact test may arguably be more useful than a chi-squared test when a cell contains a zero count. Because the exact test is "discrete" in nature and because significance levels are usually fixed (e.g., $p < .05$), the Fisher-Irwin test has been argued to be a more conservative test (Fleiss et al., 2003).

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Table 7: Number of Illicit Actors Detected per 100,000 Passengers
Transiting Checkpoint

Checkpoint Outcomes	Boston			Detroit		
	Baseline	PoC	p-value	Baseline	PoC	p-value
Number Assessed	597,550	360,281		327,640	323,383	
Number of High-Risk Passengers per 100K	1.8 (11)	15.0 (54)	0.000	0.0 (0)	2.8 (3)	n.s.
Number of Serious Prohibited / Illegal Items (SPII) per 100K	1.8 (11)	5.8 (21)	0.001	0.0 (0)	2.8 (2)	n.s.
Number of Artful Concealments per 100K	0.2 (1)	0.6 (2)	n.s.	0.0 (0)	0.3 (1)	n.s.
Number of LEO Arrests per 100K	0.0 (0)	9.7 (35)	0.000	0.0 (0)	0.3 (1)	n.s.

Statistically significant differences were observed between the Baseline and PoC outcomes for Boston. As a result, one might be tempted to conclude that Assessor is effective at identifying unknown high-risk passengers. However, *such a conclusion is not warranted on the basis of these findings*. As was discussed earlier, there are a number of other potential explanations for this pattern of results that cannot yet be ruled out. The low incidence of illicit activity viewed in combination with an unacceptably high potential for systematic bias make the interpretation of significance testing difficult at best, and possibly even unwise. Even small degrees of measurement error may yield results in which the magnitude of the effects are grossly over or underestimated, or even in the wrong direction. This means that a finding of improvement in Assessor accuracy over current procedures by a factor of 10 could actually be 5 or 15. Even worse, the finding could actually be a decrease in accuracy by the same magnitude(s).

If the error introduced into the study is random, i.e., not systematic, there would be no negative consequences to the application of inference (statistical testing). Minimizing the potential effects of systematic error through random assignment and a strong experimental design is vitally important to ensuring that appropriate conclusions can be drawn from a sample or a comparison of samples. Yet, it may be important to note that even a committed adherence to experimental best practices cannot fully overcome all of the challenges native to the problem with which TSA is wrestling. Even in a situation where the Assessor protocol was judged to be 99.9% specific to the illicit activity it was designed to detect, the very low incidence of the outcome nearly guarantees that secondary referrals will still mostly be judged to be false positives.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

In addition, at this low incidence rate small errors in classification (even if random) tend to result in significant over-estimation. Considering the number of passengers that are processed per day and the sample sizes observed, there could be a real and large measurable difference in the actual outcomes and expected outcomes of a program. The use of the odds ratios to calculate expected outcomes in terms of cost, accuracy, etc. without formally computing expected population values would still depend on those population values and would likely be misleading. This is an issue that even the reporting of traditional confidence intervals cannot overcome. In fact, Hemenway (1997) has argued that these types of rare event problems call for more informative confidence intervals that incorporate the impact of incorrect classification. For example, if you accept a 5% chance that as few as 1.5% of passengers were randomly misclassified, a 95% confidence interval for the accuracy of the impact of Assessor might show that the confidence intervals for the Baseline and PoC estimates overlap (suggesting no difference). The panel recommends that these types of issues be considered in future analyses.

Finally, it is noted that the magnitude of the difference between Baseline and PoC periods observed in Boston was not observed in Detroit, and was not judged to be significantly different from what might be expected due to chance. The establishment of the Assessor PoC at only one of the four checkpoints feeding the common Detroit sterile area may have allowed high-risk passengers to "shop" for lower risk checkpoints, diminishing the Red-1 base rate of illicit actors. No illicit actors were detected during the Baseline period.⁵ It is not known if Detroit's low rates are due to very low rates of malfeasance or to Detroit's inability to detect high-risk passengers.

Table 8. Percent of Referrals Correctly Identified at Boston and Detroit

Referral Outcomes	Boston			Detroit		
	SPOT	Assessor	p-value	SPOT	Assessor	p-value
Number Referred	74	506		48	183	
Number of High-Risk (HR) Outcomes	4 (5.41%)	52 (10.28%)	n.s.	0 0.00%	3 1.64%	n.s.
Number of Serious Prohibited / Illegal Items (SPII)	4 (5.41%)	20 (3.95%)	n.s.	0 0.00%	2 1.09%	n.s.
Number of Artful Concealments	1 (1.35%)	1 (0.20%)	n.s.	0 0.00%	1 0.55%	n.s.
Number of LEO Arrests	0 (0.00%)	35 (6.92%)	0.007	0 0.00%	1 0.55%	n.s.

⁵ The fact that no illicit activities were detected during the Baseline period in Detroit complicates the computation and interpretation of odds ratios, improvement rates, and statistical tests.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

The data shown in Table 8 provide information about positive outcomes per referral. No significant differences were shown to exist between SPOT-based referrals and Assessor-based referrals with regard to finding serious prohibited and illegal items. However, Boston showed evidence of an increase in the number of LEO arrests. A Fisher-Irwin Exact test suggests that arrest counts for the Assessor time period are sufficiently different from baseline SPOT counts as to be unlikely to be generated by chance. As with the statistically significant differences presented in Table 7, we offer the same caution. While the Fisher-Irwin Test can calculate the probability of generating a four-fold table by chance where one cell contains zero counts, it cannot address whether the Assessor protocol is responsible for the differences observed. One alternative explanation that has been offered is that the Massachusetts State Police (MSP) stationed one local LEO at Terminal A for the duration of the PoC, which likely improved LEO response. On the other hand, approximately 80% of the Boston arrests were for immigration violations, which fall under the jurisdiction of Customs and Border Protection (CBP) and Immigration and Customs Enforcement (ICE). None of the Detroit differences were significant.

Table 9 provides information about the number and type of illicit actors detected at a given checkpoint during the Baseline and PoC periods. As noted earlier, the Assessor protocol was only operated at the Red-1 checkpoint in Detroit. Further, TSA conducted another security program pilot project (called "Pre-check") at the Red-2 checkpoint at the same time the Detroit Assessor PoC was being carried out at Red-1.⁶

Table 9. Number and Type of Illicit Actors Detected at Checkpoint

Checkpoint Outcome	Boston		Detroit	
	Baseline	PoC	Baseline	PoC
High Risk Passenger				
Term A	11	54		
Red 1			0	9 ⁷
Red 2			0	5
Red 3			0	4
Red 6			0	0
TOTAL	11	54	0	18

(b)(3) 49 U.S.C. § 114(r)

⁷ Detected through any method (includes SPOT and X-Ray screening).

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

During the PoC, BDOs were responsible for the implementation of both the SPOT and Assessor protocols. A large majority of BDOs served in the Assessor PoC and may have been unavailable to staff the SPOT program. As a result, there has been speculation that the actual deployment of SPOT during the PoC may have been limited by resource availability. This hypothesis is not supported by the SPOT referral rates observed in Boston between the Baseline period and the PoC period (see Table 10a below). The SPOT referral rate increased from 12.4 to 26.4 per 100k passengers ($p < .001$); SPOT-generated LEO referrals increased from 1.3 to 3.9 per 100k passengers ($p < .05$). No increase was observed in Detroit (see Table 10b). This increase likely reflects the increased presence of SPOT at Checkpoint A. Prior to the PoC, SPOT operated at Checkpoint A about 50% of the time.

Table 10a. Frequency of Referrals at Boston Checkpoint A

Metric	BOS Baseline			BOS PoC		
Metric	SPOT	Assessor	All	SPOT	Assessor	All
Throughput		597,550			360,281	
Total Referrals	74		N/A	95	506	?
Referrals per 100k passengers	12.4		N/A	26.4	140.4	?
LEO Referrals	8		20	14	137	156
LEO Referrals per 100k passengers	1.3		3.3	3.9	38.0	43.3
Total LEO Referrals		8	20		151	156

Table 10b. Frequency of Referrals at Detroit Checkpoint Red-1

Metric	DTW Baseline			DTW PoC		
Metric	SPOT	Assessor	All	SPOT	Assessor	All
Throughput		327,640			323,383	
Total Referrals	48		?	46	183	?
Referrals per 100k passengers	14.7		?	14.2	56.6	?
LEO Referrals	0		4	3	5	21
LEO Referrals per 100k passengers	0.0		1.2	0.9	1.5	6.5
Total LEO Referrals		0	4		5	21

It is important to note that TSA policy does not allow BDOs to serve in an Assessor capacity until they pass a qualifying test designed to evaluate acquired Assessor capabilities. Individuals who did not pass the test continued to serve as BDOs in the SPOT program during the PoC. This policy may have resulted in important differences between the capabilities of the two BDO populations, with more qualified BDOs

~~SENSITIVE SECURITY INFORMATION (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Table 11 compares the frequency of and reason for LEO referrals for SPOT, Assessor, and Other Checkpoint personnel during the baseline and PoC periods for each airport. These data suggest—not unexpectedly—that Assessor-based referral reasons both differ from and overlap with SPOT-based referral reasons. The most common reasons for LEO referral by SPOT Officers (b)(3):49 U.S.C. § 114(n) (b)(3):49 U.S.C. § 114(n) The most common reasons for referrals by Assessors (b)(3):49 U.S.C. § 114(n) (b)(3):49 U.S.C. § 114(n). Other checkpoint personnel were more likely to make LEO referrals for possession of prohibited/illegal items. Assessor training provides supplemental document training. (b)(3):49 U.S.C. § 114(n)

(b)(3); 49 U.S.C. § 1114(r)

~~WARRANTS. This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

A different pattern was observed in Detroit. SPOT and Assessor reasons for referral were fairly similar.

In Detroit, other checkpoint personnel were (b)(3):49 U.S.C. § 114(r)

(b)(3):49 U.S.C. § 114(r)

It is also interesting to note that in Detroit, other checkpoint personnel made 5 times the number of LEO referrals than did either SPOT or Assessor.

The data shown in Table 12 describe the types of serious prohibited and illegal items detected in Boston during the Baseline and PoC time frames. While both airports increased the number of SPII detections, the increase was not significantly different when measured in terms of concomitant increases in passenger volume. In addition, a large proportion of SPII was discovered by other checkpoint personnel, not BDOs.

Table 12. Frequency and Type of SPII Discovered at the Checkpoint during both the Baseline and Proof of Concept Periods

Type of SPII	BOS	BOS	DTW	DTW
	Baseline	PoC	Baseline	PoC
(b)(3):49 U.S.C. § 114(r)				
TOTAL SPII	11	21	0	9

5.5 RECOMMENDATIONS FOR USE OF CURRENT DATA

Assessing Detection Capability with the Current Design

There was unanimous concern among review panel members that the limitations of the current study design are too severe to recommend the use of the effectiveness data for assessing the detection capability of the Assessor program. These data are likely to be useful only in terms of providing preliminary estimates of referral rates in the context of analyses to support other study design activities (e.g., power analyses). Referral rates in Detroit were much lower than Boston, potentially providing some information about the range of referral rates that might be observed if the study moves into other airports. The data may also serve as a future comparison point (with important caveats).

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Assessing Operational Impact with the Current Design

Review panel members were more optimistic that the operational impact data could provide useful and valid information. However, the interpretation of the cycle and wait time data in light of additional potential confounds may require some caution. Confounding influences may include passenger throughput, the number of open TDC positions and lanes, type of line (standard or first class), queue load at time of measurement, mix of documentation type typically presented by passengers (e.g., passports versus driver's license), time of day, and other staffing issues. Fortunately, the current study collected data on many of these influences. Preliminary analyses suggested that while cycle times were significantly longer under the Assessor protocol compared to standard security practices, passenger wait times did not increase overall. This seemingly counterintuitive result was true even when staffing levels were controlled (i.e., the number of passengers processed per staff position was equivalent in both groups).

As discussed in Section 5.3, further analysis revealed that TDC officers may have proceeded more slowly when more TDC positions were open, suggesting that officers may slow down or speed up depending upon queue conditions. Because the current study collected data on so many of the potentially confounding influences, panel members believe that multivariate analyses of the operational impact data are likely to yield useful and important information and should be fully analyzed. In addition, understanding the endogeneity issue described earlier (i.e., queue conditions may drive cycle times, not the reverse) is important for profitable use of the data.

As a final point, the panel notes that no data on operational efficiency were reported for the processes that occurred subsequent to the Assessor encounter. While these data may be routinely collected by other areas of the organization, it seems important to point out that to the extent that different levels of identified risk are likely to be associated with different risk mitigation practices, understanding the cost of each mitigation type will be important to the calculation of operational impact as TSA moves forward with proposed changes to the risk assessment process.

6 Assessor Evaluation: Two Field Experiments

In order to overcome the limitations described above, the Independent Review Panel recommends two field experiments, as described below. The major challenges addressed by these recommended studies are the confounding variables problem and the extremely low incidence rate for SPII and other high-risk outcomes.

6.1 STUDY 1—NEW DESIGN FOR AN OPERATIONAL STUDY OF ASSESSOR

In order to overcome the limitations of the work that has been conducted at Boston Logan and Detroit Metropolitan Wayne County Airports, we recommend a field experiment wherein passengers would be

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

randomly assigned to queues that are either operating Assessor or are not operating Assessor. The general assumption in this approach is that if passengers are randomly assigned to queues, given a sufficiently large sample, the actual rate of high-risk passengers transiting each queue type (Assessor or non-Assessor) should be about equal or at least suffer from less systematic selection biases. If the detected rate of high-risk passengers varies as a function of queue type, then it would support the conclusion that one queue type is more sensitive than the other queue type at detecting high-risk outcomes. Obviously, the hypothesis would be that detection rates of high-risk outcomes are higher in the Assessor queues than non-Assessor queues. If the opposite pattern were indicated, or no difference was found, the conclusion would be that the Assessor protocol is not more effective at detecting unknown high-risk passengers than the comparison protocol.

Fundamentally, this approach is straightforward. The challenge, of course, is that implementing such an experiment in the field will require attention to numerous details, including treatment and control group selection, randomization procedures, outcome measures, location selection, duration, other implementation procedures, and statistical analysis of resulting data. Recommendations regarding these details are made below. But first, the assumptions that guided the panel's review and recommendations are outlined.

Review Panel Assumptions

The review panel employed a set of assumptions to provide boundaries for the review and to frame recommendations. Several of the guidelines were provided by S&T program staff. For example, the panel was instructed to (1) impose the least complex design possible while ensuring the study remained scientifically defensible, and (2) defer potential operational concerns until after the scientific design phase has been completed. Resource constraints were addressed by setting study boundaries in terms of time and location. Other guidelines were adopted by the panel to provide clarification of an approach or other concern. For example, the panel felt the need to be clear in defining its preferred approach to measuring "effectiveness" (see 2a and 2b below).

Guiding principles included:

1. The Assessor Evaluation Study's primary goal is to estimate the overall effectiveness of the Assessor protocol for identifying high-risk passengers. Efficiency concerns are of secondary importance and are not addressed by the panel at this point in time.
2. The preferred measures of effectiveness are defined in terms of the following:
 - a. Of those passengers selected by Assessor, what proportion results in high-risk outcomes (Positive Predictive Value) compared to those selected by standard security procedures?

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

- b. Of the estimated true positives, what proportion was selected by Assessor (*Estimated*⁹ True Positive Rate) as compared to those selected by standard security procedures?

CAVEAT: In the absence of data enumerating the entire population of interest, true positive rates cannot be estimated without inspection results collected from a random sample of passengers referred to secondary screening. Compliance Measurement (CM) programs are often implemented in targeting domains to support such needs (among other purposes).

If concurrently collected random data cannot be made available, panel members noted that the data available from the FY2010 SPOT Base Rate Study could arguably be used to support the requirement. However, differences in the data collection timeframes make these data less than ideal for this purpose (and violate principle #6 below). Thus, use of the Base Rate Study is not recommended without some reservation. However, given analytical issues that may arise if random comparative data are not used, the Base Rate Study data may constitute the better option. These issues are discussed in more detail in the section entitled Metrics and Statistical Analysis (page 32).

3. A future evaluation study should be able to complete within 3 to 6 months from the point in time it begins.¹⁰ Realistically, however, the extremely low incidence of illicit activity increases the risk that a 3- to 6-month data collection period would be an insufficient timeframe in which to be able to provide within-airport estimates with a sufficient level of confidence. This risk may possibly be mitigated through higher random sampling rates, however.
4. Resource constraints make it unlikely that a future evaluation study will include more than 10 airports.
5. Goals of secondary importance, but which are of deep interest¹¹ include:
 - a. Estimating the effectiveness of Assessor as compared to SPOT.
 - b. Understanding the extent to which Assessor and SPOT identify different types of illicit activities.
6. A future evaluation study's control group(s) needs to be representative of the same population from which experimental group data are collected. In the current context, this requirement is achievable through the use of methods that randomly allocate passengers to treatment and control groups.
7. A future evaluation study will not require airport- or checkpoint-level comparisons to be made.

⁹ In absence of 100% inspection of the traveling public, true positive rates must be *estimated*.

¹⁰ The panel notes that this choice may make findings less generalizable. For example, if drug smugglers are especially active during the fall, and the study only includes data from the spring and summer, the study will have little to say about drug smugglers. However, if best practices are followed with regard to random assignment of passengers and TSOs to the experiment and control groups, this fact should not affect our conclusions about the effectiveness of Assessor with regard to say "fraudulent document" detection. There is also the potential that efficiency observations will be impacted. For example, winter coats through security, surges in travel around holidays, spring break, etc. will be likely to vary greatly depending on the airport.

¹¹ The panel acknowledges that the questions stated here do not represent the entire array of research questions that exist.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need-to-know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

8. Although the importance of the topic is noted, the study will not address methods for determining an optimal implementation strategy if staffing levels will not support the interaction of Assessors with 100% of passengers.
9. Due to the nature of a future evaluation Study, it is assumed that blinding¹² will not be possible.
10. It is assumed that outcome measures can be re-evaluated and potentially changed going forward. Specific recommendations are made regarding the principles that should be considered when making changes to outcome measures (see the section entitled *Outcome Measures* on page 30).
11. Given the effect sizes observed in the original SPOT Referral Report Validation Study (i.e., odds ratios between 4 and 9), it is assumed that a future Assessor evaluation study should plan for sample sizes that provide an 80% chance of detecting an odds ratio (power=0.80) of at least 3, but with other options provided. The study should utilize decision testing that results in no more than a 5% probability of making a Type I error.

Treatment and Control Group Selection

To achieve the Assessor Evaluation Study's primary goal—estimating the effectiveness of the Assessor protocol for identifying high-risk passengers as compared to standard security procedures—the review panel recommends a study design that compares targeting results for two groups: (1) Standard TSOs serving as the travel document checker (TDC) [Control Group], and (2) Standard TSOs trained in Assessor serving as the TDC (Experimental Group)—(Recommendation #2). When serving as the TDC, TSOs freely engage with passengers while remaining at a stationary podium. SPOT-trained BDOs, on the other hand, typically roam through the checkpoint, observe passengers mostly from a distance, and engage in a "casual conversation" only upon secondary referral. The job function assumed by a Transportation Security Officer trained in the Assessor technique more closely resembles the TDC function than the currently configured BDO function. This similarity in job function argues for the use of the Standard TSO serving as the TDC as the primary control group in the design of the experiment.

The experimental group should be comprised of Standard TSOs who have been trained in the Assessor technique. Two purposes are served by this proposal. First, by using TSOs in both groups, there is greater assurance that the control and experimental groups originate from the same population. Second, by removing BDOs from consideration, potentially confounding influences introduced by SPOT training and behavior detection experience are eliminated from the study.

In the PoC, BDOs were prohibited from becoming Assessors until certified. A consequence of this practice is that more experienced and capable BDOs were placed into the Assessor program; less capable BDOs remained assigned to the SPOT protocol. This practice raises the possibility that differences observed between SPOT and Assessor may have been due to officer capability differences

¹² In a double-blind study, neither the research subjects nor the persons administering the experiment are aware of the group (control or experimental) to which the subject has been assigned.

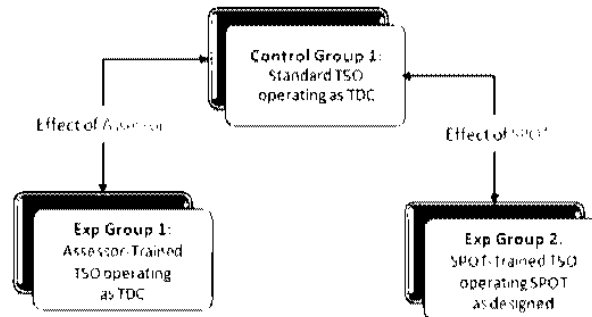
~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

rather than to the superiority of the Assessor protocol for identifying high-risk passengers. Random assignment of TSOs to experimental and control groups should ensure that unmeasured differences are spread across the two groups, mitigating the risk of bias. Thus, the review panel recommends that TSOs be randomly assigned to each group (Recommendation #3). It is further recommended that the requirement for Assessor certification be eliminated in future studies. It would be valuable to still conduct the Assessor skill assessment, but the scores should simply be recorded and made available for analysis, not used for selection (Recommendation #4). In addition, the panel recommends that additional skill assessments be provided to both groups of TSOs (Assessor-trained and non-trained) subsequent to—and possibly prior to—training for later analyses of performance (Recommendation #5). An added benefit to administering skill assessments specifically to the Assessor-trained group and keeping track of performance is that the information may be used to validate the threshold used for Assessor certification, and to also validate the use of the BDO selection instruments¹³ for the Assessor position. Examples of useful metrics to compare between the groups include: referral rates, compliance rates¹⁴, and effectiveness measures (sample sizes permitting).

The experiment will require that both the control and experimental groups operate similarly. Both groups must be provided equivalent latitude to refer passengers to secondary—a possible change to current procedures¹⁵. In addition, secondary screening procedures must be equivalent. If secondary screening practices are altered from current practices, training will need to be provided to both groups about the secondary referral process. If TDC refresher training is deemed useful to provide, it should be made available to both groups. Finally, because good data are critical to the risk assessment process, panel members recommend that (1) all data collection instruments be tailored to each officer's role in the experiment, and (2) all

Figure 2: Potential Experimental and Control Groups Operating as Designed in Relation to Effects of Interest



Note: Assumes that officers operate the Assessor protocol at the TDC and the SPOT protocol through the checkpoint. If SPOT training is provided to a standard TSO, the potential effect of prior behavior detection experience can be controlled.

¹³ BDO selection instruments represent a recently developed assessment toolset designed to select highly capable BDOs.

¹⁴ Compliance monitors should be blind as to Assessor test results.

¹⁵ It is not clear if all airports require supervisor notification before a TSO serving in the TDC capacity can refer a passenger for additional screening. Conflicting information was received from TSA on this point. Regardless, this requirement should be eliminated for all airports participating in the study.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

officers undergo in-depth training on data collection and submission procedures. It is especially important to make sure that officers understand how to document outcome measures through the use of structured database fields. (Recommendation #6).

As a point of information, the review panel considered a second experimental group (see Figure 2). The previously discussed groups (Control Group 1 and Experimental Group 1) were conceptualized to operate in the TDC position. SPOT, on the other hand, operates in a physical space much larger than the local vicinity of the TDC podium.

Since the design recommended by the panel focuses on TSOs and operates without the use of BDOs, BDOs remain available to serve in their traditional role. To the extent that the three groups can operate independently, a potential exists to make more valid comparisons of the effectiveness of Assessor as compared to SPOT, and to better understand the extent to which Assessor and SPOT identify different types of illicit activities. However, since SPOT training—under this conception—is not provided to a standard TSO within the same timeframe as the other two groups (BDOs likely received SPOT training prior to the experiment and have been operating as BDOs for some time), the potential effect of prior behavior detection experience (and other unmeasured factors) cannot be controlled. As a result, the panel would be reluctant to attribute all of the effect observed between experimental group 2 and the control group to SPOT. In addition, by allowing SPOT officers to continue to operate concomitantly with the other two groups, the risk increases that SPOT BDOs may consciously (or unconsciously) adopt some of the questions and questioning techniques utilized by the Assessors. There would be a natural inclination for this situation to occur. Therefore, **in order to further reduce the risk of study confounds, the review panel would prefer that SPOT not operate during the same time frame that an Assessor evaluation is running (Recommendation #7).**

If TDC/Assessor/SPOT comparisons are judged to be essential requirements, SPOT must be operated so that it does not impact the TSO's referral decision or the allocation of passengers to the Assessor-trained TSO or the Standard TSO lines. In addition, if a TSO decision to refer overlaps with a BDO decision to refer, appropriate data must be captured through both processes, and must be recorded independently of the other officer's data collection. It will also be important that both sets of information share a common identifier so that the two targeting events can be linked during the analysis.

Randomization Procedures

The prior section framed the importance of randomly assigning TSOs to control and experimental groups. **The review panel similarly recommends that passengers be randomly assigned to treatment and control queues (Recommendation #8).** This requirement must be implemented carefully as operational effects may have been generated by the randomization protocol chosen, and negatively affect the internal validity of the experiment. The strengths and weaknesses of two methods of

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

implementation are discussed below. We refer to one method as "Random Allocation," and to the other as "Random Rotation."

Random Allocation

The first randomization procedure envisions a scenario whereby passengers enter the security checkpoint and are assigned by a TSA representative to a particular queue. The passenger would not be allowed to select the queue he or she wishes to pass through¹⁶. One path at the divide will lead to a TSO who has been trained in Assessor, while the other divide will lead to a TSO who has not been trained in Assessor. In this scenario, the control and treatment groups operate simultaneously at each checkpoint. The panel judged the scenario where control and treatment groups operate simultaneously as the best method available to ensure control equivalence. Out of all potential methods, this method best ensures that no differences will exist between the treatment and control groups in terms of time, passenger mix, volume, or other factors. Although the initial concept would entail passengers entering a single checkpoint who are then randomly assigned to Assessor or standard TDCs at a fork or branch, the key element is random assignment of passengers to TDC type (Assessor or non-Assessor) during the same timeframe.

In this scenario, it may be necessary to account for the probable difference in wait times for Assessor versus non-Assessor queues. Data from the PoC studies suggested that the Assessor interaction process takes approximately 2.4 times longer, on average, (Boston: $44.1 \div 18.6 = 2.37$; Detroit: $40.5 \div 17.2 = 2.35$) than the traditional TDC. Ultimately, the panel's preference would be to increase staffing so that more Assessors are available to account for lengthier cycle times. However, if only a fixed number of podiums is available at a given checkpoint, in order to ensure that the TDC and Assessor lines maintain equivalent congestion patterns, cycle time differences may need to be considered when randomly assigning passengers to queues. Figures from the PoC studies suggest that it may be necessary to make approximately 7 random TDC assignments for every 3 random Assessor assignments (called "biased coin" allocation). It is important to note, however, that the 7:3 requirement may disappear when the X-Ray / AIT machine queues are full. When X-Ray / AIT machines are driving wait times, 1:1 allocation may be possible.

The panel agreed that it will be important to maintain a constant allocation ratio. This requirement is simplified if staffing of experimental and control groups can be adjusted as needed. However, in the event that the number of podiums is fixed, the requirement to maintain a constant allocation ratio will be difficult to fulfill. If the allocation ratio is implemented at 7:3 across the board, when queues are full and the x-ray/AIT lines are driving wait times, the TDC lines are likely to be longer (receiving 70% of passengers), leading to increased wait times. Passengers randomly selected to wait in the TDC line are likely to watch the Assessor line flow more quickly, increasing stress and/or frustration. If a 1:1

¹⁶ It is envisioned that all parties in a traveling group (co-travelers) would be assigned to the same random queue. Groups would not be split.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

allocation ratio is implemented across the board, then when queues are not full, the Assessor lines will be longer, again increasing anxiety, stress, and frustration for those passengers. With a fixed number of podiums and a constant allocation ratio, variability will simply be transferred to another point in the system to unknown effect.

Reviewers discussed the potential of separating the groups by a partition so that wait-time differentials could not be observed by passengers. Partitions were hypothesized to serve an additional benefit by minimizing the potential for control group contamination by the Assessor-trained TSOs. While random allocation is the gold standard for achieving control equivalence, an appropriate implementation strategy was not straightforward. **To the extent that staffing can be increased to account for the lengthier cycle times associated with the Assessor group, panel members recommend random allocation as the method to be used for ensuring passenger randomization. A gate-keeper will need to be employed with the random allocation method to ensure the protocol is administered appropriately. However, it will also be important to have a third party monitor randomization rates to ensure that they remain constant and that they produce expected passenger counts and distributions. (Recommendation #9)** As a final point, it is important to note that under the random allocation method, throughput for each treatment and control queue will need to be captured separately.

Assessor Rotation

The second potential implementation procedure—Assessor Rotation—is achieved by randomly rotating TSO groups to serve at a security checkpoint during specific time periods. In this scenario, the control and treatment groups operate separately—never at the same checkpoint at the same time (see Figure 3). All podiums during a selected timeframe would either be staffed by Standard TSOs or by Assessor-trained TSOs. Since timeframes, and also potentially passenger volume and mix, will always be different between treatment and control groups, this method was judged by the panel to be less effective at ensuring control equivalence. However, there are a number of other advantages with this method that may be important to consider, including:

1. The method does not require a TSA representative to direct passengers to an appropriate line. As a consequence, allocation procedures cannot be affected by volume fluctuations. A constant randomization rate is maintained regardless of variations in queue load conditions.
2. Experimental and control groups never operate in the vicinity of the other. The method eliminates the need to separate TDC-trained TSO stations from Assessor-trained TSO stations in order to avoid contamination.
3. While the method requires adherence to a pre-programmed random rotation schedule, it is not necessary to track throughput measures separately by Assessor vs. TDC line. Throughput can be tracked more simply at the level of the checkpoint.

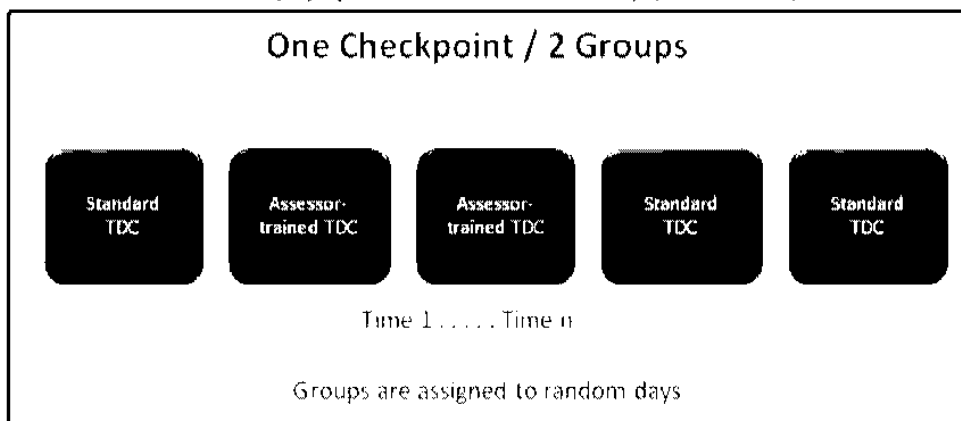
~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

4. Passengers are treated equivalently on any given day. By eliminating differential treatment, the potential for passenger complaints is unlikely to materialize. Consequently, Assessors may be less likely to alter their questions or the length of the interview if queue conditions change.
5. If SPOT operations are required, it can be operated with less complication under this method. Only one protocol outside of SPOT will be operated at a time.
6. The method more easily allows for multiple experimental groups to be included in the study. Implementation of even two groups may be a configuration challenge for many checkpoints under the random allocation design.

If the use of the Assessor Rotation method is required, the panel notes that shorter rotation periods are preferred. For example, day-on / day-off procedures are preferable to week-on / week-off methods.

Figure 3: Graphical Depiction of Passenger Rotation with 2 Groups (Experimental and Control Groups) and 1 Checkpoint



While the rotational design addresses many of the operational challenges inherent in using a method that requires the maintenance of a constant allocation rate, it introduces the potential for uncontrolled (and unmeasured) differences between the groups in terms of time, passenger mix, volume, or other factors. These differences may also create an environment that encourages passengers to select checkpoints (possibly in combination with time/volume effects) to influence detection rates. The movement of the TDC groups between checkpoints (especially in cases where two checkpoints lead into a common sterile area) may encourage passengers to "shop" for less congested screening portals, or in the case of high-risk passengers, for a less invasive screening portal. If rotations are made so that the same group type is placed in both locations on random days, the risk of "checkpoint shopping" can be mitigated. However, staffing constraints may make the implementation of the random rotation method infeasible in cases where multiple checkpoints lead to one sterile area.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need-to-know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Location and Implementation Procedures

There was not a unanimous opinion on the part of panel members regarding the ideal location for this experiment. Several felt the ideal location would be a small or medium-sized airport with few security checkpoints. These members believed that this would allow the maximum degree of control over the experiment and would most likely ensure adherence to protocol. While most members agreed that the advantages presented by this scenario were appealing, other members questioned whether the experiment could achieve the number of referrals and positive cases required from a smaller airport over a 3- to 6-month period. It was anticipated that a higher volume airport might be required.

Despite these differing views, all agreed that **in order to be included in the study, an airport must meet one of two criteria: (1) it must have a checkpoint configuration where a one-to-one relationship exists between the checkpoint and its sterile area, or (2) the study must be implemented at all of the checkpoints leading to a given sterile area (Recommendation #10).** Additionally, there was general agreement that a simultaneous implementation or evaluation of a new security procedure (e.g., TSA Pre-Check) should not be allowed to occur during the evaluation of Assessor. Such parallel efforts may present a challenge to the integrity and interpretability of the Assessor evaluation. The panel was not asked to consider an Assessor evaluation in the context of the newly proposed Risk-Based Security (RBS) approaches; however, to the extent that TSA makes significant changes to the currently used checkpoint process, the proposed design of the Assessor evaluation may need to be reconsidered in light of altered processes, priorities and/or budgets.

With regard to SPOT operations, several members felt that the preferred approach would be to implement the Evaluation Study in an airport where TSOs are *not* trained in SPOT. There was concern that when SPOT is operational, some number of prohibited items or high-risk passengers would be detected through SPOT. As a result, the number of incidents that could be detected by either TDC would be further reduced from an already low base rate. Other panel members felt that a requirement to implement the study in a non-SPOT airport might be impractical in terms of requiring longer data collection periods, and might also limit the generalizability of the results of the study. It was argued that the most important airports in terms of risk and volume are likely to have already implemented SPOT. It was felt that if the Assessor Evaluation Study was implemented in an airport without SPOT, it might not have the volume or number of threats necessary to support the study in a reasonable timeframe. On the other hand, if implemented in a non-SPOT airport, the evaluation would be conducted in a situation where personnel have little experience with behavior detection. This scenario would reduce important confounds likely to be introduced by SPOT.

Several additional data points may be important to this debate. First, only a small number of referrals occurred in the PoC where a passenger was identified as both a SPOT and an Assessor referral (N=4 in Boston and N=1 in DTW). None of the five involved a high-risk outcome. This small number of overlapping cases suggests that SPOT would not remove a significant number of incidents from the

~~Sensitive Security Information (SSI)~~

~~WARNING: THIS RECORD CONTAINS SENSITIVE SECURITY INFORMATION THAT IS CONTROLLED UNDER 49 C.F.R. PARTS 15 AND 1520. NO PART OF THIS RECORD MAY BE DISCLOSED TO PERSONS WITHOUT A "NEED TO KNOW," AS DEFINED IN 49 C.F.R. PART 15 AND 1520, EXCEPT WITH THE WRITTEN PERMISSION OF THE ADMINISTRATOR OF THE TRANSPORTATION SECURITY ADMINISTRATION OR THE SECRETARY OF TRANSPORTATION. UNAUTHORIZED RELEASE MAY RESULT IN CIVIL PENALTIES OR OTHER ACTION. FOR U.S. GOVERNMENT AGENCIES, PUBLIC DISCLOSURE IS GOVERNED BY 5 U.S.C. 552 AND 49 C.F.R. PARTS 15 AND 1520.~~

already low base rate. On the other hand, the small number of cases may be surprising from the point of view that the behavioral indicators used by SPOT and Assessor overlap to some degree. This result may be explained by the fact that Assessors engaged 100% of passengers, while the SPOT team was not always present during the PoC, and even when present, was unable to observe all passengers at all times. If both programs were fully staffed, the overlap may have been higher, making the concern about an already low base rate more pressing.

Second, base rates of illicit activity can only be estimated for airports where SPOT was in operation during FY2010. Recall that this estimate is required for an estimation of the proportion of true positives identified by Assessor¹⁷. It is not clear whether base rates of illicit activity in non-SPOT airports will be representative of SPOT airports. Thus, utilizing base rate data collected from SPOT airports to derive estimates for non-SPOT airports data seems especially problematic. Finally, a secondary goal of the Assessor Evaluation Study is to evaluate the extent to which SPOT and Assessor identify different types of illicit activities. Should this goal's priority be raised, it cannot be achieved without access to officers who have SPOT training. Evaluated as a whole, there may be good reasons to implement the Assessor evaluation in airports that currently operate SPOT.

Compliance Monitoring

In order to ensure the success of the implementation of the experiment, the review panel noted the criticality of making sure that steps are taken to monitor compliance with the Assessor protocol (Recommendation #11). If Assessors are not following the protocol consistently, any lack of an effect could be due to either the Assessor protocol, or to TSOs not adhering to performance expectations. For instance, it could be that TSOs become fatigued very quickly and are unable to sustain the necessary degree of vigilance, or that TSOs develop favored questions that they ask of all passengers and refrain from (b)(3) 49 U.S.C. § 114(r)

The panel recommended a two-pronged approach to compliance monitoring. First, supervisors should periodically observe TSOs to ensure that TSOs are administering TDC and Assessor functions appropriately and effectively. The benefits of this are that it is natural and convenient for supervisors to observe their staff. The drawback is that employees might exert greater effort when they know they are being observed. As such, the sample of performance observed by the supervisor might reflect TSOs' maximum, as opposed to typical, level of performance. One way to overcome this is to also include a sample of performance monitoring by confederates who would proceed through the Assessor queue as though they were travelers. Following their interaction with the TDC officer, they would complete a checklist reflecting the officer's adherence to protocol. Although this would be a more labor and resource intensive approach to monitoring compliance, it would maximize the likelihood that the

¹⁷ The true positive rate can only be estimated for those outcomes that were defined similarly in both studies. For example, "artful concealment" was not coded in the FY2010 Base Rate Study and so this measure could not be evaluated in this context.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

observed performance reflects the TSOs' typical performance. (A related option would be for confederates to proceed to the TDC while wearing a microphone or other audio recording device. This would obviously allow TSA to capture the interaction and code it at later time to ensure the TSO adhered to the Assessor protocol.¹⁸⁾

One of the hypotheses that has been proposed for non-compliance is that TSOs become fatigued or are otherwise unable to maintain the attention required. Some information could likely be provided on this topic if compliance monitors keep track of the time at which contact is made with the TDC. This data point can be compared to the timing of the start of the TSO's shift or to the start of the TDC assignment. Lower compliance rates might be observed for encounters made at later points in a TSO's shift. As a final implementation detail, the panel recommended that groups of passengers traveling together be randomly assigned to the same queue.

Outcome Measures

The outcome measures of interest in the PoC studies were: (1) Serious prohibited and illegal items (SPII), (2) Artful concealments, (3) LEO arrests, and (4) High-risk outcomes representing membership in any of the previous three categories, and (5) LEO Referral. Several panel members noted that the "artful concealment" measure did not previously appear in the SPOT Referral Report Validation Study. It has been presumed (although not confirmed) that the TSA research team is continuing to refine measures in an attempt to address specific criticisms that have been leveled at the outcome measures used in the SPOT Validation Study. For example, "artful concealment"—a new measure—is presumed to show evidence of "intent" to smuggle, a useful theoretical distinction.

In an effort to address previous criticisms, **the panel recommends (Recommendation #12) that each outcome measure proposed should be considered and justified in terms of the following elements:**

1. The outcome should demonstrate a reasonable level of face validity with regard to an **"intent" to smuggle or to deceive**. For example, the detection of an illegal drug has a higher likelihood of being associated with the "intent to smuggle" than say, pepper spray (which may have been simply forgotten).
2. An infraction should **be sufficiently serious**.
 - a. A sufficiently serious infraction will not allow for the return of an item to a passenger after questioning.
 - b. LEO referral data may be useful to consider as an indicator of the seriousness of the suspicion.
3. The outcome should be **attributable to the program** under study. For example:
 - a. A serious prohibited/illegal item should not be credited to the program under study if found via the X-Ray or other screening device. The premise behind this

¹⁸ The suggestion to use a microphone may not be permitted if a potential exists that passengers' conversations could inadvertently be recorded.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

recommendation is that non-behavioral-based security procedures would have detected this item without BDO involvement. It is important to understand the additional benefit gained by employing BDOs, not the overlap. TSA must be focused on understanding effectiveness within the context of both monetary and opportunity costs.

- b. Items or persons detected by employees not trained in behavior detection (e.g., the TDC) should not be attributed to the behavior detection group (a potential problem with checkpoint-level comparisons).
- c. The program under study should not be credited if the passenger surrendered the item prior to screening (although, it may be useful to track these items as a potential measure of *deterrence*).
4. The outcome should **reflect a confirmed finding**, not merely a "suspicion" of an illicit activity. That is, to the extent possible, all recorded data—including BDO remarks—should be used to confirm that the record contained an actual confirmed finding.¹⁹
5. Measurement involving **fraudulent documents should be refined** to distinguish:
 - a. Fraudulent documents used at TDC for authentication (potentially detectable by TDC without Assessor).
 - b. Fraudulent documents discovered during bag search.

The above recommendations were communicated to program analysts during a discussion held at about mid-point in the review process. TSA provided an analysis addressing the topic in a section of the final report (excerpt provided in Table 15 below).

Metrics and Analytical Considerations

The independent review panel evaluated a number of metrics that could be used to analyze the data generated by a future Assessor Evaluation Study, including:

1. **Positive Predictive Value (PPV)** - Of those selected by the Assessor, the proportion that results in high-risk outcomes as compared to those selected by standard security procedures.
2. **Estimated True Positive Rate (TPR)*** - The fraction of high-threat passengers detected from those expected to be detected. Also referred to as the Probability of Detection P(d).]
3. **Estimated False Positive Rate (FPR)*** - The fraction of expected low-threat passengers targeted as high-risk.
4. **Estimated Odds Ratio (OR)*** - The ratio of the odds of an Assessor detecting a high-threat passenger to the odds that a high-threat passenger was detected by the control group.
5. **Referral Rate** - The fraction of the traveling population that was referred to secondary screening.
6. **Probability of Encounter / Engagement**

¹⁹ In the SPOT Validation Study, numerous instances were found where BDOs marked for example, "fraudulent document" as the outcome of the referral (using the structured data fields) even when this designation was refuted by their own comments in the narrative. This situation is similar to a physician who provides an International Classification of Diseases code for a specific diagnosis, when in reality, she was only ruling it out—and it fact, it WAS ruled out.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

- a. $P(\text{Encounter})$ = The fraction of the traveling population receiving a specifically defined baseline level of scrutiny by behavior detection program staff.
- b. $P(\text{Engagement})$ = The fraction of the traveling population more significantly engaged by behavior detection program staff.

The panel noted that each of the metrics marked by an asterisk (*) above requires having access to a method allowing the estimation of expected rates of positives and negatives in the population. This requirement sparked debate among panel members. A number of panel members felt that the Estimated True Positive Rate (TPR), Estimated False Positive Rate (FPR), and the Odds Ratio (OR) should be removed as recommendations for a future Evaluation Study since there was discomfort with making a recommendation to either (1) implement a stratified random compliance measurement program or (2) use the SPOT FY2010 Base Rate Study²⁰ to estimate these metrics. This group felt that the definitional issues inherent in using the Base Rate Study, viewed in combination with the rareness of the event, the non-concurrent data collection periods, and the uncorrected complex sample design made the data too problematic to consider using for this purpose. Members from this contingent instead recommended the use of the positive predictive value (PPV) as a primary metric to evaluate differences between the experimental and control groups in a future Assessor Evaluation. The TPR and OR were only recommended for use by this group in the Red Team Study, implying the need to generate both positives and negatives in the context of the Red Team.

The other group was comfortable with the idea of evaluating the random sample of data collected during the SPOT Referral Report Validation Study, and potentially using it to estimate the other metrics of interest. This group noted that severe problems may attend the use of the PPV alone. It was argued that one could encounter a situation where a comparison of the PPV across two targeting regimes can appear to be fairly large in magnitude and significantly different in a statistical sense (see Table 13), but **neither targeting method performs any better than a simple flip of the coin**. The reason for this is that the PPV is highly dependent upon the prevalence of illicit activity in the population. As an illustration of this relationship, one might consider that as the prevalence of threat decreases, the less likely it becomes that a person exceeding threshold is a threat, and the more likely it becomes that the case will be judged a false positive. The number of threats and non-threats identified in relation to referrals is the primary essence of the PPV metric.

²⁰ The Base Rate Study collected a random sample of 71,589 passengers distributed across 43 airports over the course of a year-long period corresponding closely to FY2010 (12/1/09 - 10/31/10). Of these, 90 passengers were judged to be engaged in some form of serious illicit activity (0.126%, C.I.=0.095%, 0.166%). While the threat rate can be judged to be very low, the sample size is likely to be sufficient for certain aspects of the proposed analyses. However, it is important to note that while 43 airports participated, only 24 collected data throughout the entire 11-month period. Eighteen participated from 7/16/10 - 10/31/10.

~~Sensitive Security Information (SSI)~~

~~WARNING: This document contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this document may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

The PPVs in Table 13 represent actual PPVs for Boston Baseline SPOT (5.41%) and Boston PoC Assessor (10.28%). Using an overly simple example to make the point, one could generate a PPV that is twice the rate of regime 1, but neither is any better than random (note that the estimated true positive rate equals the false positive rate and the OR=1).

Table 13. Comparing PPV Across Two Random Targeting Regimes

Exceeds Threshold	Targeting Regime 1		
	Positive	Negative	Total
Yes	4	70	74
No	20	350	370
Total	24	420	444
Metrics			
TPR	16.67%	PPV	5.41%
FPR	16.67%	RPF	1.00
FPI	18	Refer Rate	16.67%
O.R.	1.00	Specificity	83.33%
Threat Rate	5.41%		

Exceeds Threshold	Targeting Regime 2		
	Positive	Negative	Total
Yes	52	454	506
No	100	870	970
Total	152	1,324	1,476
Metrics			
TPR	34.21%	PPV	10.28%
FPR	34.29%	RPF	1.00
FPI	9	Refer Rate	34.28%
O.R.	1.00	Specificity	65.71%
Threat Rate	10.30%		

Conversely, one could observe a scenario where the PPVs are exactly the same under each regime (see Table 14), but one regime is three times better than random (OR=3.21). In both examples, it is important to understand the role of the base threat rate in generating these outcomes. The important counterargument is that the design of the Assessor Evaluation has been staged to hold constant the underlying base threat rate. **To the extent that TSA deviates from the panel-proposed design in a manner that allows base threat rates to vary between the populations being compared, comparisons utilizing the PPV may be rendered non-interpretable.**

Table 14. Comparing PPV Across a Predictive and Random Targeting Regime

Exceeds Threshold	Targeting Regime 3		
	Positive	Negative	Total
Yes	4	70	74
No	20	350	370
Total	24	420	444
Metrics			
TPR	16.67%	PPV	5.41%
FPR	16.67%	RPF	1.00
FPI	18	Refer Rate	16.67%
O.R.	1.00	Specificity	83.33%
Threat Rate	5.41%		

Exceeds Threshold	Targeting Regime 4		
	Positive	Negative	Total
Yes	60	1,050	1,110
No	160	9,000	9,160
Total	220	9,050	10,270
Metrics			
TPR	27.27%	PPV	5.41%
FPR	10.45%	RPF	2.61
FPI	18	Refer Rate	10.81%
O.R.	3.21	Specificity	89.55%
Threat Rate	2.14%		

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

The second group raised another issue with using the PPV alone that relates to the need to compare targeting regimes to random targeting—a fundamental and standard baseline. Table 15 provides two sets of real data from the Boston Assessor Evaluation Study. The tables detail the results of two different targeting regimes: (1) targeting by SPOT of high-risk passengers in the Boston Baseline Study, and (2) targeting by Assessor of high-risk passengers in the Boston Proof-of-Concept Study.

Table 15. Comparing PPV Across Boston's SPOT and Assessor Study Segments

Exceeds Threshold	HR Passengers - Boston Baseline - SPOT		
	Positive	Negative	Total
Yes	4	70	74
No	747	596,729	597,476
Total	751	596,799	597,550
Metrics			
TPR	0.5325%	PPV	5.41%
FPR	0.0117%	RPF	45.40
FPI	18	Refer Rate	0.0124%
O.R.	45.63	Specificity	99.99%
Threat Rate	0.13%		

Exceeds Threshold	HR Passengers - Boston PoC - Assessor		
	Positive	Negative	Total
Yes	52	454	506
No	401	359,374	359,775
Total	453	359,828	360,281
Metrics			
TPR	11.481%	PPV	10.28%
FPR	0.1262%	RPF	90.99
FPI	9	Refer Rate	0.1404%
O.R.	102.66	Specificity	99.87%
Threat Rate	0.13%		

Classification matrices are standardized using the estimated Base Threat Rate of 0.13% as estimated in the SPOT Validation Study (90/71,589)²¹. If the FPR for SPOT is .0117%, then under a random regime the TPR would also be .0117%. However, the TPR that would be observed in the SPOT baseline population assuming a base threat rate of 0.13% is 0.5325%. Therefore, the RPF is 45.40. This compares to an RPF of 90.99 under a similarly adjusted Assessor targeting regime.

A comparison of the PPV for Assessor compared to SPOT demonstrates that the Assessor targeting regime is 1.9 times better than SPOT. However, if you compare both regimes to random, you see that the Assessor regime is 91 times better than random and the SPOT regime is 46 times better than random. While Assessor can still be interpreted as being about 2 times better than SPOT based on a comparison of the RPF (or OR), by comparing both against a random floor, targeting effectiveness looks very different (very good)²² in light of the extremely low base rates. Even if the random sample should have found double the number of actual positives (only 90 were found), RPFs would still be 23 and 45.

²¹ SPOT's overall base threat rate of 0.13% is utilized in each table, allowing one to compare threat-adjusted metrics. Since the study was conducted at the same airport and at the same checkpoint (although at different points in time), it would not be unreasonable to expect similar base threat rates for each segment of the study. However, it is not clear that SPOT's overall base threat rate is reflective of Boston's threat rate. Without Boston's threat rate, absolute comparisons will be less meaningful than relative comparisons (e.g., comparisons made between the two study segments).

²² Prior stated concerns regarding the interpretation of the Assessor results still stand. The data are being used to illustrate a methodological point.

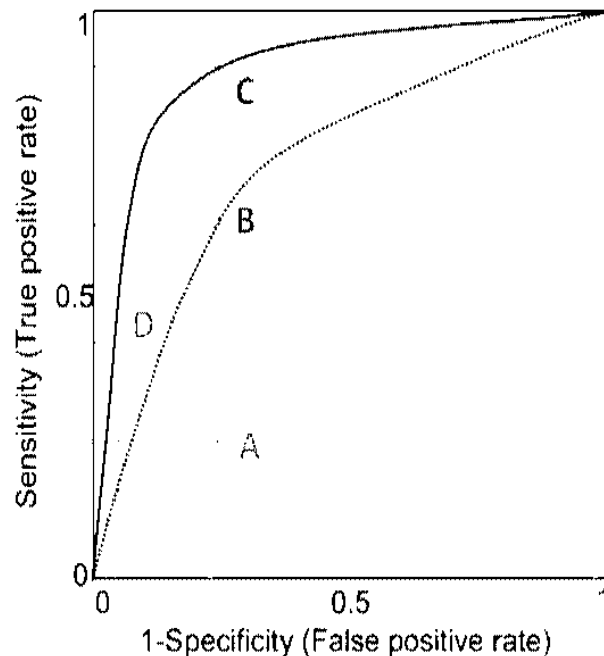
~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

These views of the data are not available with the PPV alone, or without estimates of the threat rate in the population. However, the panel noted that there may be danger in presenting artificially constructed TPRs and FPRs (as presented in Table 15). Some members felt that a reader might interpret the figures inappropriately—and might even go so far as to quote them inappropriately in a public venue. This is a real and valid concern, especially given the political concerns that surrounded the first SPOT Validation Study. However, the alternative concern is that one might also be misled by viewing comparisons without an understanding of the larger targeting context. These interpretations too could manifest inappropriately in a public venue.

In the above example, descriptions of estimated True Positive Rates were always discussed in the context of estimated False Positive Rates. There is little value in reporting the TPR without the attendant FPR. One way to think about this is to consider that the TPR and FPR serve as the two axes of the receiver operating characteristics (ROC) curve. ROC curves plot true positive rates against false positive rates for every threshold value (see Figure 4). The blue line that runs from corner-to-corner at a 45-degree angle represents the results that would be generated by a random targeting process. In this case, the TPR would equal the FPR at every threshold. If you include two targeting regimes on one ROC plot, it is important to hold the FPR constant in order to make the TPR comparison appropriately, and vice versa. In Figure 4, the points on the Y-axis associated with each point A, B, and C can be compared because all occur at one constant FPR. In order to compare the TPR of B to the TPR of D, one must consider that the TPR of D was attained with a FPR that was a fraction of the FPR of B.

Figure 4. Example of a Receiver Operating Characteristic (ROC) Curve



It may also be important to point out that data from Assessor will not allow for the construction of a full ROC curve because only one threshold is used for making referrals; and scores are apparently not generated. However, the performance of one targeting regime can still be compared against another through the use of a relative performance factor [e.g., $RPF = TPR/FPR$]. Relative performance factors are best interpreted when each rate used in the calculation is well-understood. This can be appreciated if you consider that when you capture 8% of the estimated threats by referring 2% of the traveling

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

population, the RPF is 4. If you capture 10% of the estimated threat by referring 5% of the traveling population, the RPF is 2. The first targeting regime is twice as efficient (4 versus 2), yet you actually capture more of the threat under the second targeting regime (8% vs. 10%), possibly because more of the population is referred. So, while a relative performance factor²³ supplies useful information, it will not provide a complete targeting picture. As one evaluates Assessor, even though it is assumed that the base threat rate is constant, because the protocols being evaluated either have only one threshold (Assessor) or are treated as having only one threshold (SPOT) it is important to keep in mind that other factors can vary. For every comparison made, *throughput will be different, referral rates will vary, and measures of targeting effectiveness will vary*; and while the estimated true and false positive rates may not be meaningful in an absolute sense (or might be depending upon one's evaluation of the base rate data), as demonstrated, they can be standardized to make better relative comparisons. Consequently, it will be important to quantify all of these dimensions in order to evaluate the entire context.

One of the issues raised by panel members relates to the fact that positives are not defined in an identical fashion in both the Base Rate Study and the Assessor Study. Table 16 shows that a significant overlap exists between the definitions utilized by the two studies. Approximately 80% of Assessor-identified positives would have also been identified as a positive in the SPOT Validation Study (44 out of 55=.80). Because the overlap is not 100%, one could approach the issue by making one of two assumptions. One might first assume that the definitions were close enough to simply use the SPOT base rate to estimate the number of positives in the population. The alternative approach might choose to revise the Assessor definition to correspond to the SPOT results. If a comparison of results suggests the same conclusion, then confidence in the findings may increase.

Table 16. Potential High-Risk Definitions

Definition of High-Risk Passengers	BOS (N = 506) Freq (Percent)	DTW (N = 183) Freq (Percent)
Current Assessor definition: Expanded SPII list, artful concealment, or LEO arrest	52 (10.28%)	3 (1.64%)
SPOT Validation Study definition: (b)(3) 49 U.S.C. § 114(f)	41 (8.10%)	3 (1.64%)
Proposal 1: Expanded SPII list, artful concealment, LEO arrest or Suspicious Packaging	53 (10.47%)	3 (1.64%)
Proposal 2: Expanded SPII list, artful concealment, LEO arrest or NCIC hit	54 (10.67%)	3 (1.64%)
Proposal 3: Expanded SPII list, local SPII rules, artful concealment, or LEO arrest	54 (10.67%)	3 (1.64%)
Proposal 4: Expanded SPII list, local SPII rules, artful concealment, LEO	57 (11.26%)	3 (1.64%)

²³ Note that the odds ratio is nearly equivalent to the RPF in TSA's targeting context. The RPF is used as an illustrator because it is more intuitively understandable, but the odds ratio is recommended.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

arrest, NCIC hit, or Suspicious Packaging		
Proposal 5: "Smuggling": expanded SPII list or artful concealment	21 (4.15%)	3 (1.64%)

Other assumptions that may be important to examine include variations in the base threat rate, rates of misclassification, and estimates of the probability of encounter for SPOT. There are good reasons to believe that the base rate of illicit activity as measured in the SPOT Validation Study is underestimated. Subsequent to the SPOT Referral Report Validation Study, another DHS S&T team conducted an exercise that manually dredged through all of the Operational SPOT remarks made by BDOs. The outcome of the exercise showed that 60 positives were previously misclassified as negative, and 3 negatives were inappropriately classified as positive. Leaving aside the base rate question for the moment, this resulted in no change to specificity but the O.R. went from 9.27 to 11.26, and PPV went from 1.15% to 1.40%. If it is assumed that the base rate was also underestimated (arbitrarily it was half of what it should have been), PPV and specificity are unchanged, the OR is reduced from 9.27 to 4.63, and TPR goes from .2% to .1%. If it is independently assumed that the SPOT population is 7 times too high (DHS S&T produced another analysis suggesting that a different SPOT population may be more appropriate), the OR goes from 9.27 to 9.51, TPR goes from .2% to 2.94%, and PPV remains unchanged.

Making corrections for all three problems simultaneously causes the OR to decrease from 9.27 to 5.71, the TPR to decrease from .2% to 1.78% and the PPV to decrease from 1.15% to 1.40%. Taken at face value (without appropriate confidence intervals), none of these results altered general conclusions from the study--i.e., SPOT is reliably better than random, but at the observed inspection rate, the vast majority of estimated positives are still missed. This set of results provides an example of using a multitude of metrics under a variety of assumptions to test the robustness of findings.

These scenarios, when considered jointly, bring to light the attendant problems with relying on one metric alone and with taking the FY2010 Base Rate Study at face value. As a result, panel members felt that in order to mitigate some of these risks, it would be important to collect as many measures as possible, and to apply as many analytical methods as were felt to be appropriate and reasonable in an effort to evaluate the results for concurrence in findings. To the extent that the conclusions implied by a particular set of findings are not altered under a range of different assumptions, confidence in the results could be expected to increase. A compromise position was chosen that **recommended that to the extent possible, TSA should utilize a broad array of metrics (at least the PPV, the estimated TPR, the estimated FPR, OR, and Referral Rate) and a range of assumptions (at least consider differences in the base threat rate, rates of misclassification, and definitional differences) to evaluate the performance of the targeting protocol, keeping in mind the limitations presented by the Base Rate data (Recommendation #13).**

Random Selection Processes

The panel does not recommend the implementation of a stratified random selection process at this time. However, it is noted that a permanent stratified random selection process serves several

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

functions and is considered an accepted best practice by many organizations with important targeting programs (e.g., CBP, IRS, etc.).²⁴ Examples of functions served by a stratified random selection process include:

- It allows for the **estimation of the true positive rate**, the false positive rate, the odds ratio and an ability to make comparisons specifically against random targeting.
- It allows for an assessment of **the types of illicit activities missed** by the targeting regime, if any. There is no other way to know this in the TSA context.
- It offers a **potential for making risk adjustments** between groups to allow for better comparisons. For example, if airports vary in terms of underlying threat rates, performance comparisons will be unfair if made without adjustment. This practice addresses the argument that “My airport finds fewer bad guys because there are fewer bad guys to be found.”
- Random data collection is **the gold standard data collection practice for the empirical validation** (or in a best case, optimization) of the weights assigned to targeting indicators.
 - This strategy is less straightforward in a context where proxy data are used—that is, one would not want to optimize one’s targeting methods to target activities that are not the actual activities that should be targeted. However, this is not an argument for declining to examine the data.
 - The validation process goes beyond a focus on the threat base rate; it also requires data collection related to behaviors (the independent variables).
- Random screening also **serves as a deterrent** and allows for the potential of **observing changes in underlying threat rates in response to TSA interventions**.

It may also be worth noting that an operational compliance measurement program can be made more efficient and cost-effective by utilizing a design that varies **selection probabilities according to risk**. Travelers in high-risk strata would be assigned a higher chance of selection, with travelers in lower risk strata receiving a lower probability. Population estimates can then be constructed by reweighting the sample to correct for differential sampling probabilities. Because the higher risk strata are likely to generate higher numbers of high-risk passengers, fewer passengers need to be evaluated (inconvenienced) as would be required by a method that relies on simple random sampling. Consequently, **the panel further recommends that in the absence of an appropriately structured CM program, TSA consider the utility and cost of implementing a risk-stratified probability sample to appropriately estimate baseline threat rates (Recommendation #14).**

²⁴ There is evidence to suggest that TSA engages in some form of random targeting. However, the details of this program (e.g., the methodologies used and outcomes collected) were not available to the review panel. In addition, it is not clear if results from the inspections are maintained electronically and are available for analysis.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a “need to know” as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Statistical Tests and Confidence Interval Estimation

A number of hypothesis tests can be conducted to determine whether or not the observed performance is statistically significant. For example, one can test for differences between Assessor and Non-Assessor in the proportion of secondary referrals that resulted in a positive outcome, or the proportion of secondary referrals subsequently referred to law enforcement. While these two tests are likely to serve as a minimum testing requirement, one might also test the TPR against the expected performance of random targeting that has the same FPR as the Assessor method to determine if there is a significant difference. As described earlier, under random targeting, the TPR is equal to the FPR. Thus, one could also test if the observed TPR proportion is equal to the FPR proportion.²⁵ A different approach might also consider computing the probability that a random targeting practice would achieve the observed TPR or better if it targeted the same number of passengers.

In addition to decision testing, confidence intervals can and should be constructed. For example, the confidence interval surrounding an estimated odds ratio would be important to understand. In addition, by knowing the likely range of the estimate at some level of confidence, one can bound a range of assumptions to be tested. For example, the combined outcome from the original SPOT Referral Report Validation Study has an observed rate of occurrence of 0.126%. A 99% Wilson confidence interval (with Continuity Correction) suggests that the proportion of high-risk passengers based on this outcome is likely to be in the range [0.095%, 0.166%]. Without a confidence interval, one might decide to use an arbitrary range of fixed multipliers to evaluate the sensitivity of metrics to variations in the base threat rate. If the same level of confidence is not represented by each multiplier chosen, then this choice might not yield the results expected. A multiplier of 2 may be sufficient for one outcome, but may be too small for another. This example uses confidence intervals as a way to bound the assumptions being tested. However, as mentioned earlier, confidence intervals can also be used to better report the impact of certain assumptions. Confidence intervals typically include only the impact of sampling error on the range of estimates likely to be observed. A confidence interval could also be constructed to incorporate not only sampling error, but also misclassification error (or other types of error). For example, we previously described the suggestion from Hemenway (1997) for how to evaluate the effect of a percentage of passengers who were randomly misclassified on the chances that Assessor and non-Assessor confidence intervals overlap.

Duration of Study

The independent review panel considered the size of the samples that would be required to ensure that a future evaluation study collects enough data to make inferences about aviation risk in the traveling population. Size requirements were based on a number of assumptions. First, it was assumed that

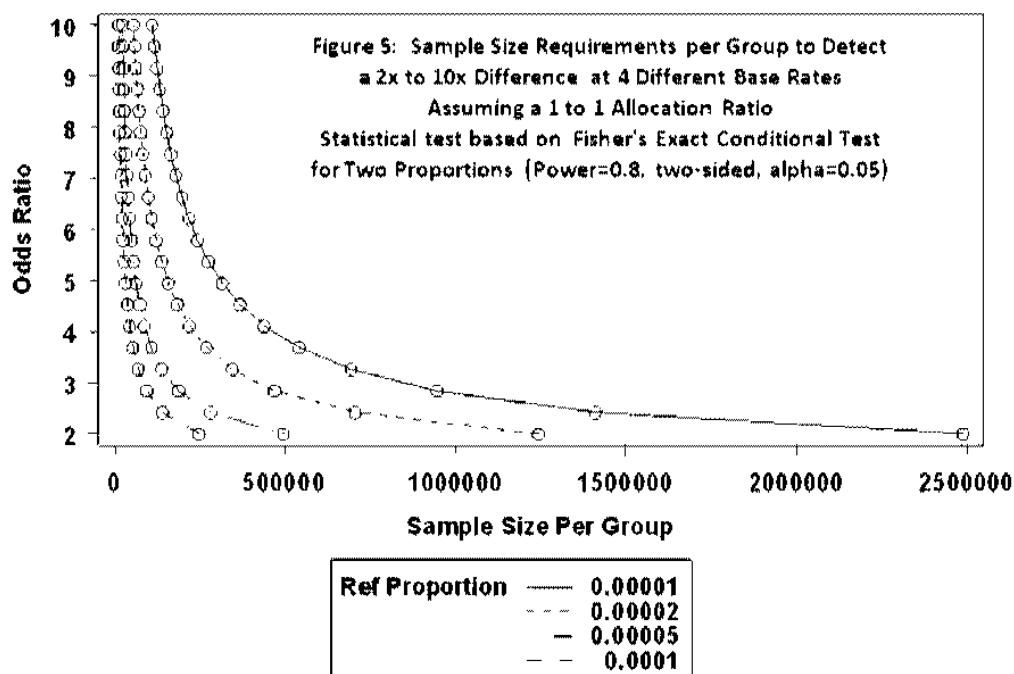
²⁵ Note that since both targeting methods are evaluating the same population, a paired test of correlated proportions should be used (e.g., McNemar's test or Liddell's exact test). The latter is probably favored because of its exact nature.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need-to-know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

100% of passengers would encounter Assessor-trained TSOs. Second, requirements assume an 80% chance of detecting a true difference (power=0.80) with 95% confidence (alpha=0.05, two-sided). Decision testing assumes Fisher's Exact Conditional Test for two proportions at different incidence levels and effect sizes. Using the SPII incidence rates from Boston's Baseline period (.002%) and Proof-of-Concept period (.006%) as a gauge of the incidence rates that are likely to be observed, the analysis evaluated sample size requirements based on incidence rates of 0.001%, 0.002%, 0.005%, and 0.01%. Effect sizes are assumed to be odds ratios and are allowed to vary between 2 and 10 (the SPOT Validation Study observed odds ratios that were typically between 4 and 9).

From Figure 5, if one assumes an incidence of .005% and a 1 to 1 allocation of passengers between the two queues, it would be necessary to process approximately 170,000 passengers through each queue (340,000 passengers total) to detect a statistically significant difference between an odds ratio of 3 and 1 (no effect). In terms of the duration of the study, assuming that Assessor/TSO staffing levels would permit approximately 3000 passengers per day to be processed, the study would need to run for approximately 4 months. To the extent that additional locations can be included, the amount of time required to complete the study will decrease.



~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

One may arrive at a different set of conclusions about sample size requirements if using referrals as the denominator. Table 17 provides a number of parameters estimated from the Boston and Detroit Proofs-of-Concept. On average, Boston screened approximately 700 passengers for every referral made; Detroit screened nearly 1800 passengers per referral. One out of every 10 Assessor referrals in Boston resulted in detection of a high-risk passenger. In Detroit, one out of every 50 referrals resulted in detection of a high-risk passenger. If we assume (for the sake of argument) that the non-Assessor trained TSO detection rates can be expected to operate at the 5% Boston SPOT rate and Assessor rates operate at 2 times the base rate (Boston Assessor rate is 10%), one is likely to have an 83% chance of detecting the effect (with 95% confidence) with a sample size of approximately 500 referrals per group (N=1000).

Table 17: Characteristics of the Boston and Detroit Proofs-of-Concept

Metric	Boston	Detroit
Begin Date	8/15/2011	10/17/2011
End Date	12/15/2011	12/15/2011
Pilot Days	120	58
Throughput	360,281	323,383
Referrals	506	183
Number of Passengers Screened per Referral	712	1,767
Number of High-Risk Passengers Detected	52	3
Percent of HRP (per Referrals)	10.3%	1.6%
Avg. Number of Passengers / Day	3,002	5,575
Avg. Number of Referrals / Day	4.2	3.2
Referrals Required (Power=.80)	500	500+
Pilot Days Required	119	156

If one can assume that referral rates will remain similar to the referral rates seen above²⁶, Boston would need 119 days of Assessor referrals (357,000 passengers) and Detroit would need 156 days of Assessor referrals (870,000 passengers) to generate the referral numbers required for the Assessor group alone (N=500). Assuming an equivalent number would be required for the non-Assessor trained TSO group would result in a requirement of 715,000 and 1.7 million for Boston and Detroit, respectively. Results from the PoC studies have suggested that it is unlikely that the non-Assessor trained TSO will maintain referral rates as high as the referral rates of Assessor-trained TSOs. Consequently, these sample size requirements are likely too low. The magnitude of the difference in the estimates produced under the first power analysis compared to the second is surprisingly large. This investigation demonstrates how

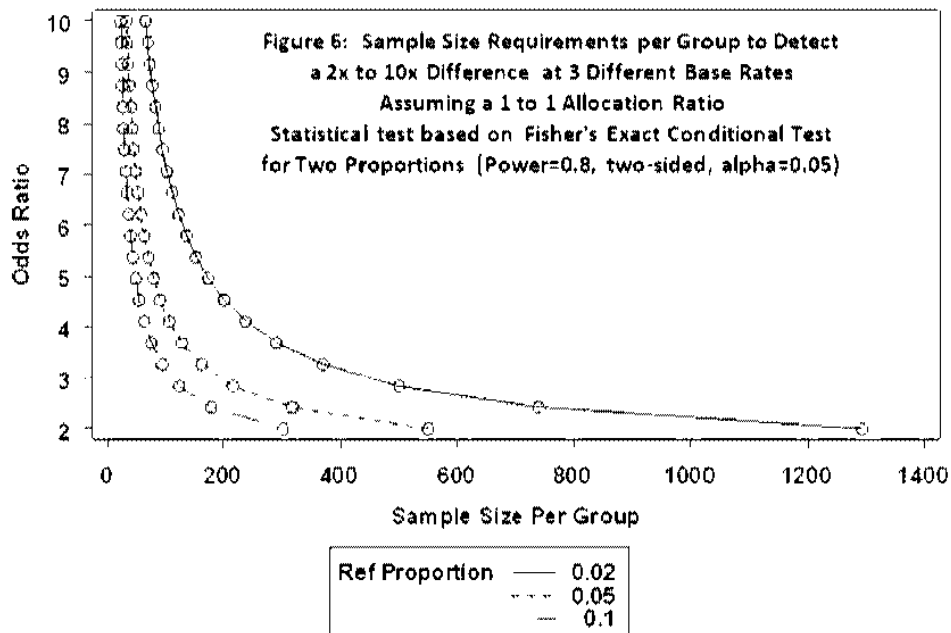
²⁶ Note that the panel stressed the importance of making certain that secondary screening referrals only reflect reasonably suspicious passengers. Referral rates should not be altered to increase sample sizes (**Recommendation #15**).

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

small differences in study parameter assumptions can create very large differences at the population level.

From Figure 6, the sample size required per group (x-axis) to detect a given odds ratio (y-axis) is plotted for three different reference proportions. Sample size requirements decline as base rate proportions move closer to 0.5. Note that the availability of staff has not been factored into estimates of the number of study days required. All of these estimates are tenuous at best; sample size recommendations should be interpreted as "ballpark" estimates.



Benefits and Limitations of Study 1

Although the Proof-of-Concept studies have provided valuable information and have yielded numerous lessons about the implementation of Assessor, the presence of confounding variables has made it difficult to isolate the effectiveness (or lack thereof) of Assessor. A new experimental design for an Operational Study of Assessor has been proposed to more effectively control confounding influences—the panel's primary concern. Additional benefits of the new design include an ability to evaluate the protocol in an operational context that allows for the observation of real (not contrived) passenger

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

behavior, and an increased potential for a more valid comparison of SPOT results to Assessor results.²⁷ The new design also supports a cross-validation of the BDO selection instruments as applied to the Assessor position.

The new design will not overcome all known limitations. Sample size requirements—driven by low base rate events—will need to be attended to through the implementation of a large-scale study (i.e., a long implementation period or a multi-airport study). In addition, the most effective design calls for the concurrent estimation of base rates of illicit activity, and requires that a random selection of passengers be referred to secondary for more intensive screening. While the design will inconvenience a large number of innocent travelers, in the absence of this design feature, key effectiveness measures cannot be estimated. At the end of the day, though, even if all other recommendations could be carried out perfectly, the risk remains that known serious data quality problems will introduce more error into the study than can be reasonably tolerated. This limitation is not unique to the proposed new design; but, rather becomes increasingly important in studies of low base rate phenomena and consequently, must be addressed. Key changes to TSA's information technology (IT) infrastructure and training programs will be required.

6.2 STUDY 2—A FIELD EXPERIMENT UTILIZING RED TEAMS

An ongoing challenge is the fact that the outcome of interest is a low base rate event and is generally unknown (i.e., undetected events are never known). Although it is possible to identify the rate at which “flagged” individuals are discovered to be high risk or in possession of SPII, it remains unknown how many “non-flagged” individuals might have been high risk or in possession of SPII. Two approaches are usually taken to overcome this limitation. The first—random sampling—has been discussed in the context of the first study. The second—a covert or “red team” approach—is recommended by the review panel to serve as a supplementary study. A field experiment utilizing red teams will serve to address the rare event problem, and also allow for a better understanding of the strengths and weaknesses of the security measures currently in operation (Recommendation #16).

Purpose of Red Teaming

Red teaming is a relatively generic term that encompasses a broad range of activities that often depend on context and operational needs. At a high level, red teams are typically established to test or challenge a system or plan, as viewed through the eyes of an adversary. Red teaming provides an operational team an opportunity to better understand the environment, to identify vulnerabilities, and to reveal how adversaries might adapt or create counter measures to evade detection. The practice allows one to generate hypothetical adversarial behavior for analysis, and to challenge deliberately assumptions held by the operational team. These activities as a whole have the potential to reduce risk

²⁷ As noted previously, in order to lessen the risk of additional confounders entering a new study, the independent panel prefers that SPOT not be operated concurrently with a future Assessor evaluation.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a “need to know” as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

through the avoidance of some of the pitfalls associated with "groupthink," a style of thinking that may be employed when individuals are deeply involved in a cohesive group (Janis, 1972).

TSA has a history of engaging in "red teaming" exercises (e.g., Aviation Screening Assessment Program (ASAP)). "Artful concealment" has been an important component of threat insertion in this context. Examples of threats that have been simulated by ASAP include:

(b)(3):49 U.S.C. § 114(r)

The intent of the proposed red teaming effort is not to duplicate ASAP efforts. Rather, the goal would be to design a supplementary study capable of modeling behavioral and document components that could be detected by Assessor. The Assessor protocol has been designed to achieve the related goals of deterring and detecting threats to a flight, and assuring that all items making it onto the plane are safe. To realize these goals, Assessors are instructed to focus on three sets of signature events: (1) appearance and behavior signs, (2) documentation signs, and (3) questioning strategies that allow for the detection of deception and other inconsistencies (ISDS Aviation Security Program, 2011). Of these three types of events, documentation may be the most straight-forward to implement. All panel members agreed that evaluating the extent to which fraudulent documents and/or documents exhibiting various threat signatures can be detected by Assessors would be valuable to assess. Behavioral components were judged to be more challenging. Citing research that has investigated the extent to which an increase in cognitive load is associated with deception and whether reliable indicators exist that would allow its detection (Vrij, Granhag, Mann, & Leal, 2011; Vrij, Granhag, & Porter, 2010), panel members generally agreed that attention should be paid to appearance and behavior signs, and an attempt should be made to detect items that would not be detected by the TDC or the WTMD/AIT equipment.

Panel members suggested that the design of the red team study should introduce some type of smuggling, whether it be drugs, knives, or documents. The term "smuggling" implies that there would be some attempt at concealment. There was also discussion of not only sending through persons who were instructed to smuggle something, but also sending through "innocents" to understand the extent to which confederates with nothing to hide might be detected. It was imagined that confederate passengers who were not engaged in concealment may still need to lie as a by-product of the red team study—for example, most will not actually fly to the destination indicated on the boarding pass, and if asked, a story might need to be constructed. In this regard, they may not actually be "true negatives." Such issues would need to be addressed, and potential effects evaluated.

(b)(3):49 U.S.C. § 114(r)

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

(b)(3):49 U.S.C. § 114(f)

(b)(3):49 U.S.C. § 114(f)

The analogy to this in the Assessor context is that if Assessors are using the Assessor protocol appropriately, and if test items exhibit the appropriate signature or the signature can be elicited by the Assessor, the test should be a fair one. (b)

(b)(3):49 U.S.C. § 114(f)

(b)(3):49 U.S.C. § 114(f)

Since the Assessor protocol looks for signals that are outside of the SPOT protocol, it is possible that Assessor will ultimately demonstrate a higher rate of detection than SPOT. However, at the present time, the existence and potential size of this effect is unknown. Considered as a whole, these facts call into question the idea of equating Assessor-based red teaming to ASAP red teaming. (b)(3):49 U.S.C. § 114(f)

(b)(3):49 U.S.C. § 114(f)

To make the Assessor red team analogous to the ASAP red team, changes would need to be incorporated into the test. Signs present in the documentation offered to the TDC should be objectively detectable. However, other items are not likely to be detectable by the TDC or by the WTMD/AIT equipment. Fraudulent documents stashed in the carry-on, for example, would require that other signs (appearance, behavior, deception, and inconsistencies) be available for detection. In order for a red team exercise to be effective at detecting non-documentation-based threats, it would be necessary for confederates to engage in some type of lie or to respond with some form of inconsistency in order to establish a signature that could be detected. Ultimately, panel members preferred the idea of attempting to generate situations that would be likely to spawn authentic reactions on the part of confederate passengers. However, this desire would need to be balanced against the need to generate detectable signatures.

Confederate Selection

Panel members evaluated two approaches to the recruitment of confederates. The first method sought to recruit airport passengers into the study, asking the subject to conceal an item through security. It was envisioned that passengers would be randomly assigned a concealment status. The argument for using passengers is that they would be untrained and inexperienced and would have certain heightened anxieties while participating. In addition, such passengers were thought to be more representative of the target population. However, several panel members imagined that many passengers would almost certainly refuse to participate based on not being able to allay concerns that recruiters were illegitimate. Ultimately, this issue drove the panel to abandon the idea of recruiting passengers and move toward **recommending a design that recruited confederates who are unaffiliated with TSA and the airport (Recommendation #17)**. Again, panel members noted the importance of attempting to build realism into the experiment. It was argued that by randomly assigning passengers to different roles, the strategy may serve to increase natural anxieties. Participants would also be incentivized to successfully complete the screening process without detection.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

The panel discussed the extent to which red teamers could truly be expected to act the part. There was concern that it might not be reasonable to expect red team members to be able to simulate terrorist behavior or to portray specific emotions. Responses to this question focused on the extensive array of research in social psychology and sociology that documents the precedent for the use of confederates in studies where information is concealed or withheld from others. In addition, it was pointed out that even in the case of fraudulent document detection, red teamers, by definition, must lie about their trip stories, and respond with lies when answering questions. These lies are real. Thus, there may be logic to support the idea that an authentic signature would be present and detectable. On the other hand, past research has found that there are individual differences in human abilities to conceal emotions, with individuals high in self-monitoring (Snyder, 1974) better able to conceal their emotions than individuals low in self-monitoring (Friedman & Miller-Herringer, 1991). Since it is vital that reliable signatures be available for the test, panel members stressed the importance of suitable confederate recruitment and careful development of the study protocol.

Study Protocol

A full study protocol was not developed by the review panel at this time. Most panel members felt that the experimental design for Study 2 was sufficiently complex that the group would need additional time to work through the many issues that were likely to arise. Instead, panel members decided to describe as much of the discussion as possible in the review panel report, while **recommending that a group of subject matter experts be convened to continue working on the red team study design in the near future (Recommendation #18).**

In terms of the red team protocol, it was generally envisioned that confederates would be recruited from outside of TSA, and randomly assigned to one of a couple of different groups. Some would be assigned to a control group (would be asked to carry nothing), some would be assigned to a group that was instructed to smuggle items, and others would be assigned to a group that would carry problem documents. The panel anticipated that the manipulation of the documents would be relatively easy to accomplish. The manipulation of behaviors was felt to be the larger challenge, but importance was placed on resisting the urge to train actors to behave in certain ways. Additionally, it was noted that there might be some value in assigning actors to carry problem documents, but also leaving them unaware that the documents were problematic. Members of this group would be told that they were assigned to the "innocent" group.

As with Field Experiment 1, panel members recommended that research confederates be randomly assigned to queue type.²⁸ Under this design, the effectiveness of Assessor is determined by comparing the rate at which the Assessor correctly identifies confederates to the rate at which the standard TDC correctly identifies them. The panel expects that the standard TDC will be likely to detect some of the

²⁸ Since the experimental and control queues must be operating under the proposed design, Field Experiment 2 would be most efficiently administered if it was conducted concurrently with Field Experiment 1.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

threats seeded into documents, so smaller differences might be observed between the TDC and Assessor groups than would be expected with other indicators. The Assessor group is expected to detect higher rates of concealed items, behavioral signs, and inconsistencies than the TDC group.

Location and Implementation Procedures

The red team approach as currently conceived does not require a modification to the layout of Assessor and non-Assessor queues. The only important factor is that research confederates be randomly assigned to an Assessor queue or non-Assessor queue. For this reason, the Study 2 protocol could be implemented in a setting where Assessor is already in place, such as Boston or Detroit. This would prevent the need to train an entire airport location in Assessor just for the purpose of the second study. Alternatively, the panel suggested that the red team study could be operated concurrently with the operational study; however, a small pilot in one airport was suggested as a starting point.

As with Field Experiment 1, in order to ensure the integrity of the experimental manipulation (i.e., the implementation of Assessor) it is critical that steps be taken to monitor compliance with the Assessor protocol. If Assessors are not following the protocol in a consistent manner, then the apparent lack of an effect could be due to Assessor not being effective, or to TSOs not adhering to performance expectations. The same approach taken to monitor compliance in Study 1 is also recommended for Study 2. As described above, supervisors and/or confederates should periodically observe TSOs to ensure TSOs are administering Assessor appropriately and effectively. If confederates are used to monitor compliance with the Assessor protocol, these individuals should not also serve as red team research confederates. In fact, it may be important to limit the number of observers (compliance monitors or otherwise) altogether. (b)(3):49 U.S.C. § 114(r)

(b)(3):49 U.S.C. § 114(r)

(b)(3):49 U.S.C. § 114(r)

Other issues such as pilot program signage and staffing issues were not discussed, but are likely to be important to consider.

Duration of Study

In terms of the duration of the experiment, the panel anticipated that it would take place over the course of approximately two months. Theoretically, the study only needs to last long enough to send a sufficient number of confederates (approximately 60, assuming 2 groups) through the queues. However, if all red team members were sent through during a one-week period, and a significant number were detected, it would be apparent to TSOs that an unusual number of passengers with false documents (for example) were attempting to traverse security. As such, the red team members would need to be sufficiently spaced out so as not to make the observation obvious to the TSOs (and BDOs if SPOT is in operation). The panel discussed sending one confederate through every three to four days. However, if there are existing TSA guidelines on red teaming, panel members suggested deferring to

²⁰ Personal conversation with a BDO.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

such guidelines for the frequency and procedural details. If 60 red team attempts are made, and they are spaced out by a few days between each attempt, the study could take at least six months to carry out in one airport. If several airports are used, then time to completion can be reduced, but the scale of the study then becomes larger. The panel felt that a smaller pilot study would be preferable as a first step. Ultimately, however, the costs and benefits of various study choices would need to be weighed.

Statistical Analysis

In terms of the statistical analysis, the approach to analyzing Study 2 should look fairly similar to Study 1. Specifically, it is anticipated that the data would be reported in a contingency table (2 x 2 table) format and a Fisher-Irwin Exact Test of the association between queue type (Assessor or non-Assessor) and disposition (Red Team Member Flagged versus Not Flagged) would be conducted. **Because the panel recommended that the experiment be seeded with both threats and non-threats, a full 4 x 4 classification matrix can be used in the assessment (Recommendation #19).**

It may be useful to note that Fisher's exact test assumes that the row and column totals are fixed. In most experiments, row and column totals are free to vary. Red teaming exercises require that the marginals of the classification matrix remain fixed, creating a situation where sampling error simply does not apply; all of the population parameters are understood *a priori*. Consequently, it was not completely clear that a statistical approach would be warranted. However, since one will still be interested in understanding the extent to which the patterns generated could have been generated by chance, the context may be nearly ideal for the application of Fisher-Irwin's Exact test. In practice, though, a larger number of cases may make the chi-square results indistinguishable from the exact test, negating the need for the exact test. To the extent that additional groups are examined or other comparisons made, additional statistical tests or analytic approaches may need to be considered.

Benefits and Limitations of Study 2

All prior work (the SPOT Validation Study and the Boston and Detroit Proofs-of-Concept) has been limited by the extremely low base rate of illicit activity. The primary purpose of the red team study is to address the low base rate issue through an alternative method of analysis. Using this approach, evaluators can control the base rate of illicit activity allowing a precise calculation of true and false positive rates. It should be noted, however, that while precise true and false positive rates can be now be calculated, these data cannot establish validity. By serving as an alternative method of analysis, important insights can nevertheless be gained. In this regard, Study 2 may identify previously unknown strengths and weaknesses of the Assessor protocol, as well as highlight instances of unanticipated effects or other bias. To the extent that different threat signatures with overlapping indicators can be inserted into the study, the study may also be able to address the reliability of specific indicators. At a high level, the study may indeed support an assessment of the Assessor protocol from an adversary's point of view, and lead to the development of a more robust security protocol.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Despite its many advantages, the proposed red teaming exercise should be viewed as exploratory—perhaps even as a proof-of-concept—and as a complement to the Operational analysis proposed as Study 1. The red teaming study evaluates the Assessor protocol in the context of experimentally manipulated behaviors. It is not clear how well the results of the study can be generalized from the unique experimental setting to the required operational setting. There are many current unknowns associated with Study 2; consequently, it will be important to take the time to thoroughly vet the protocol finally chosen. The quality of the information obtained from Study 2 will be highly dependent on the quality and experience of the team, and the approach taken.

6.3 A SUMMARY OF REVIEW PANEL RECOMMENDATIONS

The Independent Review Panel made nineteen recommendations with regard to two new field experiments. Recommendations primarily covered topics relating to the design of the study and the analysis of data produced by the study. Recommendations are summarized below.

Study 1 Recommendations

1. The Independent Review Panel recommends that interpretations of the effectiveness data collected in the Boston and Detroit Proofs-of-Concept be made very cautiously (observed differences might not be due to Assessor), and that future designs attempt to resolve confounding factors.
2. The review panel recommends a study design that compares targeting results for two groups: (1) Standard TSOs serving as the travel document checker (TDC) (Control Group), and (2) Standard TSOs trained in Assessor serving as the TDC (Experimental Group).
3. The review panel recommends that TSOs be randomly assigned to each group (Experimental and Control groups).
4. It is recommended that the requirement for Assessor certification be eliminated in future studies. It would be valuable to still conduct the Assessor skill assessment, but the scores should simply be recorded and made available for analysis, not used for selection. Compliance monitors should remain blind as to Assessor test results.
5. The panel recommends that additional skill assessments be provided to both groups of TSOs (Assessor-trained and non-trained) subsequent to—and possibly prior to—training for later analyses of performance.
6. The review panel recommends that (1) all data collection instruments be tailored to each officer's role in the experiment, and (2) all officers undergo in-depth training on data collection

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

and submission procedures. It is especially important to make sure that officers understand how to document outcome measures through the use of structured database fields.

7. In order to further reduce the risk of study confounds, the panel would prefer that SPOT not operate during the same time frame that an Assessor evaluation is running.
8. The review panel recommends that passengers be randomly assigned to treatment and control queues.
9. To the extent that staffing can be increased to account for the lengthier cycle times associated with the Assessor group, panel members recommend random allocation as the method to be used for ensuring passenger randomization. A gate-keeper will need to be employed with the random allocation method to ensure the protocol is administered appropriately. However, it will also be important to have a third party monitor randomization rates to ensure that they remain constant and that they produce expected passenger counts and distributions.
10. The panel recommends that to be included in the study, an airport must meet one of two criteria: (1) it must have a checkpoint configuration where a one-to-one relationship exists between the checkpoint and its sterile area, or (2) the study must be implemented at all of the checkpoints leading to a given sterile area.
11. To ensure the integrity of the implementation of the experiment, the review panel noted the criticality of making sure that steps are taken to monitor compliance with the Assessor protocol.
12. The panel recommends that each outcome measure proposed should be considered and justified in terms of the following criteria: (1) the measure should demonstrate a reasonable "intent" to smuggle or deceive; (2) the infraction should be judged sufficiently serious, (3) the outcome should be attributable to the program under study, (4) the outcome should reflect a confirmed finding, and (5) the outcome should reflect multiple refined definitions of fraudulent documents.
13. The panel recommends that to the extent possible, TSA should utilize a broad array of metrics (at least the PPV, the estimated TPR, the estimated FPR, OR, and Referral Rate) and a range of assumptions (at least consider differences in the base threat rate, rates of misclassification, and definitional differences) to evaluate the performance of the targeting protocol, keeping in mind the limitations presented by the Base Rate data.
14. The panel recommends that in the absence of an appropriately structured CM program, TSA consider the utility and cost of implementing a risk-stratified probability sample to appropriately estimate baseline threat rates.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

15. The panel stressed the importance of making certain that secondary screening referrals only reflect reasonably suspicious passengers. Referral rates should not be altered to increase sample sizes.

Study 2 Recommendations

16. A covert or "red team" approach—wherein confederates attempt to pass through security undetected—is recommended by the review panel as a supplementary study.
17. The panel recommended the use of actors not affiliated with TSA or the airport in the red team study.
18. The panel acknowledged the complexity of designing the red team study appropriately, and recommended that a group of subject matter experts be convened to continue working on the red team study design in the near future.
19. The panel recommended that the red team experiment be seeded with both threats and non-threats to allow for analysis of the full 4 x 4 classification matrix.

7 References

- Fleiss, J., Levin, B., & Paik, M. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, New Jersey: Wiley-Interscience.
- Friedman, H., & Miller-Herringer, T. (1991). Nonverbal display of emotion in public and in private: Self-monitoring, personality, and expressive cues. *Journal of Personality and Social Psychology*, 61, 766–775.
- George, P., & Hole, G. (1995). Factors influencing the accuracy of age estimates of unfamiliar faces. *Perception*, 24(9), 1059–73.
- Hemenway, D. (1997). The myth of millions of annual self-defense gun uses: A case study of survey overestimates of rare events. *Chance*, 10, 6–10.
- ISDS Aviation Security Program. (2011). Assessor Course Workbook.
- Janis, I. L. (1972). *Victims of Groupthink: a Psychological Study of Foreign-Policy Decisions and Fiascoes*. Boston, MA: Houghton Mifflin.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

- Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of personality and social psychology*, 30(4), 526.
- Vrij, A., Granhag, P., Mann, S., & Leal, S. (2011). Outsmarting the Liars: Toward a Cognitive Lie Detection Approach. *Current Directions in Psychological Science*, 20(1), 28.
- Vrij, A., Granhag, P., & Porter, S. (2010). Pitfalls and Opportunities in Nonverbal and Verbal Lie Detection. *Psychological Science in the Public Interest*, 11(3), 89.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

APPENDIX A: ACRONYMS AND ABBREVIATIONS

Table 10—List of Acronyms and Abbreviations used in the Report

Acronym	Definition
AIT	Advanced Imaging Technology
ANOVA	Analysis of Variance
APA	Aviation Passenger Assessor
BDAD	Behavior Detection and Analysis Division
BDO	Behavior Detection Officer
BOS	Boston Logan International Airport
CONOPS	Concept of Operations
CBP	Customs and Border Patrol
DCL	Data Collection Lead
DHS S&T	Department of Homeland Security Science and Technology
DTW	Detroit Metropolitan Wayne County Airport
ETD	Explosives Trace Detection
FAM	Federal Air Marshal
FTD	Fraudulent Travel Document
HRP	High-Risk Passengers
ICAO	International Civil Aviation Organization
ID	Identification
ISDS	International Security Defense Systems
JFK	John F. Kennedy International Airport
LEO	Law Enforcement Officer
NCIC	National Crime Information Center
NDF	National Deployment Force
OJT	On-the-Job Training
PASS	Passenger Assessment Screening System
PMIS	Performance Measurement Information System
PoC	Proof-of-Concept
QUAD-S	Secondary Security Screening Selection
RBS	Risk-Based Security
SME	Subject Matter Expert

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need-to-know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Acronym	Reference Term
SOP	Standard Operating Procedures
SPOT	Screening of Passengers by Observation Techniques
SPII	Serious Prohibited or Illegal Items
SSSS	Secondary Security Screening Selection
STSO	Supervisory Transportation Security Officer
TDC	Travel Document Checker
TSA	Transportation Security Administration
TSI	Transportation Security Inspector
TSM	Transportation Security Manager
TSO	Transportation Security Officer
WTMD	Walk Through Metal Detector

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

APPENDIX B: POTENTIAL AIRPORTS FOR A FUTURE STUDY

Airport throughput data were available for the 43 airports represented in the FY2010 Base Rate Study. These data were used to examine an airport's potential for inclusion in a future Assessor evaluation study. Data were collected from the 43 airports between December 1, 2009 and October 31, 2010. The analysis assumed a passenger sample size of 800,000 and a 3 to 6 month study completion constraint.

Table 12—Daily Passenger Throughput and Number of Checkpoints for 18 Airports Meeting Volume and Data Consistency Requirements
[Numbers in brackets express the 95% confidence interval]

Airport Name	Airport Code	Number Of Check Points	Daily Passenger Throughput (Sample Size)	Days To Achieve 800,000
Baltimore-Washington International	BWI	5	26,508 [25,891, 27,125]	31 [30, 31]
Logan International	BOS	11	43,378 [42,614, 44,141]	19 [19, 19]
Burbank-Glendale Pasadena	BUR	2	6,780 [6,662, 6,897]	119 [116, 121]
Cincinnati/Northern Kentucky International	CVG	3	7,312 [7,187, 7,438]	110 [108, 112]
Denver International	DEN	3	38,559 [37,892, 39,226]	21 [21, 22]
Detroit Metro Wayne County	DTW	7	25,249 [24,847, 25,651]	32 [32, 33]
Honolulu International	HNL	6	21,119 [20,875, 21,363]	38 [38, 39]
Houston Hobby	HOU	1	9,478 [9,281, 9,675]	85 [83, 87]
Memphis International	MEM	4	6,259 [6,148, 6,370]	128 [126, 131]
Newark International	EWR	9	43,848 [43,184, 44,513]	19 [18, 19]
Metropolitan Oakland International	OAK	2	14,395 [14,114, 14,676]	56 [55, 57]
Orlando International	MCO	4	49,866 [49,179, 50,553]	17 [16, 17]
Philadelphia International	PHL	7	32,063 [31,417, 32,708]	25 [25, 26]
Phoenix Sky Harbor International	PHX	7	37,349 [36,746, 37,952]	22 [22, 22]
Portland International	PDX	3	16,650 [16,328, 16,972]	49 [48, 49]
San Antonio International	SAT	2	10,941 [10,749, 11,132]	74 [72, 75]
San Diego Intl-Lindbergh Field	SAN	6	23,875 [23,471, 24,279]	34 [33, 35]
Tampa International	TPA	4	24,543 [24,144, 24,942]	33 [33, 34]

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

While 25 of the 43 *airports* reported passenger throughput for at least 300 of the 335 available days, 20 of the 25 consistently reported passenger throughput for at least 300 of the 335 available days at each *checkpoint* as well. Of the 20 airports with consistent checkpoint data, 18 have enough passenger throughput between the core hours of operation (5am to 9pm) to achieve the 800,000 passenger mark within a 6 month timeframe. The airports that met both the volume and data consistency requirements are listed in Table 11 above. It is important to note, however, that to be eligible for inclusion, airports must also have at least one available checkpoint where a one-to-one relationship exists between the checkpoint and its sterile area—or enough staff need to be available to run the study simultaneously at all checkpoints feeding the common sterile area. Hence, this list may be reduced upon further review. Additional checkpoint information is provided in Table 12.

Table 12—Checkpoint Information on the 18 Potential Airports

Airport Name	Airport Code	Checkpoint
Baltimore-Washington International	BWI	Pier A Pier B Pier C Pier D Pier E
Logan International	BOS	A1 B1 B2 B5 B6 B7 Passenger (Air Canada) C1 C2 C3 E1 E2
Burbank-Glendale-Pasadena	BUR	BUR01 BUR02
Cincinnati/Northern Kentucky International	CVG	FIS T2-MCP T3-MCP
Denver International	DEN	A Bridge North

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Airport Name	Airport Code	Checkpoint
		South
Detroit Metro Wayne County	DTW	Blue-1 Blue-2 Red 1 Red 2 Red 3 Red 5 Red 6
Honolulu International	HNL	Checkpoint 1 COMMUTER TERMINAL CPT 2 HNL03 HNL04 HNL05
Houston Hobby	HOU	CENTRAL
Memphis International	MEM	Int'l Arrivals MEM-A MEM-B MEM-C
Newark International	EWR	Terminal A/1 Terminal A/2 Terminal A/3 Terminal B/1 Terminal B/2 Terminal B/3 Terminal C/1 Terminal C/2 Terminal C/3
Metropolitan Oakland International	OAK	TERM1 TERM2
Orlando International	MCO	EAST Checkpoint EAST FIS WEST Checkpoint WEST FIS
Philadelphia International	PHL	A East A West Sec 5 A West Sec 7 Checkpoint B Checkpoint C Checkpoint D/E Checkpoint F
Phoenix Sky Harbor International	PHX	T-2 T-3 North

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Airport Name	Airport Code	Checkpoint
Portland International	PDX	T-3 South
		T-4 A
		T-4 B
		T-4 C
		T-4 D
		PDX A/B/C
		PDX D/E
San Antonio International	SAT	PDXFIS
		SAT01
		SAT02
San Diego Intl-Lindbergh Field	SAN	Checkpoint 1
		Checkpoint 2
		Checkpoint 3
		Checkpoint 5
		Checkpoint 6
		Commuter
Tampa International	TPA	TPA A
		TPA C
		TPA E
		TPA F

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15, and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

APPENDIX C: INDEPENDENT REVIEW PANEL

MEMBERS

Dr. Kevin Bradley: Dr. Kevin Bradley is a Senior Scientist at the Human Resources Research Organization (HumRRO). His expertise is in the area of test development and validation. His primary areas of research and experience are in the development, administration, and maintenance of job knowledge tests; the implementation of job analysis and competency modeling projects; and the validation of personality tests and structured interviews for making employee selection and promotion decisions. In addition, he has extensive experience managing complex databases and utilizing various statistical analysis techniques. He has conducted and presented research in job analysis and the use of structured interviews and personality inventories in employment decision-making, and the impact of faking on the psychometric properties of personality inventories. He earned his master's degree from Rensselaer Polytechnic Institute and his Ph.D. from Virginia Tech.

Dr. Karen Carver: Dr. Karen Carver is an independent statistical consultant and systems analyst with over 20 years of experience supporting government and commercial clients with statistical requirements on high-performance, mission-critical systems. Dr. Carver is a demographer by training, completing an NICHD-funded post-doctoral research fellowship at the Carolina Population Center (CPC) of the University of North Carolina (UNC) Chapel Hill. She has held several academic positions, including positions at the Washington State University and the Population Research Institute (PRI) at the Pennsylvania State University. Much of her work has focused on the development of statistical approaches to risk-assessment in security-related contexts, with services provided to the New York Stock Exchange (NYSE), the Department of Homeland Security (DHS), the United States Agency for International Development (USAID), the Naval Research Laboratory (NRL), the Defense Advanced Research Projects Agency (DARPA), the Transportation Security Administration (TSA), and Customs and Border Protection (CBP), among others.

Dr. Cheryl Hendrickson: Dr. Cheryl Hendrickson is a Principal Research Scientist at the American Institutes for Research (AIR). She provides human capital consulting services to clients in the areas of

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

job and occupational analyses, competency modeling, test development and validation, performance management, organizational surveys, workflow analysis, and training needs assessment and evaluation. Dr. Hendrickson is currently leading a large-scale, criterion-related validation study for DHS and NRL for the purposes of improving the hiring system for a specialized, security-related job. Dr. Hendrickson is also currently conducting competency modeling work for a Department of Defense agency, and conducting a strategic job analysis and training needs assessment for the Federal Aviation Administration. The goal of the work is to identify the job and training requirements for the job of air traffic controller in the year 2018. Dr. Hendrickson is a member of the American Psychological Association (APA), the Society for Industrial and Organizational Psychology (SIOP), the Human Factors and Ergonomics Society (HFES), and the Personnel Testing Council of Metropolitan Washington (PTC/MW). She received her Ph.D. in Industrial and Organizational Psychology from the University of Nebraska.

Dr. Kathleen Kiernan: Dr. Kathleen Kiernan is a 29-year veteran of Federal Law Enforcement and is the CEO of an international consulting firm that supports federal and civil clients. She serves as a Senior Fellow at the George Washington University Homeland Security Policy Institute, and as the Director of Intelligence Analysis Programs at Johns Hopkins University. She currently chairs the National Board of Directors of the InfraGard National Members Alliance, an FBI-affiliated public-private partnership representing over 26,000 nationwide members whose mission is infrastructure protection. Dr. Kiernan previously served as the Assistant Director for the Office of Strategic Intelligence and Information for the Bureau of Alcohol, Tobacco, Firearms, and Explosives. She is a senior member on the International Association of Chiefs of Police (IACP) Terrorism subcommittee, an Intelligence Fellow (2001) and graduate of the FBI's National Executive Institute, and a graduate of DNI's SHARP Program. She has presented training on Critical Incident Management and Explosives response protocol on a worldwide basis, and is currently serving as a special advisor to the Director of the Combating Terrorism Task Force within the Department of Defense.

Dr. Richard Legault: Dr. Richard Legault is the Program Manager for Social and Behavioral Modeling in the Human Factors / Behavioral Sciences Division of the Science & Technology Directorate at the U.S. Department of Homeland Security. Prior to this position, he was an associate research scientist and co-director of the Terrorism and Preparedness Data Resource Center (TPDRC), Empirical Analysis of

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Parts 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Improvised Explosive Devices Project, and a statistical and methodological advisor for a number of other projects at the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland. Dr. Legault has published several scientific research articles, book chapters, and reports as well as one book on a range of issues addressing quantitative analyses of social science data including terrorism, survey analysis, policy evaluation, data usage and measurement, and violence-reduction strategies.

Dr. Christian Meissner: Dr. Christian Meissner is the Program Director for Law and Social Sciences at the National Science Foundation, and an Associate Professor of Psychology at the University of Texas at El Paso. Dr. Meissner's research examines the social and cognitive psychological processes that underlie the interviewing of individuals in forensic settings, including: (1) factors that influence eyewitness memory and lineup identification; (2) the detection of deception in forensic interviews; and (3) techniques employed in the interrogation of suspects and other non-cooperative individuals that lead to confessions. He is interested in understanding how best to apply basic theory in social and cognitive psychology to develop interview protocols and identification methods that will improve the diagnostic value of information obtained by law enforcement and intelligence personnel.

Elizabeth Sparano: Elizabeth Sparano is an Engineering Research Psychologist at the US Naval Research Laboratory. Currently she is the Principal Investigator on the Person Borne Improvised Explosive Device Indicators study, aimed at developing a scientifically grounded list of pre-incident behavioral indicators of suicide attacks that can be used by Government-employed security personnel to identify potentially "high-risk" individuals. Additionally, she is working with various performers on deception detection models and simulation efforts, as well as a behavior detection course for government clients. Elizabeth is certified in the Facial Action Coding System and has extensive experience working with and analyzing Action Units, particularly related to deception. Prior to her work at the Naval Research Laboratory, Ms. Sparano worked as a human factors analyst, writing profiles for a federal client, in addition to conducting classified research and analysis for numerous government agencies. Ms. Sparano has a concentrated educational background including B.S. in Psychology and Criminal Justice from Baldwin-Wallace College and an M.A. in Forensic Psychology from Marymount University.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know," as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~

Dr. Amber Story: Dr. Amber Story currently serves as the Deputy Director of the Division of Behavioral and Cognitive Sciences (BCS) at the National Science Foundation (NSF). Prior to this position, she served for six years as the director of NSF's social psychology program. Dr. Story received her Ph.D. in social psychology from Cornell University and held several academic positions before joining NSF.

Dr. Ruth P. Willis: Dr. Ruth P. Willis is the Head of the Adversarial Modeling and Exploitation Office within the Information Technology Division of the Naval Research Laboratory (NRL), Washington, D.C. She is responsible for formulating scientific goals for adversarial modeling and collaborating with a community of interest to conduct long-range research. Dr. Willis holds a Doctorate in Industrial/Organizational Psychology from the University of South Florida with a concentration in selection and performance appraisal. Previous assignments include Technical Lead for the Human Behavior Representation Program at the Defense Modeling and Simulation Office, Evaluation Team Lead for the Synthetic Theater of War-Europe, Principal Investigator for Distributed Teamwork. She is on the editorial board for the Journal of Defense Simulation, a previous Chair of the Computer Generated Forces and Behavior Representation Conference (1999-2001), the U.S. representative on multiple NATO and TTCP teams focusing on human behavior issues, and a US Patent holder. Awards include Department of the Navy Top Scientist for 2006, the Office of the Secretary of Defense Award for Excellence, Meritorious Unit Citation in support of NGA's Human Terrain and Socio-Cultural Dynamics Initiatives, as well as an NRL Award of Group Merit for the Clearwater/TINA program.

ACKNOWLEDGEMENTS

The Department of Homeland Security Science and Technology Directorate Human Factors Division acknowledges the work of the Independent Review Panel for Assessor Evaluation whose recommendations formed the basis of this report. The division also acknowledges the work of the scientists at the Naval Research Laboratory's Adversarial Modeling and Exploitation Office for substantive contributions, coordination activities, and oversight of the production of the report. The review panel would like to thank NRL and Carol Chang for data analysis and statistical support.

~~Sensitive Security Information (SSI)~~

~~WARNING: This record contains sensitive security information that is controlled under 49 C.F.R. Parts 15 and 1520. No part of this record may be disclosed to persons without a "need to know" as defined in 49 C.F.R. Part 15 and 1520, except with the written permission of the administrator of the Transportation Security Administration or the Secretary of Transportation. Unauthorized release may result in civil penalties or other action. For U.S. Government Agencies, Public disclosure is governed by 5 U.S.C. 552 and 49 C.F.R. Parts 15 and 1520.~~