# Revised Test and Evaluation Plan for Determining Screener Training Effectiveness

Brenda A. Klock
Joshua Rubinstein, Ph.D.

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

August 2000

Revised Test and Evaluation Plan (Rev. 5)

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 available to the public through the National Technical Information Service (NTIS), Springfield, 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. This document does not constitute FAA certification policy.

PB2001-100122

| 1. Report No.<br>DOT/FAA/AR-00/53 | 2. | 3. Recipient's Catalog No. | |
|---|---|---|---|
| 4. Title and Subtitle<br><br>Revised Test and Evaluation Plan for Determining Screener Training Effectiveness | | 5. Report Date<br>August 18, 2000 | |
| | | 6. Performing Organization Code<br>AAR-510 | |
| 7. Author(s)<br>Brenda A. Klock and Joshua Rubinstein, Ph.D. | | 8. Performing Organization Report No.<br>DOT/FAA/AR-00/53 | |
| 9. Performing Organization Name and Address<br>Federal Aviation Administration<br>William J. Hughes Technical Center<br>Atlantic City International Airport, NJ 08405 | | 10. Work Unit No. (TRAIS) | |
| | | 11. Contract or Grant No. | |
| 12. Sponsoring Agency Name and Address<br>Federal Aviation Administration<br>Associate Administrator for Civil Aviation Security, ACS-1<br>800 Independence Ave., S.W.<br>Washington, DC 20591 | | 13. Type of Report and Period Covered<br><br>Revised Test & Evaluation Plan (Rev.5) | |
| | | 14. Sponsoring Agency Code<br>ACS-1 | |
| 15. Supplementary Notes<br>This draft test and evaluation plan was prepared by Joshua Rubinstein, Ph.D., Federal Data Corporation | | | |

16. Abstract
The efficacy of Computer-Based Training (CBT) programs, potentially useful for security checkpoint screener training, will be evaluated at three different airports. Candidates will be trained with one of three CBT programs or the Air Transport Association-approved classroom training program. The Screener Readiness Test, designed to assess screening-related knowledge, will then be used to evaluate the effectiveness of these programs.

| 17. Key Words<br>Computer-Based Training<br>Screener Readiness Test<br>Aviation Security | | 18. Distribution Statement<br>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<br><br>This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161 | |
|---|---|---|---|
| 19. Security Classif. (of this report)<br>Unclassified | 20. Security Classif. (of this page)<br>Unclassified | 21. No. of Pages<br>15 | 22. Price |

**Form DOT F 1700.7 (8-72)**     **Reproduction of completed page authorized**

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# EXECUTIVE SUMMARY

In April 1997, the Federal Aviation Administration (FAA) approved the use of a Computer-Based Training (CBT) program for initial screener training prior to on-the-job training. Because the variety of training options is growing, the FAA has developed a single uniform measure of mastery of initial training, the Screener Readiness Test (SRT).

The SRT has utility to discriminate between alternative training programs. That is, using the SRT as a standard post-training measure of effectiveness, the SRT will highlight criterion-based differences between the different screener-training programs. This test and evaluation plan describes a comparison of four training programs: three CBT programs and the Air Transport Association (ATA)-approved classroom-training program, using performance on the SRT as a measure of training efficacy.

One hundred ninety two security screener candidates who complete the various training programs will participate in this study. Some candidates will receive an SRT pretest followed by CBT or ATA-classroom training and an SRT posttest. This design will provide data on the absolute effectiveness of each program (i.e., how much effect did each training program have on SRT performance). The remaining candidates will not receive a pretest, they will only receive CBT or ATA-classroom training followed by an SRT posttest. This design will provide data on the relative effects of training (i.e., what are the relative differences in SRT performance following training).

In addition, an overall success measure will be calculated for each training program, taking into account both the percentage of screener candidates who complete training and the percentage of screener candidates who attain a minimum SRT cutoff score. This will provide a metric for evaluating training efficiency (as measured by the program dropout rate) and training effectiveness (as measured by the SRT performance) for each screener-training program.

## ACRONYMS

| | |
|---|---|
| ACSSP | Air Carrier Standard Security Program |
| ATA | Air Transport Association |
| ATL | Atlanta William B. Hartsfield International Airport |
| CBT | Computer-Based Training |
| COIC | Critical Operational Issues and Criteria |
| DTW | Detroit Metropolitan Wayne County International Airport |
| FAA | Federal Aviation Administration |
| HFE | Human Factors Engineer |
| ICTS | International Consultants on Targeted Security |
| MOP | Measure of Performance |
| OJT | On-the-Job Training |
| OSM | Overall Success Measure |
| PC | Percentage Complete |
| PP | Percentage Pass |
| SEA | Seattle-Tacoma International Airport |
| SME | Subject-Matter Expert |
| SRT | Screener Readiness Test |

# 1. INTRODUCTION

The effectiveness of the national civil aviation security system is highly dependent upon the people who are employed as checkpoint screeners. The training of these individuals is critical to their performance on the job. The Federal Aviation Administration (FAA) is very interested in enhancing screener training and further improving their readiness for the job.

## 1.1 Background

According to Federal Aviation Regulations § 108.17 (Use of X-ray systems), there shall be a program for initial and recurrent training of operators of X-ray systems that includes 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 who perform screening and security functions. For many years, the only FAA-approved training was that developed by the Air Transport Association (ATA). This 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) program for initial screener training prior to OJT. This training is also based on Section XIII of the ACSSP. There are other training programs currently being developed for initial screener training.

As additional programs are offered for initial screener training, each is expected to include a different test to assess mastery prior to OJT and screener certification. Because the variety of training options is growing, the FAA has developed a single uniform measure of mastery of initial training, the Screener Readiness Test (SRT) [1,3]. This preparedness evaluation also contains multiple-choice questions on the major checkpoint screening tasks of walk-through and hand-held metal detectors, pat downs, hand searches, X-ray operation, trace detector operation, and monitoring the exit lane. The SRT contains X-ray images to be resolved for threat articles such as improvised explosive devices, the FAA modular bomb set, hand grenades, guns, and knives.

The SRT has been validated for its content by subject matter experts (SMEs) and field assessments. Initially, FAA SMEs familiar with both initial X-ray screener training content and the ACSSP reviewed all test items. FAA SMEs familiar with IEDs evaluated the image items to ensure that they represented a full range of threats and difficulty levels. These reviews amounted to SME validation of the SRT content. Subsequent item analyses were also conducted to evaluate the test construction, usability, and question reliability.

In addition, the SRT was field validated using certified airport screeners. An initial assessment of face validity and usability of the test was conducted with all 18 screeners at the Atlantic City International Airport. Further, the test was administered to more than 349 certified airport screeners at multiple US airports. The two field validations demonstrated content validity of the test.

As a valid measure of the mastery of initial screener training, the SRT has the utility to discriminate between alternative training programs. That is, using the SRT as a standard post-

training measure of effectiveness, the SRT will highlight criterion-based differences between the different training programs. In this way, a common standard of comparison (i.e., the SRT) provides an objective aid in identifying the training program likely to yield the most desirable training outcomes.

## 1.2 Purpose

This test and evaluation plan describes the evaluation of four different training programs using performance on the SRT as a measure of training efficacy. It describes the overall training-program examination strategy and validation criteria to be used in evaluating three candidate CBT programs and the ATA-classroom training program. The primary measure of training effectiveness for this analysis will be screener candidate performance on the content questions and image tests of the SRT following training. The programs will be evaluated at Detroit Metropolitan Wayne County International Airport (DTW), Atlanta William B. Hartsfield International Airport (ATL), and Seattle-Tacoma International Airport (SEA).

## 2. METHODOLOGY

### 2.1 Test Milestones

Table 1 shows the milestones for planning and reporting the project.

### TABLE 1. MILESTONES

| MILESTONE | DATE | RESPONSIBLE ORGANIZATION |
|---|---|---|
| Project Plan | April 9, 1999 | FAA |
| Revised Test and Evaluation Plan | September 15, 2000 | FAA |
| Start Test and Evaluation | Oct 1, 2000 – Feb 1, 2001 | FAA |
| Data Analysis | Feb 1 – Mar 15, 2001 | FAA |
| Preliminary Test and Evaluation Report | Mar 15 – May 15, 2001 | FAA |
| Draft Test and Evaluation Report | July 15, 2001 | FAA |

### 2.2 Test Sites and system description

All three sites, DTW, ATL, and SEA, will be involved in all phases of this study. Additional sites may be added if screener-candidate availability is too low. At an initial visit to each site the SRT will be installed on designated computers at the airport.

2

The three CBT programs to be evaluated in this study are from International Consultants on Targeted Security (ICTS), SafePassage International, Ltd., and Smart Approach, Ltd. Representatives of each CBT manufacturer will install their CBT programs at each test site. The ATA-approved classroom training program will also be evaluated in this study.

The SRT has undergone a human factors evaluation and redesign to minimize any computer related operator complexities. Consequently, no computer skills are required to take the test. Clear instructions and examples lead the test taker through each section of the SRT. Responses are made by pressing one of the number keys. The computer mouse is never used during the test.

## 2.3 Participants

One hundred ninety two security screener candidates who complete the various training programs will participate in this study. Because some percentage of screener candidates who begin training do not complete training, it is assumed that the number of screener candidates who initially participate in this study will be somewhat greater.

## 2.4 Procedures

This test and evaluation will be conducted with screener candidates serving as subjects provided by different contracted security companies. Thus, the logistics of the testing will require careful coordination and communication between airport security personnel, security company personnel, FAA security personnel, and HFEs. An HFE will be present at each site to administer the SRT and implement the experimental protocol for this study, instruct the screener candidate and answer any questions pertaining to taking the SRT. Initial data collection at each site will be contingent on the availability of screener candidates and will continue until 16 screener candidates for each of the four training programs (i.e., 64 screener candidates per airport) have completed the experimental protocol. Candidates will be randomly assigned to one of the four training programs. If a candidate discontinues training, the HFE will simply assign the next available candidate to the now vacant subject slot. This will insure that no training program lags behind in terms of completing the study due to a higher dropout rate. This will also provide a written record of training-program dropouts for later analysis. The SRT will be used as the measure of screener knowledge and candidates will be tested with the SRT following completion of training.

Data collection will continue until all 192 candidates across the four training programs have taken the SRT. Because data collection is contingent on candidate availability, a strict procedural timeline cannot be specified.

A mixed within-subjects/between-subjects test design will be used to mitigate the effects of individual differences while adequately addressing all critical issues. In its present form, the SRT contains a large number of items. It is possible to test candidates before and after training without repeating many items. Such an approach provides a sensitive design in which individual differences in knowledge, skills, and ability before training may be controlled. Other candidates

will not take the SRT before training, allowing the determination of whether the pre-training test affected post-training performance.

Screener candidates will be randomly assigned to a training program. For each program at each site, (see figure 1) eight of the 16 candidates will be randomly assigned to the within-subjects comparison. The remaining eight candidates per training program per site will be assigned to the between-subjects comparison.
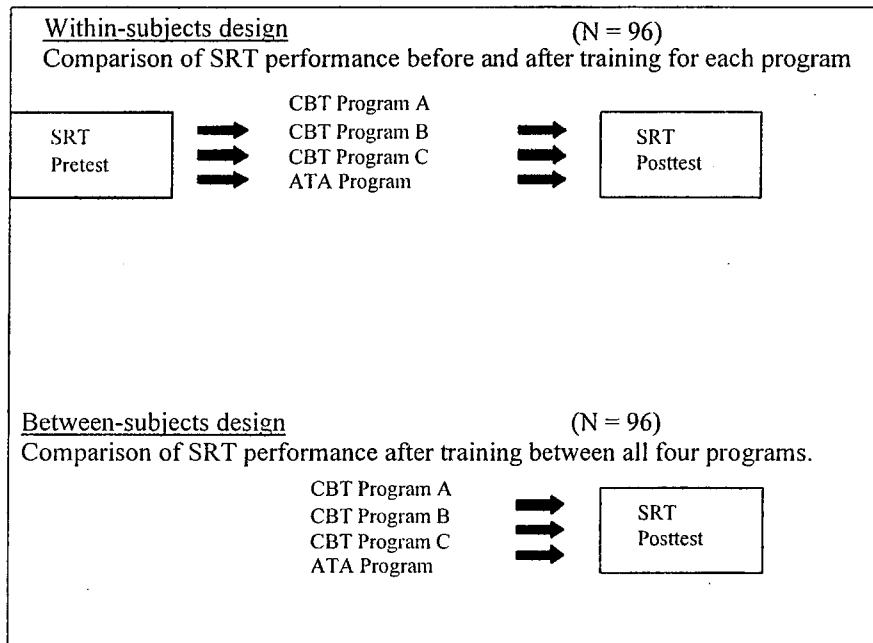


FIGURE 1. EXPERIMENTAL DESIGN FOR THE ANALYSIS OF THE FOUR TRAINING PROGRAMS

2.4.1        Within-Subjects Comparison

Eight candidates for each training program from each airport (n = 96) will participate in the within-subjects comparison. Each of these candidates will be given an SRT pretest. Following the pretest, each candidate will begin training. Upon completion of training, each candidate will be given an SRT posttest. If a candidate does not complete training or does not take the SRT posttest, that candidate will be replaced.

2.4.2        Between-Subjects Comparison

The remaining eight candidates for each training program from each airport (n = 96) will be used for the between-subjects comparison. This posttest-only comparison will provide an evaluation of the training programs while avoiding possible unexpected influences of having taken the pretest. Each of these candidates will be trained by one of the four programs. Upon completion

of training, each candidate will be given an SRT posttest. If a candidate does not complete training or does not take the SRT, that candidate will be replaced.

## 2.5    Data Collection

Data for the SRT will be collected and stored in a computerized database. The SRT will automatically store responses for all content and image questions for each screener candidate who takes the test. The database will contain screener candidate background information, CBT or ATA-classroom program identification, and SRT scores for each section. The data for each candidate will be transferred into electronic media and stored in an Excel 5.0 database after testing is completed at each airport site. When data collection is complete, the HFE will collect the data and debrief the security trainers.

## 2.6    Data Analysis

Descriptive and inferential statistics will be calculated for all data. Mixed-design analyses of variance will be used to evaluate the SRT differences between the training programs and test sites. Additional analyses will be conducted on the different types (i.e., gun, knife, MBS, grenade, IED) of Image Questions (five levels).

### 2.6.1    Within-Subjects Comparison

Test Order (pretest, posttest) will serve as the within-subjects independent variable. Training Program (four levels) and Test Site (three levels) will serve as the two between-subjects independent variables for this comparison. This results in a 2 x 4 x 3 mixed within-and between-subjects experimental design.

The dependent variable will be the SRT scores. A repeated-measures analysis of variance will be performed on these data. A main effect of the Training Program will be used to assess overall differences in SRT performance between the four training programs. The interaction between Test Order and Training Program will address whether the training programs produce differential learning effects as measured by the SRT. Post hoc analyses will be used to determine if all four programs produce significant learning effects and to rank order the different programs according to the magnitude of their training efficacy.

### 2.6.2    Between-Subjects Comparison

Training Program (four levels) and Test Site (three levels) will serve as the two between-subjects independent variables for this comparison. This results in a 4 x 3 between-subjects experimental design.

The dependent variable will be the SRT scores. An analysis of variance will be performed on these data. A main effect of Training Program will serve to evaluate the overall differences in SRT performance between the four different training programs. Post-hoc analyses will be used to rank order the different programs according to the magnitude of SRT proficiency.

5

An additional analysis will be performed comparing the within-subjects posttest scores to the between-subjects posttest scores. This analysis will determine if screener candidate training was influenced by the SRT pretest.

## 2.6.3 A Cutoff Score for the Screener Readiness Test

SRT data are being collected from over 100 newly trained checkpoint screeners in a separate study [2] and will be used to establish an SRT cutoff score. The range of scores will be used to establish this cutoff score as a measure of minimum acceptable performance. The SRT data used to evaluate the various training programs will be evaluated with respect to this cutoff score. Each training program will be evaluated according to the number of its candidates who exceed this score.

## 2.6.4 Item Analyses

For the SRT content questions, a descriptive distribution of the accuracy rate for all questions will be established. The specific knowledge for questions with high accuracy rates and high error rates will be identified.

## 2.6.5 Additional Analyses

A high dropout rate for a given training program could cause a sampling bias for candidates from that program who finally take the SRT. It could also indicate inefficient training materials and/or techniques. For these reasons, an Overall Success Measure (OSM) will be calculated for each training program taking into account both the percentage of candidates who complete training and the percentage of candidates who pass the SRT cutoff score. This overall performance measure for each training program will be

$$OVERALL\ SUCCESS = PC \times PP$$

where PC equals the percentage of candidates that complete training and PP equals the percentage of candidates that pass the SRT cutoff score. A low score on this measure, signifying poor training performance, could be due to either a low completion rate or a low SRT pass rate.

## 2.7 Critical Operational Issues and Criteria

The Critical Operational Issues and Criteria (COIC) are those necessary to evaluate the training programs. The strategies for evaluating these COICs and their associated Measures Of Performance (MOPs) are discussed in the following subsections.

## 2.7.1 Issue 1 - Absolute Training Effectiveness

Do screeners acquire sufficient knowledge with each training program to progress to OJT?

Criterion 1-1. Investigative in nature.

MOP 1-1-1. SRT pretest and posttest scores associated with each training program.

MOP 1-1-2. The difference between pre-training and post-training SRT scores for each individual training program.

MOP 1-1-3. The percentage of screener candidates who successfully complete each training program.

MOP 1-1-4. The percentage of screener candidates who exceed a minimum SRT cutoff score.

MOP 1-1-5. A training-program Overall Success Measure (OSM) derived as the product of the percentage of screener candidates who complete training and the percentage of screener candidates who pass the minimum SRT cutoff score.

## 2.7.2 Issue 2 - Relative Training Effectiveness

Do the training programs differ in their training effectiveness?

Criterion 2-1. Investigative in nature.

MOP 2-1-1. Post-training differences between training programs on the SRT score.

MOP 2-1-2. Differences between training programs on the SRT pretest/posttest difference score.

MOP 2-1-3. Post-training differences between training programs on the SRT content and image questions.

MOP 2-1-4. Differences between training programs on the SRT pretest/posttest difference scores on content and image questions.

## 2.8 Limitations

A potential limitation of the results interpretation involves a possible sampling bias that might be introduced by the training programs themselves. Candidates will be randomly assigned to the training programs; however, unequal dropout rates could create a bias for one or more of these programs. For example, one training program might be harder to complete, resulting in a higher dropout rate. The candidates who do complete the training and subsequently take the SRT might, consequently, have a greater overall aptitude level. The resulting SRT performance showing higher test scores would suggest a better quality of screener-readiness training for that particular program. In reality, however, the differences in SRT performance would be due to a stricter selection criterion, which would eliminate the less able screener candidates prior to