

DOT/FAA/AR-00/01

Office of Aviation Research  
Washington, DC 20591

# Screeners Readiness Test Validation

J. L. Fobes, Ph.D.  
E. Neiderman, Ph.D.

Aviation Security Research and Development Division  
Federal Aviation Administration  
William J. Hughes Technical Center  
Atlantic City International Airport, NJ 08405

September 1999

This report is approved for public release and is on file at the William J. Hughes Technical Center, Aviation Security Research and Development Library, Atlantic City International Airport, New Jersey 08405.

This document is also available to the U.S. public through the National Technical Information Service, Springfield (NTIS), Virginia 22161.



U.S. Department of Transportation  
**Federal Aviation Administration**

## **NOTICE**

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the objective of this report.

**Technical Report Documentation Page**

<b>1. Report No.</b> DOT/FAA/AR-00/1		<b>2. Government Accession No.</b>		<b>3. Recipient's Catalog No.</b>	
<b>4. Title and Subtitle</b>  Screener Readiness Test Validation				<b>5. Report Date</b> September 1999	
				<b>6. Performing Organization Code</b> AAR-510	
<b>7. Author(s)</b> J. L. Fobes, Ph.D. and Eric C. Neiderman, Ph.D.				<b>8. Performing Organization Report No.</b> DOT/FAA/AR-99/XX	
<b>9. Performing Organization Name and Address</b> Federal Aviation Administration William J. Hughes Technical Center Atlantic City International Airport, NJ 08405				<b>10. Work Unit No. (TRAIS)</b>	
				<b>11. Contract or Grant No.</b>	
<b>12. Sponsoring Agency Name and Address</b> Federal Aviation Administration Associate Administrator for Civil Aviation Security, ACS-1 800 Independence Ave., S.W. Washington, DC 20591				<b>13. Type of Report and Period Covered</b>  Final Report	
				<b>14. Sponsoring Agency Code</b> ACS-1	
<b>15. Supplementary Notes</b> Draft Report prepared by William Maguire, Ph.D., Steven Siegel, Ph.D., and Melissa Dixon, Ph.D., Federal Data Corporation					
<b>16. Abstract</b>  The Screener Readiness Test was administered to 349 individuals at four major airports. Responses to verbal and image items were recorded and analyzed. Test items showed moderate difficulty and good reliability. The test was more difficult for those who were not native speakers of English but did not have adverse impact on racial or gender groups once this factor was controlled. A test structure and test scoring method is proposed that will allow valid and reliable assessment of initial screener readiness.					
<b>17. Key Words</b> Screener Readiness Test Aviation Security Human Factors				<b>18. Distribution Statement</b>	
<b>19. Security Classif. (of this report)</b> 14 Unrestricted		<b>20. Security Classif. (of this page)</b> Unrestricted		<b>21. No. of Pages</b> 23	<b>22. Price</b>

## TABLE OF CONTENTS

	Page
ACRONYMS	iv
1 INTRODUCTION	1
1.1 Background	1
1.2 Purpose	1
2 METHOD	2
2.1 System description	2
2.2 Test overview	2
2.3 Participants	3
2.4 Equipment	3
2.5 Data collection	3
2.6 Test procedure	4
2.7 Test scoring	4
3 CRITICAL ISSUES AND CRITERIA	5
4 RESULTS	6
4.1 ISSUE 1. Do the test items have acceptable psychometric properties?	6
4.2 ISSUE 2. Is there an adverse impact of test on any group?	6
4.3 ISSUE 3. How effective are the two SRT modules?	12
4.4 ISSUE 4. How 'user-friendly' is the SRT?	12
4.5 ISSUE 5. How 'trainable' is the SRT?	13
4.6 ISSUE 6. Do the Rapiscan and EG&G X-ray image sets result in equal performance?	13
4.7 ISSUE 7. Do black and white and color images result in equivalent performance?	14
5 IMPLICATIONS FOR THE STRUCTURE OF THE SRT	14
6 CONCLUSIONS	17
7 REFERENCES	17

## LIST OF TABLES

Tables	Page
1. Scoring Weights for X-ray Image Items	5
2. Number of Screeners for Each Major Racial Group as a Function of Native Language and Gender	7
3. Mean SRT Raw Scores by Gender	7
4. SRT Raw Scores by Native Language Status	7
5. Groups in Current Sample and in the Target Screener Population	8
6. Sample Means and Weighted Estimates of the Population Means for the SRT Raw Scores	8
7. Mean SRT Composite Scores by English Language Status as a Function of Race and Gender	<b>Error! Bookm</b>
8. Mean Percent Correct for Verbal Test Items by Question Content Category	9
9. Number of Difficult Questions by Explanation Category	10
10. Number of Questions Showing Adverse Impact for Non-Native English Speakers by Explanation Category	12
11. Mean Detection and False Alarm Rates and Composite Scores for Matched Color and B&W Image Sets	14
12. Estimated Reliability and Completion Time of the Proposed Test Structure	15
13. Composite Scores and Associated Percentiles for Experienced Screeners	17

## ACRONYMS

$\alpha$	Coefficient Alpha
ACSSP	Air Carrier Standard Security Program
ANOVA	Analysis of Variance
B&W	Black and White
C	Composite Test Score
CBT	Computer-Based Training
CSS	Checkpoint Security Supervisor
CV	Correct Verbal Percentage
FAA	Federal Aviation Administration
FAR	Federal Aviation Regulation
IAD	Washington-Dulles International Airport
JFK	J. F. Kennedy International Airport
MBS	Modular Bomb Set
MCO	Orlando International Airport
OJT	On-the-Job Training
$P_d$	Probability of Detection
$P_{fa}$	Probability of False Alarm
RT	Reaction Time
SD	Standard Deviation
SFO	San Francisco International Airport
SRT	Screener Readiness Test

## **1 INTRODUCTION.**

Federal Aviation Regulations (FAR) require air carriers to provide for the safety and security of passengers and their property. Air carriers do this with security equipment and trained personnel to screen passengers and their baggage before they board the aircraft. The Federal Aviation Administration (FAA), working with the U.S. aviation industry, is developing new equipment and procedures to improve aviation security. Investigating human factors is critical to the success of these efforts. The White House Commission on Aviation Safety and Security, and the General Accounting Office, recognized this need and recommended a greater focus on human factors and training to complement advanced technologies.

According to FAR § 108.17 (*Use of X-Ray Systems*), there shall be a program for initial and recurrent training of operators of X-ray systems. This program shall include training in the efficient use of X-ray systems and the identification of weapons and other dangerous articles. Section XIII of the Air Carrier Standard Security Program (ACSSP) presents the standards for training and testing of persons performing screening and security functions. The ACSSP also describes English language proficiency requirements for screeners [1: Section XII.B.1.f, g].

### **1.1 BACKGROUND.**

For many years, the only FAA-approved training for screener personnel was developed by the Air Transport Association. Their 12-hour, initial screener-training program includes 40 multiple choice questions and 40 X-ray images to assess mastery prior to On-the-Job Training (OJT). In April 1997, the FAA also approved the use of a Computer-Based Training (CBT) system for initial screener training prior to OJT. This CBT system by Safe Passage has a library of test questions and the trainee is presented with unit tests, a 50-item content mastery test, and a 50-item threat image interpretation test to assess mastery.

Screener training options have increased and are likely to continue to grow in the future. The FAA's Aviation Security Human Factors Program will soon test screener candidates after their initial training with one of four different CBT systems for initial screener training [2]. The issue of determining screener knowledge will be accomplished with a valid, reliable, and non-biased test for initial screener training, the Screener Readiness Test (SRT) the FAA's Aviation Security Human Factors Program has developed. Most importantly, the SRT will be used to determine whether or not a screener candidate has sufficient knowledge to proceed to the next step in their training. A critical step in the development of the SRT was the assessment of the proposed test items reported here.

### **1.2 PURPOSE.**

The Test and Evaluation Plan for this effort to develop an appropriate test of screener readiness is contained in [3]. This Test and Evaluation Report addresses the findings on the critical operational issues investigated during field testing of the SRT. The main goal of the project was to develop a computerized SRT that can run on both MacIntosh and PC platforms. The SRT should have the following characteristics:

- a. job-relevant, performance-oriented testing of trainees' knowledge,

- b. accurate sampling of the entire range of screeners' job functions and roles, and
- c. large pool of test questions to minimize the likelihood of cheating.

These requirements were met by the development of a large set of multiple choice written and X-ray image items that broadly test the knowledge that should be acquired during initial screener training. The development of these items was described in an earlier document [4]. Other requirements of the SRT include:

- a. psychometric qualities of internal and criterion-related validity,
- b. fair and unbiased against specific population groups (as defined by the Equal Employment Opportunity Act under Title VII of the Civil Rights Act of 1964), and
- c. 'user friendly' with minimal demands for administration.

These requirements were addressed in the field test described in this document.

## **2 METHOD.**

The initial item set for the SRT and the computer interface used to administer the test were evaluated. A large set of multiple choice questions (264) and X-ray images of threat and non-threat bags have been developed for both Rapiscan (1,363 images) and EG&G (1,367 images) X-ray machines as described in a previous document [5]. An interface was developed to present items on a desktop computer using standard web browser software. This interface served as the prototype for the final computer-based SRT. The interface and items were tested on a large pool of screeners selected so that diverse ethnic and gender groups were represented. These tests were conducted at John F. Kennedy International (JFK), Orlando International (MCO), San Francisco International (SFO), and Washington Dulles International (IAD).

### **2.1 SYSTEM DESCRIPTION.**

The test library of items evaluated consisted of 264 multiple-choice items and 2,730 X-ray image items. The composition of the image sets is described in an earlier document [4]. Because there were 2,730 X-ray images in the image library, 429 representative images from the various image classes were chosen to evaluate the critical operational issues. Written items were sampled from four content areas.

### **2.2 TEST OVERVIEW.**

Initial testing of the modules took place at the Atlantic City International Airport and emphasized interface usability. That is, could individuals from the target screener population understand the instructions, use the equipment, and complete the test module in a reasonable period of time? The results of that test are described in a previous document [6].

The second phase of testing was conducted at four sites with numerous screeners that responded to a question module and an image module. Responses were collected by computer and the

results were incorporated into a database, with data exported for computerized statistical analyses.

Data were collected on accuracy and latency of item responses, as well as demographic information about individual screeners. Detailed information was collected about individual items so that the length and composition of the SRT could be finalized based upon specific criteria for fairness and reliability of the overall test.

### **2.3 PARTICIPANTS.**

A total of 349 screeners from the four airport sites (JFK, MCO, SFO, and IAD) took the test prototype. All were currently working as certified screeners.

### **2.4 EQUIPMENT.**

The test software was installed on four computers. The test modules were JavaScript applications running on a standard PC-desktop computer with Netscape Communicator 4.08. The applications included computer usage and test taking instructions, an information sheet to include demographic information, timing and time limits for individual items, and automatic data collection.

### **2.5 DATA COLLECTION.**

Data were collected on accuracy and latency of item responses, as well as demographic information about individual screeners. Detailed information was collected about individual test items so that the length and composition of the SRT could be finalized based upon specific criteria for fairness and reliability of the overall test.

Individual item responses and their associated response times were recorded by the computer. An Excel database was created to take the item responses and compare them to an answer key as part of the data analysis. The data were exported to the Statistical Package for the Social Sciences for most analyses. The set of 264 textual test items were divided into 8 presentation groups. The questions were grouped so that the full heterogeneity of content was represented in each group and the groups did not differ in content. The set of X-ray image items (2,730 images) is described in Appendix A of [1]. The images were grouped based on a number of variables (e.g., X-ray machine type, level of clutter, bag type, and threat status). A subset of these images was chosen for the item validation. There were 429 images used and they were assembled into 12 image sets.

There were also some special image sets. One set of images was Black and White (B&W) and consisted of exactly the same images as one of the sets of color images. These two sets were specifically designed to evaluate the effect of color versus B&W images on performance. Four other sets were a mixture of EG&G and Rapiscan images to compare performance across machines. The sets were organized as matched pairs where every EG&G image in one was matched by a Rapiscan image of the identical bag in the other image set, and vice versa.

## **2.6 TEST PROCEDURE.**

Before the test was administered, a verbal item set and an image item set were chosen at random for presentation to the participant as a demonstration. The test, the purpose of the testing was explained to each participant, after which their verbal consent to participate was obtained. Each screener was seated in front of the computer with the first page of instructions on the screen and asked to proceed. Each participant completed a full test module consisting of both verbal content and image-based performance items with responses in a multiple-choice format. Individual test sessions consisted of random selection of items from a larger item pool. All questions were answered by keyboard entry by the screeners and scoring of the test and response latencies were recorded automatically by the computers.

## **2.7 TEST SCORING.**

The test consists of both image and verbal components and, in practice, will yield a single score for each screener.

### **2.7.1 A Composite Scoring Algorithm and Analysis of Performance for the Composite Score.**

It is intended that the SRT be fielded as an easy-to-use and easy-to-score test. Each screener's score will be calculated by computer, but scoring should be easy to understand and evaluate. For this reason, a composite scoring technique was devised as the scoring method to be used when the SRT is fielded. The composite score can vary from 0 to 100. The pass/fail status of any completed test will be based upon a comparison of the composite score to a cutoff score to be determined in the future. The composite score was constructed from a verbal score that varied from 0 to 100 and from an image score that varied from 0 to 100. The following scoring algorithm meets these criteria.

The verbal test is scored in a straightforward manner in terms of the percentage of verbal items that are answered correctly. For example, a screener who answers 40 out of 50 items correctly would obtain a score of 80.

The image test algorithm is more complex because there are two types of items: threats and non-threats. It is further complicated because there are three possible responses: not a threat, possible threat, and definite threat. The first step in scoring the image test was attributing a point value to each item using the point assignments shown in table 1.

TABLE 1. SCORING WEIGHTS FOR X-RAY IMAGE ITEMS

Item	Response Choice		
	Not a Threat	Possible Threat	Definite Threat
Non-Threat	+2	-.67	-2
Threat	-3	+1	+3

Because threats and non-threats are scored differently, maximum and minimum raw scores were determined by the relative proportion of threats and non-threats. The scoring weights above were chosen specifically for a test that has 60% non-threat images and 40% threat images. Randomly guessing will yield an average item score with an expected value of 0. This occurs if the proportion of questions is 60/40 and the item weights determined as shown in table 1. The maximum item score is 2.5, and the minimum item score is -2.5.

To create an image score that varies from 0 to 100, the following formula was used:

$$(1) \text{ IMAGE SCORE} = 20 * (\text{ITEM AVERAGE}) + 50.$$

To derive a composite score, the verbal scores and image scores are combined with equal weights. If both parts are weighed equally:

$$(2) C_{50-50} = (.50 * \text{IMAGE SCORE}) + (.50 * \text{VERBAL SCORE}).$$

The scores of all the screeners who participated in the evaluation were converted to composite scores using the above algorithm.

### **3 CRITICAL ISSUES.**

There were seven critical issues identified for this test and evaluation [1].

- Issue 1 - Do the test items have acceptable psychometric properties?
- Issue 2 - Is there an adverse impact and test bias?
- Issue 3 - How effective are the SRT's two modules?
- Issue 4 - How 'user friendly' is the SRT?
- Issue 5 - How trainable is the SRT?
- Issue 6 – Do the Rapiscan and EG&G image sets produce equivalent performance?
- Issue 7 – Do B&W and color images produce equivalent performance?

## **4 RESULTS.**

### **4.1 ISSUE 1. DO THE TEST ITEMS HAVE ACCEPTABLE PSYCHOMETRIC PROPERTIES?**

Each screener's data were initially summarized as three raw scores. The three scores were the percentage of Correct Verbal (%CV) questions on the verbal test, the percentage of threats correctly identified as possible or definite threats (Probability of Detection [ $P_d$ ]), and the percentage of non-threat images incorrectly identified as possible or definite threats (Probability of a False Alarm [ $P_{fa}$ ]). The mean CV was 0.66, Standard Deviation ( $SD$ ) = 0.17, the mean  $P_d$  was .72,  $SD$  = 0.17, and the mean  $P_{fa}$  rate was 0.31,  $SD$  = 0.22. The composite score  $C_{50-50}$  was also calculated and the mean score was 67.4,  $SD$  = 10.7. These means and standard deviations indicate that the difficulty of the item sets chosen are consistent with good test practices for this population [7].

#### **4.1.1 Reliability of the Verbal Tests.**

The reliability of each of the verbal test modules was calculated using Coefficient Alpha ( $\alpha$ ) as the measure. Reliability of individual modules ranged from  $\alpha$  = 0.69 to  $\alpha$  = 0.89, and the average  $\alpha$  = 0.84. The reliability of the verbal test items is consistent with good test practices.

#### **4.1.2 Performance on Image Test Items.**

The  $P_d$  rate was different for the various classes of threats. Separate 2 x 2 Chi-Square analyses revealed that screeners were better at detecting (definite threat or possible threat) the presence of both Modular Bomb Set (MBS) images (72%) and guns (75%) compared to detecting knives (64%) or grenades (64%),  $\chi^2 p < .01$ .

Clutter affected performance and there was a significant increase in  $P_{fa}$  for high clutter bags (mean = 0.42) in comparison with low clutter bags (mean = 0.27) [ $t(236) = 5.99, p = .001$ ]. There was no difference in  $P_d$  for low and high clutter bags.

#### **4.1.3 Reliability of the Image Test Items.**

The reliability was calculated for each of the image test modules using  $\alpha$ . Reliability of individual image modules ranged from  $\alpha$  = 0.36 to  $\alpha$  = 0.90, and the average  $\alpha$  = 0.76. The reliability of the image test items is consistent with good test practices.

### **4.2 ISSUE 2. IS THERE AN ADVERSE TEST IMPACT ON ANY GROUP?**

There was a very large number (51%) of screeners who reported that English was not their native language. This factor was included along with race and gender in the overall analysis of the results. The numbers for each group represented is shown in table 2.

TABLE 2. NUMBERS OF SCREENERS FOR EACH MAJOR RACIAL GROUP AS A FUNCTION OF NATIVE LANGUAGE AND GENDER

	English 1 <sup>st</sup> Males	English 1 <sup>st</sup> Females	English 2 <sup>nd</sup> Males	English 2 <sup>nd</sup> Females	Total
<b>Asian</b>	21	10	45	47	123
<b>Black</b>	12	42	2	6	62
<b>Hispanic</b>	3	6	16	35	60
<b>Other</b>	3	22	8	5	38
<b>White</b>	30	23	4	9	66

#### 4.2.1 Statistical Analyses Based on Current Sample.

The three raw scores analyzed were obtained for each screener (CV, P<sub>d</sub>, and P<sub>fa</sub>) in a factorial multivariate Analysis of Variance (ANOVA). The independent variables in the analysis were gender, race, and native language status. Only the variables gender [ $F(3,327) = 3.26, p = .02$  (Table 3)] and native language status [ $F(3,327) = 11.2, p = .001$  (Table 4)] were significant in this analysis. The univariate post-hoc tests for gender showed a significantly higher score for males on the CV [ $F(1,329) = 4.08, p = .04$ ] and a significantly lower score for males on P<sub>d</sub> [ $F(1,329) = 4.66, p = .03$ ]. The second language speakers scored significantly lower on the CV [ $F(1,329) = 31.02, p = .00$ ] and significantly higher on P<sub>fa</sub> [ $F(1,329) = 5.24, p = .02$ ]. The main effect of native language accounted for 9% of the score variance, whereas gender accounted for only 3%. For this reason, English language status as a covariate was used in the main analysis of adverse impact in the weighted analysis.

TABLE 3. MEAN SRT RAW SCORES BY GENDER

	CV	P <sub>d</sub>	P <sub>fa</sub>
<b>Male</b>	.69	.69	.33
<b>Female</b>	.64	.75	.31

TABLE 4. SRT RAW SCORES BY NATIVE LANGUAGE STATUS.

	CV	P <sub>d</sub>	P <sub>fa</sub>	C
<b>English 1<sup>st</sup></b>	.72	.72	.25	71.50
<b>English 2<sup>nd</sup></b>	.60	.73	.36	63.90

#### 4.2.2 Weighted Statistical Analyses Based on Target Population.

The sample test sites were chosen to acquire a sufficient number of screeners in each important demographic group. The sample demographics, however, were not the same as the screener population at all major airports as determined by a previous FAA census. Therefore, sample statistics were calculated and population means estimated by adjusting the sample for the

proportion of racial groups in the population determined by the 1994 FAA census of major airports.

First, the proportion for each racial by gender group represented in the target population was calculated using the population statistics from the 1994 FAA census. Given these population proportions and the number of individuals in each group of the sample, a proportional weight was derived for each racial by gender group in the current sample. For example, Asian females represented 10% of the target population and 18% of the current sample, The sample mean can be expressed as a sum: (Sample Mean =  $w_1$  \* Mean of Group 1 +  $w_2$  \* Mean of Group 2 + ...) where the weights  $w_n$  represent the proportion of the sample that the group represents, .18 in the case of Asian females. The weighted population mean can be expressed similarly (Weighted Population Mean =  $w_1$  \* Mean of Group 1 +  $w_2$  \* Mean of Group 2 + ...) where the weights  $w_n$  represent the proportion of the population that the group represents, .10 in the case of Asian females. The final result of weighting the scores is a distribution of scores that better resembles the scores that are likely to be obtained in the target population. The proportions of each racial group represented in the sample and in the 1994 FAA Census are included in table 5. The corresponding mean scores for the three raw test scores are shown in table 6.

TABLE 5. GROUPS IN THE CURRENT SAMPLE AND IN THE TARGET SCREENER POPULATION

<b>Ethnic Group</b>	<b>Sample Proportion</b>	<b>Census Proportion</b>
<b>Asian</b>	.35	.22
<b>Black</b>	.18	.42
<b>Hispanic</b>	.17	.15
<b>Other</b>	.11	.03
<b>White</b>	.19	.17

TABLE 6. SAMPLE MEANS AND WEIGHTED ESTIMATES OF THE POPULATION MEANS FOR THE SRT RAW SCORES

	<b>Sample Mean</b>	<b>Weighted Population Estimate</b>
<b>CV</b>	.66	.71
<b>P<sub>d</sub></b>	.72	.69
<b>P<sub>fa</sub></b>	.31	.34

In addition to the analysis of adverse impact on raw scores as presented in section 4.2.1, an evaluation of whether the test has an adverse impact on any particular racial group, data for the composite scoring method discussed in section 2.7 were used. Recall that the composite score is defined as  $C_{50-50} = (.50 * \text{IMAGE SCORE}) + (.50 * \text{VERBAL SCORE})$ . Note that racial group and English language status are not independent factors;  $\chi^2 = 121.77, p < .001$ . Therefore, a weighted analysis was performed on the composite scores with race and gender as independent variables and English language status as a covariate. The impact of demographics on the scores was evaluated with Bonferroni weighted contrasts [8]. Four of the contrasts compared the mean

performance of a specific group against the weighted means of the other groups. Another contrast evaluated the effect of gender on performance. Separate analyses were performed for each of the two versions of the composite score. None of these analyses showed a significant adverse impact for any group.

TABLE 7. MEAN SRT LANGUAGE-ADJUSTED COMPOSITE SCORES AS A FUNCTION OF RACE AND GENDER

	<b>Group Average</b>	<b>Weighted Average All Others</b>	<b>Score Difference</b>
<b>Race</b>			
Asian	65.5	67.3	-1.8
Black	65.8	67.7	-1.9
Hispanic	69.7	66.4	3.3
Other	67.3	66.9	0.4
White	69.3	66.4	2.9
<b>Gender</b>			
Female / Male	66.8	67.1	-0.3

#### 4.2.3 Detailed Analysis of the Original Test Items.

The 264 verbal questions were originally grouped into four content categories: (1) Background, Responsibilities, and Operations, (2) Identifying the Threat, (3) Procedures for Passenger Screening, and (4) Atypical Passengers and Special Situations. A one-way ANOVA on the percent correct for the text items revealed no differences across the four question content categories. This suggested that no particular category had an adverse impact on screeners' performance. The mean percent correct for the verbal items in each of the four question content categories are listed in table 8.

TABLE 8. MEAN PERCENT CORRECT FOR VERBAL TEST ITEMS BY QUESTION CONTENT CATEGORY

<b>Question Content Category</b>	<b>Correct Verbal Mean</b>	<b>Standard Deviation</b>
<b>(1)</b> Background, Responsibilities, and Operations	63.3	24.5
<b>(2)</b> Identifying the Threat	65.4	20.7
<b>(3)</b> Procedures for Passenger Screening	66.6	22.9
<b>(4)</b> Atypical Passengers and Special Situations	67.9	21.9

An examination of the verbal test items revealed considerable variation in item difficulty. Sixty-one of the 264 questions (23%) were identified as difficult questions defined as questions answered correctly by equal to or less than 50% of all screeners. Thirty-four of the 264

questions (13%) were identified as easy questions, questions that were answered correctly by equal to or greater than 90% of all screeners. Examination of these questions suggests that most of these questions test important content regarding screeners' jobs. They frequently refer to very common procedures that are emphasized more during training and reinforced with practice.

A mixture of difficult and easy questions is desirable in a test if the source of difficulty is fair and relevant to job performance. Further analyses to help explain the reason for item difficulty was performed. One concern was that certain questions may have been difficult to understand because of their reading level. Although the Flesch-Kincaid index was developed to be used with prose passages rather than with multiple choice questions, it does provide a tested and accepted measure of readability. The mean Flesch-Kincaid readability score for the test questions was 9.17. However, individual questions varied greatly, as did the readability of any short sample of text, because the measure is only sensitive to sentence and word length. (Readability for each question and its correct answer were calculated with the Flesch-Kincaid Readability Measure; Grade Level =  $(L * 0.39) + (N * 11.8) - 15.59$ , where L is the average sentence length and N is the average number of syllables per word.) A Pearson correlation was performed to investigate the relationship between the readability of the 264 questions and the error rates for these questions. The analysis revealed no relationship between these two measures, suggesting that factors other than readability (e.g., question content or training) are related to the screeners' poor performance on certain text questions.

Five general categories were created to classify the questions identified as being difficult. Table 9 presents the number of questions that can be grouped into each category, with the possibility that some questions could be categorized in more than one group. Category 1 describes questions that screeners may have answered incorrectly not necessarily because the content was difficult but because the question was interpreted on a different level. For example, the correct answer for the question, "Who is primarily responsible for maintaining and testing checkpoint security equipment?" is that the airlines are responsible. Most screeners responded that the Checkpoint Security Supervisor (CSS) is responsible. They may have responded in this manner because the CSS is the person to whom they report and thus is seen as the person responsible for the checkpoint.

TABLE 9. NUMBER OF DIFFICULT QUESTIONS BY EXPLANATION CATEGORY.

<b>Explanation for Difficulty</b>	<b>Number of Questions</b>
<b>(1)</b> Correct answer versus practiced procedure	6
<b>(2)</b> Interpretation of wording	14
<b>(3)</b> Negative answer expected	12
<b>(4)</b> Difficult content	40
<b>(5)</b> Erroneous wording	2

NOTE: The number of questions does not add to 61 because some fit into more than category.

Category 2 describes questions and/or their answers that were poorly or ambiguously worded, causing the screeners to misinterpret the question. For example, the correct answer for the question, "What common carry-on items produce X-ray images that may resemble explosives?" is "food". Most of the screeners incorrectly said that tools may resemble explosives. Although food may produce similar X-ray images because of the materials that compose them, screeners are taught that explosives may be disguised as tools, and for this reason they may have misinterpreted the phrase "may resemble explosives."

Category 3 includes questions that ask the screeners to choose a negative or contradictory answer. For example, in the question, "Which of the following is *not true* of a malfunctioning X-ray system," the word "not" could be missed when reading this question. This could result in the screeners answering with what is *true* of a malfunctioning X-ray system rather than what is *not true*. In fact, most screeners answered that "unusual looking images may indicate a malfunction," which suggests that they missed the word "not" in the question.

Category 4 identifies questions about difficult or very specific content material. For example, a commonly incorrect answer for the question, "If you cannot resolve an alarm which occurs on the right forearm of a person using the hand-held metal detector" was to do a whole body pat-down search rather than to do a limited pat-down search of the forearm only. Here, the question requires the screeners to differentiate the specifics of doing a pat-down search.

Finally, Category 5 identifies two questions that contained typos in the test and therefore may not accurately reflect the screeners' actual knowledge. One of these questions was, "Which of the following *is statements* about OJT is false?" If read as "Which of the following is...," the reader may have neglected the phrase "is false" and answered with a positive answer. If on the other hand, it was read as "Which of following statements...," the reader may have correctly answered with a negative answer. The second of these questions was "Which of the following *is true* regarding whole body pat down searches?" This question was intended to ask "Which of the following *is not true* regarding whole body pat-down searches?" These two questions will be edited to read as originally intended.

There was particular concern with questions that exhibited differential difficulty for those screeners who were not native English speakers. Separate Chi-Square analyses were conducted for each of the 264 questions to compare whether native and non-native English speakers produced similar proportions of correct and incorrect answers for each question. These Chi-Square analyses revealed that 55 of the written test questions may have an adverse impact on the performance of non-native English speakers relative to native English speakers. Altogether, 8 (14.5%) of these 55 questions were answered correctly by less than or equal to 50% of all screeners, indicating that these questions were difficult for everyone.

A similar categorization method was used to group these 55 questions that showed an adverse impact on non-native English speakers. An examination of the content and the commonly incorrect answers for each of the questions showed that approximately 30% of the questions may have been answered incorrectly due to the nature of the wording or sentence construction. Approximately 85% of the questions may have been answered incorrectly due to a lack of knowledge about specific procedures or threat characteristics (table 10).

TABLE 10. NUMBER OF QUESTIONS SHOWING ADVERSE IMPACT FOR NON-NATIVE ENGLISH SPEAKERS BY EXPLANATION CATEGORY.

<b>Explanation for Difficulty</b>	<b>Number of Questions</b>
(1) Correct answer versus practiced procedure	0
(2) Wording / Interpretation of question	12
(3) Negative answer expected	5
(4) Difficult content	47
(5) Erroneous wording	0

Note: The number of questions does not add to 55 because several questions could be placed into more than category.

An ANOVA was conducted on the number of questions showing adverse impact towards non-native English speakers as a function of question content. The four question categories were: (1) Background, Responsibilities, and Operations, (2) Identifying the Threat, (3) Procedures for Passenger Screening, and (4) Atypical Passengers and Special Situations. The number of questions showing adverse impact towards non-native English speakers significantly differed across the four categories of question content [ $F(3, 263) = 3.08, p = .03$ ]. Bonferroni post-hoc analyses revealed a difference between the "Procedures for Passenger Screening" category and the "Background, Responsibilities, and Operations" category [ $t(165) = 2.63, p = .05$ ] with a greater number of difficult questions in the former. No differences in the number of difficult questions were found for any other category comparisons.

### **4.3 ISSUE 3. HOW EFFECTIVE ARE THE TWO SRT MODULES?**

For each module, (multiple-choice and image tests), data were collected to examine task execution times. There was an upper limit with items presented for up to 30 seconds for each image item and 45 seconds for each verbal item. Originally, both portions of the test presented test items for 30 seconds each; however, it was discovered that more screeners, especially those reporting English as a second language, ‘timed-out’ on the verbal items than on the image items. Reaction Times (RT) for each interface were calculated (verbal items RT: mean = 28.2 sec,  $SD = 8.24$ ; image item RT: mean = 7.23 sec,  $SD = 4.59$ ). Both interfaces can be considered effective. Longer reaction times and the increased time-outs associated with the verbal items can be explained by the large numbers of screeners reporting English to be their second language.

### **4.4 ISSUE 4. HOW ‘USER-FRIENDLY’ IS THE SRT?**

Behavioral observations were carried out during the initial checkouts of the SRT. These included noting any step repetitions, errors, assistance needed, and verbal or non-verbal complaints. Incorrect navigation was the most frequent usability problem in earlier versions of the test. These mostly related to screeners’ difficulty using the mouse, as a number of screeners were only minimally familiar with computers. As a result, later versions of the SRT had all test responses being made via keyboard entry, using the number keys to answer questions in a

multiple-choice format. Earlier versions of the test had the screeners answer using the letter keys A, B, C, or D. It was found that screeners continued to have some difficulty navigating the use of the keyboard (e.g., some had trouble finding the letter keys or they timed out before they found the letters).

The present test version was changed from letter key entry to number key entry, in anticipation of an easier use of the sequential 1, 2, 3, or 4 number keys. This proved more user friendly, especially to those with little computer experience. “How to use the keyboard” instructions, which included a picture of the keyboard with the appropriate number keys to be used highlighted in red, were also provided at the beginning of the test. Screeners were instructed to press the spacebar on the bottom of the keyboard to move to the next screen. An introduction to the SRT test was presented to the screeners and an example of the test questions was provided. The multiple choice portion of the test then began.

After the verbal items were answered, a new set of instructions presented the screener with the appropriate set of number keys to be used for the image items. Screeners received instruction on which number keys corresponded to the responses of “Definite Threat,” “Possible Threat,” and “No Threat.” When each image was presented, number keys highlighted in different colors appeared at the bottom of each image. Screeners were instructed to respond “1” (Definite Threat), “2” (Possible Threat), and “3” (No Threat). When the image items were answered, the screeners were then given a message on screen that they had completed the test.

The only other observed screener difficulties were infrequent pressing of wrong keys and misunderstanding the test example. This was interpreted to be more a factor of the English language skills of the particular screener than the usability or user-friendliness of the SRT. Overall, the SRT is successful in being user-friendly.

#### **4.5 ISSUE 5. HOW ‘TRAINABLE’ IS THE SRT?**

There was little or no observed difficulty in the trainability of the SRT. The test was designed to have easy-to-understand instructions so that the test can be taken with minimal supervision. There is no anticipated difficulty in training the new screeners in navigating through the SRT in a self-paced manner.

#### **4.6 ISSUE 6. DO THE RAPISCAN AND EG&G X-RAY IMAGE SETS RESULT IN EQUAL PERFORMANCE?**

The special matched sets of EG&G and Rapiscan images were analyzed in a separate analysis. Performance for each image was first converted to an average performance by averaging each subject's points scored on the image test using the scoring weights in table 1, section 2.7. That score was then converted to an image score on a 0 to 100-point scale. Scores for Rapiscan and EG&G images were compared in a paired samples *t*-test. The mean Rapiscan image score was 66.8 and did not differ significantly from the mean EG&G image score of 65.8.

#### **4.7 ISSUE 7. DO BLACK & WHITE AND COLOR IMAGES PRODUCE EQUIVALENT PERFORMANCE?**

One of the color image sets (set 6) was matched with a B&W version (set 12), and the performance of screeners on these two sets was compared in separate *t*-tests of independent means. Neither the  $P_d$ , the  $P_{fa}$ , nor the composite scores showed a significant difference in performance between the color and B&W image sets (table 11).

TABLE 11. MEAN DETECTION AND FALSE ALARM RATES, AND COMPOSITE SCORES FOR MATCHED COLOR AND B&W IMAGE SETS

	<b>Mean</b>	<b>SD</b>
<b><math>P_d</math></b>		
Color	0.69	0.18
B&W	0.68	0.21
<b><math>P_{fa}</math></b>		
Color	0.39	0.23
B&W	0.42	0.25
<b>Image Composite</b>		
Color	64.96	9.75
B&W	63.03	7.78

Note: None of these comparisons statistically differed.

### **5 IMPLICATIONS FOR THE STRUCTURE OF THE SRT.**

#### **5.1 LENGTH OF THE TEST.**

The internal validity of both verbal and image modules was found to be good. Based upon these data, a 40-item verbal test should have good reliability ( $\alpha = 0.86$ ). The image modules had a slightly lower internal validity but can be completed more quickly, therefore a 50-item image test should have internal validity ( $\alpha = 0.83$ ). The test as a whole is predicted to have a reliability of  $\alpha = 0.91$ . The reliability of the image test can be maximized by having a fixed internal structure for the number of high and low clutter bags or the type and number of threats, as these variables are known to affect accuracy.

The amount of time to complete the test was estimated from the average RTs obtained for the verbal and image test items (table 12). On average, it took the screeners less than half hour to complete the test. Based on these data, each verbal item can be presented for 45 seconds and each image item for 15 seconds, yielding a maximum of 42.5 minutes to complete the test.

TABLE 12. ESTIMATED RELIABILITY AND COMPLETION TIME OF THE PROPOSED TEST STRUCTURE

Test Module	Reliability	Mean Time to Complete (min)
40-item verbal module	.86	18.80
50-item image module	.83	6.03
Total test	.91	24.80

## **5.2 SCORING THE TEST.**

The scoring system in section 2.7 has a number of desirable features. While separate verbal and image scores may provide information about deficiencies that exist in training, the 0 to 100-point combination scoring format is easy to understand and to evaluate. Image scoring does not reward particular strategies for favoring one type of response over another. As discussed in section 4.2, extensive analyses of the test using this composite scoring method did not show significant adverse impact for either race or gender. Although some impact was found for nonnative English speakers, English comprehension is a job requirement [1] and this provides the FAA with its first measure of whether or not this requirement is being met.

## **5.3 PASSING SCORE.**

This document does not recommend a particular passing score. The choice of a cut-off score will be influenced by a number of factors including the practical impact of a particular cut-off on meeting manpower requirements, assuring the preparedness of new hires, and creating the criteria that determine whether and how the test may be re-taken. It is assumed that the cut-off score for pass/fail will correspond to some percentile to be selected in the future. Bear in mind that screeners in this sample were experienced. The percentiles reported here therefore do not necessarily predict the rate at which newly trained screener candidates would pass or fail.

Table 13 lists each SRT composite score and the the accompanying percentile. For example, the 25<sup>th</sup> percentile corresponds to a SRT score of 64.5. If this percentile was used as the cut-off score, the bottom 25% of the experienced screeners taking the SRT would be rejected. Similarly, selecting the 75<sup>th</sup> percentile score of 77 would eliminate the lowest 75% of the test takers. Table 13 represents SRT scores for those speaking English as their first language and therefore enforces the ACSSP English language requirements.

These percentiles were adjusted to reflect the ethnic proportions in the screener population who would be expected to score at or below the corresponding SRT score (table 5). Percentiles do not exactly correspond to the percentage of individuals in the sample (table 5) that scored below a certain score because they were calculated with weighted percentages of each ethnic group in the sample. Calculating the percentiles associated with any SRT scores in table 13 began by determining how many individuals in each ethnic group had scores below a particular cut-off. This number was then converted to a percentage of that group in the sample. For example, if 12 blacks scored below 65, they represented 22% of the sample of 54 blacks. The overall percentile was then calculated by weighting the proportion of each group that fell below a certain score by their proportion in the population. For example, blacks represent 42% of the screener census

population. Therefore, if the 22% that fell below the cut-off score of 65 represented 9.24% of the population (.42 \* .22). The percentiles in table 13 were calculated by adding weighted percentiles for each group to estimate the percentage of the screener population above and below this cut-off score.

TABLE 13. SRT COMPOSITE SCORES AND ASSOCIATED PERCENTILES FOR EXPERIENCED SCREENERS

SRT Score	Percentile	SRT Score	Percentile	SRT Score	Percentile
43	0.00	61	0.18	79	0.82
44	0.00	62	0.20	80	0.84
45	0.01	63	0.23	81	0.92
46	0.03	64	0.24	82	0.92
47	0.03	65	0.26	83	0.94
48	0.03	66	0.30	84	0.97
49	0.03	67	0.36	85	0.98
50	0.03	68	0.38	86	0.98
51	0.03	69	0.41	87	0.98
52	0.03	70	0.48	88	0.98
53	0.04	71	0.53	89	0.98
54	0.04	72	0.57	90	0.99
55	0.05	73	0.60	91	0.99
56	0.09	74	0.62	92	0.99
57	0.12	75	0.64	93	0.99
58	0.12	76	0.67	94	0.99
59	0.14	77	0.75	95	0.99
60	0.16	78	0.76		

#### **5.4 VERBAL TEST ITEMS.**

Labeling questions as difficult does not imply that they should be removed from the test. They help to differentiate between those who know the material well and those who do not. Rather, each question should be examined individually to determine if screeners could benefit from it being reworded. Changing too many questions might also adversely effect the internal validity of the test. In particular, questions asking for a negative response should not be removed or totally reworded but that the negative word(s) (e.g., not, cannot, false, and never) should be emphasized. For example, "Which of the following is NOT true?"

The majority of the 55 questions, showing an adverse impact on the performance of non-native English speakers, were judged to have been answered incorrectly due to a lack of knowledge rather than to poor item construction. It is assumed that non-native English status resulted in reduced learning during training. Non-native English speakers might therefore benefit from additional training and re-testing in order to achieve the same proficiency as native English

speakers on the specific details of procedures and threat characteristics. Ten questions have been identified, showing adverse impact towards non-native English speakers, that are likely to benefit from rewording of either the question or answer choices.

## **6 CONCLUSIONS.**

The SRT item sets have a sufficient range of difficulty to assure score distributions that were approximately normal with reasonable variance. Item difficulty was relatively uniform across different subject areas. Item intercorrelations were sufficient to produce test modules of moderate length with good internal validity.

A composite scoring method combined verbal and image test performance into a single score with a range of 0 to 100 points. When this scoring method was applied to the current test data, it was found that English language status was the most influential demographic factor on overall test scores. When English language status was held constant across groups, there were no significant racial or gender differences in performance as measured by the composite score. It is important to note that English proficiency is a legal requirement for the position.

The test instructions and interfaces were adequate for the large majority of screeners taking the test, suggesting that the fielded test can be successfully self-administered and navigated. There were no differences in performance with the Rapiscan and EG&G image sets nor for the matched full-color and B&W image sets.

Based on analyses of the current SRT data, the test to be fielded as a determination of whether screeners have been adequately trained, will be composed of 40 verbal items (for a maximum of 45 seconds each) and 50 image items (for a maximum of 15 seconds each). This gives a maximum total test time of 42.5 minutes minus instructions. This particular test structure should yield an extremely reliable ( $\alpha = .91$ ) test of the knowledge of newly trained hires. Most importantly, the proposed composite scoring system does not result in an adverse impact on racial and gender groups.

## **7 REFERENCES.**

- [1] Air Carrier Standard Security Program (1976, through change #54). Washington, D.C., Federal Aviation Administration.
- [2] Klock, B. K. & Fobes, J. L. (1999). *Test and Evaluation Plan for Determining Screener Training Effectiveness* (DOT/FAA/AR-99/42). Atlantic City International Airport, NJ: DOT/FAA William J. Hughes Technical Center.
- [3] Fobes, J. L. & Neiderman, E. C. (1999). *Test and Evaluation Plan for Recreener Readiness Test Validation* (DOT/FAA/AR-99/21). Atlantic City International Airport, NJ: DOT/FAA William J. Hughes Technical Center.
- [4] Fobes, J. L. & Neiderman, E. C. (1999). *Validating the Computer-based Training Process for Aviation Security Screeners* (DOT/FAA/AR-99/40). Atlantic City International Airport, NJ: DOT/FAA William J. Hughes Technical Center.

[5] Fobes, J. L., Neiderman, E. C. & Klock, B. A. (1999). *Screening Readiness Test items* (DOT/FAA/AR-99/1). Atlantic City International Airport, NJ: DOT/FAA William J. Hughes Technical Center.

[6] Neiderman, Eric C., Ph.D., & Fobes, J. L., Ph.D. (1999). *Screening Readiness Test - Validation Pilot Testing* (DOT/FAA/AR-99/37). Atlantic City International Airport, NJ: DOT/FAA William J. Hughes Technical Center.

[7] Nunnally, J. C. (1978). *Psychometric testing*. New York: McGraw Hill.

[8] Neter, J., Wasserman, W., & Kutner, M. (1985). *Applied linear statistical models*, second edition. Homedale, IL: Irwin Inc.