PB99-145054

DOT/FAA/AR-99/23

Office of Aviation Research
Washington, D.C. 20591

Evaluation of the Screener Aptitude Test Battery Items

Eric C. Neiderman, Ph.D.
J. L. Fobes, Ph.D.

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

February, 1999

Final Report

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

REPRODUCED BY
U.S. DEPARTMENT OF COMMERCE
NATIONAL TECHNICAL
INFORMATION SERVICE
SPRINGFIELD, VA 22161

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 the information is essential to the objective of this report.

Technical Report Documentation Page

1. Report No.
DOT/FAA/AR-99/23

2

PB99-145054

3. Recipient's Catalog No.

4. Title and Subtitle

Evaluation of the Screener Aptitude Battery Test Items

5. Report Date
February, 1999

6. Performing Organization Code AAR-510

7. Author(s)

Eric C. Neiderman, Ph.D.
J. L. Fobes, Ph.D.

8. Performing Organization Report No.

9. Performing Organization Name and Address

U.S. Department of Transportation, Federal Aviation Administration
William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

12. Sponsoring Agency Name and Address

U.S. Department of Transportation, Federal Aviation Administration
Associate Administrator of Civil Aviation Security, ACS-1
800 Independence Ave., S.W.
Washington, DC 20590

13. Type of Report and Period Covered

14. Sponsoring Agency Code
ACS-1

15. Supplementary Notes: Draft prepared by:
William Maguire, Ph.D. & Michael Snyder
Federal Data Corporation
Science and Engineering Division
500 Scarborough Drive
Egg Harbor Township, NJ 08234

16. Abstract

Six perceptual and cognitive tests were fielded at 18 major US airports to develop a screener aptitude test to reliably, validly, and fairly predict future performance of checkpoint security screener candidates. Performance measures included both knowledge acquisition and training transfer to the field. Some tests correlated with computer-based training performance or threat image projection detection data (Hidden Figures Test, Hidden Patterns Test, Spatial Relations Test) while others did not. Some of these three tests showed adverse impact for minorities and women (Hidden Figures Test, Spatial Relations Test). While reliability, validity, and fairness concerns were associated with individual tests at this stage, their limitations may be overcome by a test battery constructed from revisions of the particular tests found to predict learning and training transfer.

17. Key Words
Screener Aptitude Battery (SAB), Hidden Figures Test, Hidden Patterns Test, Spatial Memory Test, Spatial Relations Test, Visual Closure Test, Visual Discernment Test

18. Distribution Statement
This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 33	22. Price

Reproduction of completed page authorized

PROTECTED UNDER INTERNATIONAL COPYRIGHT
ALL RIGHTS RESERVED.
NATIONAL TECHNICAL INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE

CONTENTS

CONTENTS (Continued)

Illustrations

Figures **Page**

Tables **Page**

ACRONYMS

ATL	Atlanta Hartsfield International Airport
c	Operator's Response Threshold
CBT	Computer-Based Training
d'	D prime sensitivity measure
FAA	Federal Aviation Administration
FAA-VST	Federal Aviation Administration Visual Skills Test
HFT	Hidden Figures Test
HPT	Hidden Patterns Test
K-S	Kolmogorov-Smirnov
M	Mean
MCO	Orlando International Airport
SAB	Screener Aptitude Battery
SD	Standard Deviation
TIP	Threat Image Projection

1. INTRODUCTION

1.1 Background

Airline baggage screeners perform a critical role in assuring passenger and aircraft safety. In fact, the security of the whole National Airspace System depends on the perceptual and cognitive skills of the screener. Attrition of these individuals is costly to security companies and means they must spend additional money for recruitment, screening, testing, selecting, hiring, and training new candidates. Companies could benefit greatly by being able to identify individuals who have aptitude for threat detection (i.e., detection of guns, knives, bombs, and hazardous materials). Nevertheless, there are no validated tests available to identify people who may have special aptitude for security screening before training.

Because screeners are directly involved with the safety of airline passengers and equipment, it is crucial that only the most competent individuals are selected for training. The first step in selecting potential screeners is to develop a selection test, or battery of tests, that can identify such individuals. The Federal Aviation Administration (FAA) has undertaken a project to develop a Screener Aptitude Battery (SAB) to predict screeners' performance with Computer-Based Training (CBT) and their threat detection performance as measured by the Threat Image Projection (TIP). The SAB will be unbiased, fair, valid, reliable, and easy to administer. Furthermore, it will be easily transportable from one computer platform to another to ensure maximum usage.

1.2 Scope

The FAA has developed and fielded a number of potential selection instruments. Screeners being trained on Safe Passage's CBT system are first given one of a series of selection tests, on a random basis, before training begins. This report presents the analysis of the selection test data collected to date, as well as, an evaluation of the suitability of each of the tests as a selection instrument.

1.3 Description of the Tests Fielded

Six tests were evaluated (see Table 1). The Hidden Figures Test (HFT) and the Hidden Patterns Test (HPT) are fielded in two alternate forms (A and B). They are designed to measure a variety of perceptual and cognitive abilities and are presently deployed at 18 airports. These tests have been fielded at some airports since November 1997. Before a candidate screener begins any training, they are given either the HFT (form A or B), the HPT (form A or B), or the FAA's Visual Skills Tests (FAA-VST) (tests 3a, 3b, 3c, and 3d). Included with each test or set of tests is an introduction to using a computer mouse, mouse practice, test instructions, and sample test problems. In addition, the candidate is asked to supply demographic data including age, gender, ethnic background, and education.

The HFT and the HPT are described in Fobes and Neiderman (1997) and Neiderman (1997). The FAA-VST consists of four short tests; the Spatial Memory Test, the Spatial Relations Test, the Visual Closure Test, and the Visual Discernment Test.

Table 1. Perceptual/Cognitive Tests Currently Under Investigation

	Test	**Type of Test**	**Approx. Length of Test** (including instructions)
1.	Hidden Figures Test (Form A and B)	Power/Speed	14 min.
2.	Hidden Patterns Test (Form A and B)	Speed	5 min.
3.	FAA Visual Skills Test (FAA-VST)		
3a.	Spatial Memory Test	Power	4 min.
3b.	Spatial Relations Test	Power	4 min.
3c.	Visual Closure Test	Power	4 min.
3d.	Visual Discernment Test	Power	4 min.

Each of the visual skills tests is described below and they all contained ten items. In each of the tests, a moving time bar displays the amount of test time remaining.

1.3.1 Spatial Memory Test

Individuals are presented a four-by-four pattern of squares with two or three boxes blackened (see Figure 1). The individual is given three seconds to study the pattern before it disappears. Shortly thereafter, they are presented four possible choices and have 10 seconds to select the exact pattern that was presented earlier (see Figure 2).

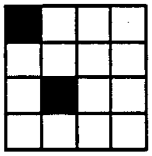

Figure 1. Spatial Memory Test Item Sample

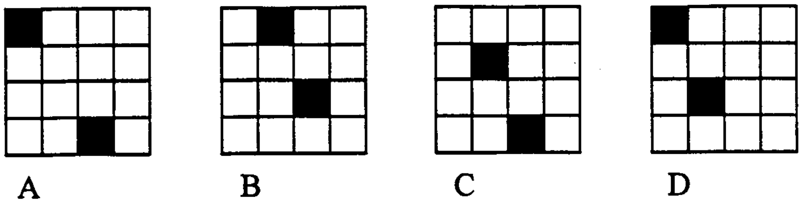

Figure 2. Spatial Memory Test Answer Sample. Pattern "D" is the correct answer

1.3.2 Spatial Relations Test

Individuals are presented a four-by-four pattern with three blackened squares and the upper-left corner is cut out (Figure 3). Directly below the target pattern are four alternatives. Individuals are to select the pattern that matches the target pattern but has been rotated counter-clockwise one quarter turn (see Figure 4). Individuals have 10 seconds to respond.

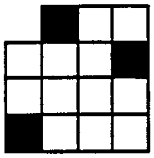

Figure 3. Spatial Relations Test Item Sample

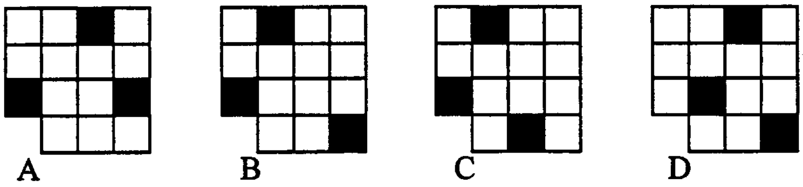

Figure 4. Spatial Relations Test Answer Sample. Pattern "B" is the correct answer.

1.3.3 Visual Closure Test

Each screen contains a simple pattern such as a circle, square, or some other geometric shape. The target pattern contains three to five lines, as shown in Figure 5. Note that the spacing between lines is not equal. Below the target pattern is a row of four alternatives 'broken' with gaps in their lines (Figure 6). Only one of the alternatives matches the target pattern. Both the target pattern and correct alternative contain the same number of lines and spacing; the only difference is that the lines in the alternative pattern are incomplete. In this test, individuals have 15 seconds to respond.

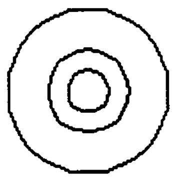

Figure 5. Visual Closure Test Item Sample

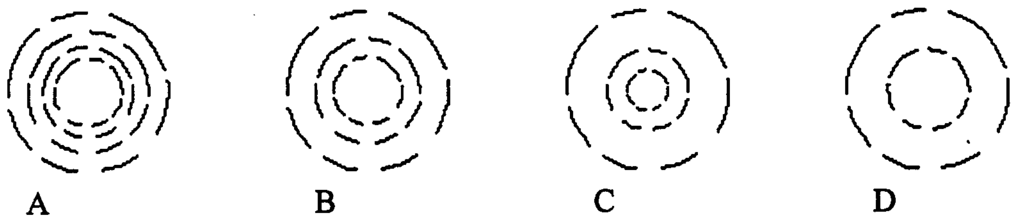

Figure 6. Visual Closure Test Answer Sample. Pattern "C" is the correct answer.

1.3.4 Visual Discernment Test

Each screen contains a simple figure (Figure 7) and a row of four complex patterns (see Figure 8). The complex patterns are made of many lines and shapes with some shapes light while others are dark. One of the alternatives contains the test shape; however, the test shape can be smaller than is shown in the sample pattern. It can also be different in terms of white or black. It is the individual's task to indicate which of the four alternatives contains the test shape, ignoring its size and brightness. Individuals have 10 seconds to respond.

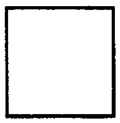

Figure 7. Visual Discernment Test Item Sample

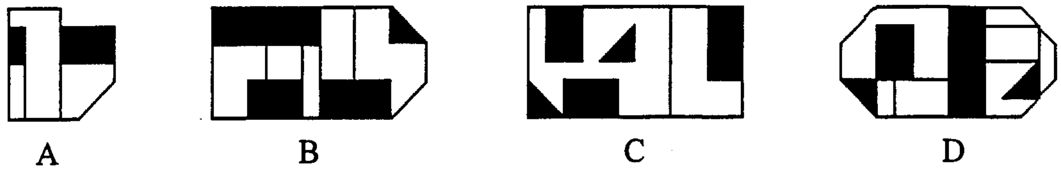

Figure 8. Visual Discernment Test Answer Sample. Pattern "B" is the correct answer.

1.3.5 Test Scoring

The scoring system recommended by the Educational Testing Service, holder of the copyright on the HFT and HPT, was followed. For the HFT, one point was awarded for each correct answer and no points were awarded for each incorrect answer. For this test, ETS does not correct scores for guessing. However, this is not the case for the HFT where multiple choice answers had five alternatives. An individual accordingly has a one-in-five chance (20%) of *guessing* the correct answer. Every correct answer was awarded one point and, for every incorrect response, a fourth of a point (1/4) was subtracted.

Each of the Visual Skills Tests was scored using the same scoring scheme as for the HFT. Multiple choice answers on Spatial Memory, Spatial Relations, Visual Closure, and Visual Discernment tests had four alternatives. Therefore, a third of a point (1/3) was deducted for each incorrect response and full point was awarded for each correct answer.

1.4 Item Analysis Strategy

The detailed strategy for the item analysis is described in Fobes and Neiderman (1998). A brief summary is provided here.

The analysis of each test consisted of the following procedures.

1. A Kolmogorov-Smirnov (K-S) significance test was calculated on the test scores to determine if the distribution of scores for each test is skewed.
2. Item difficulty was measured by the percentage who answered correctly on each item.
3. Item discrimination was measured by a corrected correlation (i.e., influence of the item score on the overall score was removed) of the item and the overall test. The item values were not binary (right or wrong), but rather triple (right, wrong, or skipped). The weights assigned to items in scoring were used to calculate the correlations.
4. The measurement of test reliability differed from test to test. The FAA-VST tests are power / speed tests. Reliability was measured by calculating coefficient alpha (r_{nn}) of each test which is equivalent to average split-half reliability. The HFT is a hybrid speed/power test and the HPT is basically a speed test (Fobes & Neiderman, 1997). Estimating reliability for these two tests is more difficult. The specific approach used is presented with the results for these particular tests.

5. The criterion-related validity of each selection test was measured against both CBT and TIP performance. CBT performance was measured by classifying trainees into those who had successfully completed the CBT (i.e., pass CBT) and those who had not completed the CBT (fail CBT). Selection tests scores of these two groups were compared to determine if they differed. Scores on the final exams were not used because a significant fraction of screeners failed to complete the CBT, causing significant restriction of range in those measures.

6. TIP performance was measured by hit rates, false alarm rates, screener sensitivity (d′), and screener criterion (c) (refer to See, Warm, Dember, & Howe, 1995 for a discussion of response bias measures). Selection test scores were correlated with these measures.

7. The fairness of each of the tests was analyzed by statistically assessing adverse impact on minority groups and women. This was assessed by a two-way factorial analysis of variance on test scores using gender and racial grouping, identified from the Personal Information Form, as factors. A more detailed analysis is provided for tests that showed potential adverse impact.

1.5 Characteristics of the Selection Test Sample

With data collected through June 1998, the screener selection database contained 1696 records of screeners who had 'taken' a selection test. From that group, a significant portion (758) did not answer any of the questions. The database was examined in order to determine the characteristics of this non-responsive group. The most salient characteristic of non-responsive records was the airport site. Of 323 records collected from Atlanta Hartsfield International Airport (ATL), only 8 were unresponsive. Of 199 records collected from Orlando International Airport (MCO), 185 were non-responsive. Airports with a very high (> 80%) unresponsiveness included Detroit Metropolitan Wayne County Airport, Honolulu International Airport, MCO, and San Francisco International Airport. It is very likely that at these sites trainers instructed screeners to click right through the test. Despite the significant loss of data, it does not represent any form of self-selection from screeners and, therefore, is not a critical problem.

Data were plentiful for all tests and analyses are based upon the numbers given in Table 2. The third column provides the number of trainees who supplied demographic information in addition to test data.

If an individual did not answer any questions for a given test, then that test data was not included. The four Visual Skills Sub-tests were given as a battery. If an individual provided answers for some sub-test(s) but not for others, the sub-test taken was included and the one(s) skipped was not included. This is the reason that the N for each of the Visual Skills Sub-tests slightly differs as shown in Table 2.

Table 2. Number of Participants and Demographic Information for Each Test

Test	Number of Participants	Number Providing Demographics
Hidden Figures Test 1	181	172
Hidden Figures Test 2	145	144
Hidden Pattern Test 1	195	187
Hidden Pattern Test 2	176	172
Spatial Memory Test	199	189
Spatial Relations Test	206	196
Visual Closure Test	221	210
Visual Discernment Test	213	199

1.6 Characteristics of the Validity Samples

CBT data and accompanying selection test data were obtained for 204 screeners from Dallas/Fort Worth International Airport, ATL, and Seattle-Tacoma International Airport. These data were collected between July 1998 and November 1998. TIP data and accompanying selection data were obtained for 111 screeners. Most of these had worked at ATL during the same period of time, with a small number from John F. Kennedy International Airport.

2. RESULTS

For each selection test, data were first analyzed to see whether the distribution of scores was skewed. For this calculation, the K-S test was used. Table 3 provides the Mean (M), Standard Deviation (SD), and results of the K-S test for normality for all score distributions.

Table 3. Mean Score, Standard Deviation, and Normality Results for Each of the Tests

Test	Mean Score	Standard Deviation	K-S Z Value	Probability Value
Hidden Figures Test 1	1.24	2.28	1.36	.049
Hidden Figures Test 2	0.74	2.08	1.47	.026
Hidden Pattern Test 1	24.20	22.80	1.75	.004
Hidden Pattern Test 2	26.90	22.20	1.38	.043
Spatial Memory Test	6.69	3.25	2.39	.0001
Spatial Relations Test	2.39	3.06	2.06	.0001
Visual Closure Test	6.77	2.61	2.18	.0001
Visual Discernment Test	7.32	2.31	1.84	.002

Note: All probability values significant at the .05 level or greater.

2.1 Hidden Figures Test

2.1.1 Test Scores and Times

The distribution of scores for the HFT-1 was not normal (K-S $z = 1.36, p = .049$). Similarly, the distribution of the scores for the HFT-2 was not normal (K-S $z = 1.47, p = .026$). In both cases, the distribution is skewed to the right.

The HFT-1 and the HFT-2 are characterized by low scores and, because the test is time-limited, the majority of screeners do not finish it. In that sense, it has elements of a speed test even though individual items are quite difficult. There were 64 screeners who completed the HFT-1 and 40 screeners who completed the HFT-2 test in the time given. Since screeners often skipped items, the test was considered to be completed if one of the last four items was answered.

Hidden Figures Test 1

The M score for the HFT-1 was 1.24, *SD* = 2.28 and scores ranged from –4.0 to 8.75. The M number correct was 3.17 and the M number incorrect was 7.71. The M time to finish the test was 523 seconds with 40% of the trainees using the full time allotted (720 seconds).

The proportion correct for the HFT-1 ranged from 0.02 to 0.46 (see Table 4). An average rate of 0.20 can be attained by guessing. Only 5 of the 16 items on the HFT-1 were answered at a correct rate exceeding 0.30.

Item discrimination was measured by corrected item-total score correlations. These ranged from -0.09 to 0.42. The item discrimination parameter for each item is given in Table 4.

Table 4. Proportion Correct and Item Discrimination Values for the Hidden Figures Test 1

Item Number	Proportion Correct	Item Discrimination	Proportion Skipped
Item 1	.46	.18	.12
Item 11	.42	-.05	.32
Item 5	.41	-.07	.28
Item 2	.40	.22	12
Item 8	.38	.09	.41
Item 4	.30	.42	.16
Item 6	.28	.19	.20
Item 14	.28	.12	.45
Item 13	.27	.04	.49
Item 7	.23	.22	.32
Item 12	.21	.06	.39
Item 15	.21	-.01	.46
************ Item 10	************ .19	************** -.09	********** .38
Item 3	.18	.17	.20
Item 16	.09	.06	.49
Item 9	.02	.24	.34

Hidden Figures Test 2

The M score for the HFT-2 was 0.74, *SD* = 2.08 with scores ranging from –4.0 to 9.0. The M number correct was 2.64 and the M number incorrect 7.57. The average amount of time o complete the test was 522 seconds and 39% of the trainees took the full amount of time.

For the HFT-2, the proportion correct ranged from 0.08 to 0.48 (see Table 5). Only 4 of the 16 items on the HFT-2 are answered at a correct rate exceeding 0.30.

Item discrimination was measured by corrected item-total score correlations and these correlations ranged from -0.17 to 0.35. The item discrimination parameter for each item is given in Table 5.

Table 5. Proportion Correct and Item Discrimination Values for the Hidden Figures Test 2

Item Number	Proportion Correct	Item Discrimination	Proportion Skipped
Item 13	.48	-.02	.47
Item 10	.40	.05	.41
Item 16	.36	-.17	.60
Item 14	.32	-.01	.50
Item 3	.30	-.01	.25
Item 6	.28	.11	.29
Item 5	.27	.24	.26
Item 1	.26	.35	.11
Item 2	.24	.16	.19
Item 12	.24	.15	.45
Item 9	.22	.20	.38
Item 11	.21	.09	.44
***************	*************	*****************	**************
Item 7	.19	.17	.30
Item 4	.18	.25	.23
Item 8	.17	.06	.35
Item 15	.08	-.06	.55

2.1.2 Reliability of the Hidden Figures Tests

The split-half reliability coefficients were calculated for both tests using separate scores for odd- and even-numbered items. The split-half reliabilities were low. The HFT-1's split-half reliability was 0.44 and the HFT-2's split-half reliability was 0.43.

2.1.3 Adverse Impact of the Hidden Figures Test

In the demographic self-descriptions provided on the Personal Information Form, screeners described themselves as Asian, black, Hispanic, white, or other. Test fairness was analyzed by examining the effect of race and gender on the test scores using a factorial analysis of variance. There was no effect of race or gender for the HPT-1.

There was, however, a significant effect of race on performance for the HFT-2 [*F*(4,134) = 3.31, *p* = .013]. Post testing using Dunnett's C (because the data violated homogeneity of variance assumption and the design was unbalanced) did not show significant adverse impact for any racial group relative to the white group. Mean scores for each demographic group are listed in Table 6.

In addition to race, there was also a significant effect of gender on HFT-2 performance [*F*(1,134) = 4.61, *p* = .034]. Females performed more poorly than did males on this test ($M_{females} = .38$, $M_{males} = 1.21$) (see Table 7). Therefore, there is an adverse impact for females associated with the Hidden Figures Test.

Table 6. Test Scores by Demographic Group

	Mean					**F-Value**	**df**	**Prob-ability**
Test	**Asian**	**Black**	**Hispanic**	**White**	**Other**			
Hidden Figures 1	1.25	1.34	1.10	0.85	1.68	0.24	4, 162	.913
Hidden Figures 2	0.07	0.50	0.98	1.82	2.40	3.31	4, 134	.013*
Hidden Patterns 1	20.61	22.29	28.42	28.81	36.50	1.16	4, 175	.326
Hidden Patterns 2	28.38	29.17	31.69	41.24	27.10	1.88	4, 161	.115
Spatial Memory	6.37	6.96	5.79	7.71	4.22	1.70	4, 180	.152
Spatial Relations	2.54	1.96	1.91	4.25	0.78	3.61	4, 187	.007*
Visual Closure	7.35	6.81	6.48	7.35	3.67	1.56	4, 200	1.19
Visual Discernment	7.30	7.48	8.20	7.27	5.41	0.77	4, 189	.543

Note: * indicates significant at the .05 level or greater.

Table 7. Test Scores by Gender

Test	Mean		F-Value	df	Probability
	Male	Female			
Hidden Figures 1	1.47	1.06	0.42	1, 162	.517
Hidden Figures 2	1.21	0.38	4.61	1, 134	.034*
Hidden Patterns 1	27.8	23.2	0.07	1, 175	.788
Hidden Patterns 2	31.7	24.6	2.35	1, 161	.127
Spatial Memory	6.7	7.0	0.00	1, 180	.963
Spatial Relations	3.0	1.9	5.15	1, 187	.024*
Visual Closure	7.0	6.7	0.43	1, 200	.511
Visual Discernment	7.2	7.6	0.00	1, 189	.982

Note: * indicates significant at the .05 level or greater.

2.1.4 Validity of the Hidden Figures Test

In the CBT validation group, 30 screeners took the HFT-1 and 37 screeners took the HFT-2. Because a number of screeners failed to complete the training, the correlations with final CBT scores involved a smaller sample and were not significant. The criterion measure, which includes the full range of screeners, is the simple categorization of screeners into pass CBT and fail CBT groups. The M HFT-1 score of screeners who failed the CBT was 0.08, while the M HFT-1 score of those who passed it was 1.9 ($r_{pb} = 0.32, p = .03$).

For the HFT-2, the M score of screeners who failed the CBT was 0.34, while the M HFT-2 score of those who passed it was 2.34 ($r_{pb} = .32, p = .03$). As shown in Table 8, M differences between passing and failing groups are significant. It is also worth noting that the point-biserial validity coefficients are really quite large relative to the reliability coefficients in this case.

Using TIP data to look at validation of the selection tests, attention was restricted to those screeners who were exposed to at least 10 TIP images on the job, in order to increase the overall reliability of the TIP criterion measure. Using this criterion, TIP data were available for 14 screeners who took the HFT-1 and 12 screeners who took the HFT-2. Both d' [$r(11) = .54$, p=.07] and hit rate [$r(11) = .67, p$=.02] were related to the HFT-2 but not the HFT-1. On the other hand, there was a significant correlation between HFT-1 scores and c̲ /$r(13), p = .023$]. Note that the relationship between HF2 and TIP accounted for 29% of the variance..

Table 8. Mean Differences in Selection Test Scores
Between Those Who Passed and Failed CBT

Test	N	Mean		*t*-Value	df	Signif.
		Failed	Passed			
Hidden Figures 1	30	0.08	1.90	2.23	28	.034*
Hidden Figures 2	37	0.34	2.34	2.27	35	.029*
Hidden Patterns 1	37	9.00	35.00	3.51	35	.001*
Hidden Patterns 2	44	28.95	26.83	0.32	42	.754
Spatial Memory	47	6.51	7.16	0.66	45	.511
Spatial Relations	46	1.19	3.49	2.99	44	.005*
Visual Closure	48	7.16	7.81	0.97	46	.336
Visual Discernment	48	7.63	7.60	0.04	46	.967
Full Battery	45	23.50	26.20	1.38	43	.176

When the results from the two tests were combined, only hit rate, $r(25) = .56, p < .01$, was found to be related to HFT performance. Table 9 presents the correlations of selection tests scores with TIP data for all of the tests.

Table 9. Correlations of Selection Test Scores With TIP Performance

Test	N	Hit Rate	FA Rate	d'	c
		Correlation	Correlation	Correlation	Correlation
Hidden Figures 1	14	.49	.44	.12	.60 (*p* = .023)
Hidden Figures 2	12	.67 (*p*=.02)	.07	.54 (*p* =.05)	.46
Hidden Patterns 1	22	.27	.18	.17	.25
Hidden Patterns 2	15	.55	-.25	.63 (*p* = .01)	.00
Spatial Memory	16	-.20	.33	-.33	-.05
Spatial Relations	17	.11	-.07	.17	.01
Visual Closure	18	-.40	-.16	-.16	-.24
Visual Discernment	17	-.40	.32	-.44	-.37

For the HFT, there were sufficient data on gender performance, for both the selection tests and CBT, to look more closely at adverse impact and explore issues of test bias. An analysis of both the HFT-1 and the HFT-2 scores was performed with test score as the dependent variable and gender and CBT status (passed or failed) as independent variables. Only the effect of CBT status was significant [$F(1,63) = 8.9, p = .004$]. The difference in scores between females who failed the CBT and those who passed the CBT was 2.5, while the M difference in scores between males who failed the CBT and males who passed the CBT was 1.1. Although there is evidence of adverse impact with this test, there is little evidence of test bias because the interaction was not significant. Also, the differences between Ms did not indicate that the test is less valid as a predictor of female's performance than of male's performance.

2.2 Hidden Patterns Test

2.2.1 Test Scores and Times

Hidden Patterns Test 1

Like the HFT, the HPT was divided into two parts, HPT-1 and HPT-2. The M score for the HPT-1 was 24.2, *SD* = 22.8. Scores ranged from –6.0 to 85.0. The M number correct was 28.2 and the M number incorrect was 3.93. The average amount of time to complete the test was 174 seconds and 95% of the trainees took the full time of 180 seconds. The distribution of the scores is not normal, K-S $z = 1.752$, $p = .004$, but is rightward skewed.

Hidden Patterns Test 2

The M score for the HPT-2 was 26.9, *SD* = 22.2 with scores ranging from –7.0 to 87.0. The average number correct was 30.6, and the average number incorrect was 3.7. The M time to complete the test was 172 seconds with 90% of the trainees taking the full 180 seconds. The distribution of scores was not normal, K-S $z = 1.39$, $p < .043$; it is skewed toward the right.

2.2.2 Reliability for the Hidden Patterns Test

Because the HPT is a speed test, the preferred method used to evaluate reliability is to use equivalent forms. This means the same screeners would need to take alternate forms of the test. Currently, there are not sufficient data available that provide reasonable measures to calculate reliability.

2.2.3 Adverse Impact of the Hidden Patterns Test

For the HPT-1, there was no effect of race or gender. For the HPT-2, the main effects of race and gender again did not significantly differ. However, there was a significant interaction of race and gender for the HPT-2 [$F(4,175) = 3.70$, $p = .007$]. The interaction resulted from the Asian group having a much higher score among females ($M = 38.0$) than among males ($M = 8.2$). Only seven females and six males were in these groups. Scores for each demographic group and gender are listed in Tables 6 and 7, respectively.

2.2.4 Validity for the Hidden Patterns Test

Of the CBT validation group, there were 37 screeners who took the HPT-1 and 44 who took the HPT-2. For the HPT-1, the scores of those who failed the CBT ($M_{\text{failed}} = 9.0$) were significantly lower than those who passed the CBT ($M_{\text{passed}} = 35.0$), $t(35) = 3.51$, $p = .001$. For the HPT-2, there was no difference ($M_{\text{failed}} = 28.9$, $M_{\text{passed}} = 26.8$).

TIP scores (>10 projected threats) were available for 22 screeners who took the HPT-1 and 15 who took the HPT-2. Only the correlation between HPT-2 scores and d' was significant [$r(14) = .63, p = .01$] and this relationship accounted for 40% of the variability.

2.3 Spatial Memory Test

2.3.1 Test Scores and Times

The M score on the Spatial Memory Test was 6.69, *SD* = 3.25 with test scores ranging from –3.0 to 10. The M number correct was 7.24, and the M number incorrect was 1.63. The average amount of time spent on an item was 5.66 seconds. The distribution of scores is not normal, K-S $z = 2.39, p = .0001$ but skewed to the left.

For this test, the item proportion correct ranged from 0.69 to 0.89 (see Table 10). The correlation of the test item to the overall test score is indicated in the item discrimination column. For the Spatial Memory Test, these values ranged from 0.34 to 0.59. As can be seen in the table, relatively few participants skipped or did not have enough time to answer any single test item.

Table 10. Proportion Correct and Item Discrimination Values for the Spatial Memory Test

Item Number	Proportion Correct	Item Discrimination	Proportion Skipped
Item 1	.88	.40	.17
Item 2	.69	.46	.12
Item 3	.89	.58	.14
Item 4	.87	.50	.09
Item 5	.83	.59	.10
Item 6	.80	.52	.10
Item 7	.82	.46	.09
Item 8	.83	.56	.11
Item 9	.85	.34	.11
Item 10	.71	.47	.13

2.3.2 Reliability of the Spatial Memory Test

The Spatial Memory Test had good reliability with a coefficient alpha of 0.81. This value is very high considering that this is a short ten-item test that would be part of a larger test battery used for selection.

2.3.3 Adverse Impact of the Spatial Memory Test

There was no effect of race or gender on scores. Therefore, there is no adverse impact of the Spatial Memory Test for minority groups or females. Scores for each demographic group are listed in Table 6.

2.3.4 Validity of the Spatial Memory Test

There were 47 screeners who took the Spatial Memory Test in the CBT validation group. There was no difference in test scores between those who failed the CBT ($M_{\text{failed}} = 6.5$) and those who passed the CBT ($M_{\text{passed}} = 7.2$).

There were 16 screeners who took the Spatial Memory Test and were exposed to 10 or more TIP images. There were no significant correlations of test scores with hit rate, false alarm rate, d', or c.

2.4 Spatial Relations Test

2.4.1 Test Scores and Times

The M score on the Spatial Relations Test was 2.39 (*SD* = 3.06). Test scores ranged from –2.33 to 10. The M number correct was 3.56, and the mean number incorrect was 3.51. The average amount of time spent on an item was 7.78 seconds. The distribution of scores was not normal (K-S $z = 2.06, p = .0001$), but skewed to the right.

Item difficulty, as measured by proportion correct, ranged from 0.28 to 0.71. The values for individual items are listed in Table 11. Item discrimination correlations for the Spatial Relations Test, as compared to the Spatial Memory Test, are slightly lower. They ranged from 0.27 to 0.50.

Unlike the previous test, participants appeared to have more difficulty with this test as indicated by the proportion skipped. The range for this test was 0.14 to 036, as compared to the range of 0.09 to 0.17 on the Spatial Memory Test.

Table 11. Proportion Correct and Item Discrimination Values for the Spatial Relations Test

Item Number	Proportion Correct	Item Discrimination	Proportion Skipped
Item 1	.34	.49	.32
Item 2	.71	.27	.22
Item 3	.68	.32	.31
Item 4	.43	.34	.29
Item 5	.60	.37	.33
Item 6	.60	.45	.27
Item 7	.37	.50	.34
Item 8	.28	.44	.14
Item 9	.47	.42	.36
Item 10	.57	.34	.36

2.4.2 Reliability of the Spatial Relations Test

The internal validity of the test (coefficient alpha) was 0.73. This value may be acceptable considering that this is a short 10-item test if it would be part of a larger test battery whose overall score would be used for selection.

2.4.3 Adverse Impact of the Spatial Relations Test

There was a significant effect of race [$F(4,187) = 3.61, p = .007$] and gender [$F(1,187) = 5.15, p = .024$] on Spatial Relations Test scores. Because there was significant nonhomogeneity of variance [Levene's test, $F(8,187) = 4.09, p = .001$)], Dunnett's C was used in post hoc tests. There was a significant adverse impact for black respondents ($M = 1.96$) relative to white respondents ($M = 4.25$). There was also a significant adverse impact for females ($M = 1.94$) relative to males ($M = 3.04$). Scores for all groups are listed in Tables 6 and 7.

The possibility of test bias for gender was examined with an analysis of variance using gender and CBT status as factors. The effect of CBT pass/fail status is significant by this analysis [$F(1,39) = 6.7, p = .01$], as expected from the earlier analysis above. Gender effects in the CBT validation group were not significant, but they were in the same direction and magnitude as for the item analysis group. The M score for females was 2.3, and the M score for males was 3.4. The interaction of CBT status and gender was not significant, and the M differences found did not support a hypothesis of test bias. Females who failed the CBT had a M test score of 0.88 while those who passed the CBT had a M test score of 3.7. Males who failed the CBT had a M score of 2.7, while those who passed it had a M score of 4.0.

2.4.4 Validity of the Spatial Relations Test

There were 46 screeners who took the Spatial Relations test in the CBT validation group. Test scores of those who failed the CBT (M_{failed} = 1.19) were significantly lower than those who passed the CBT (M_{passed} = 3.49) [$t(44) = 2.99, p =.005$].

There were TIP image data and Spatial Relations Test scores available for 17 screeners. Correlations with hit rate, false alarm rate, d', and c were not significant.

2.5 Visual Closure Test

2.5.1 Test Scores and Times

The M score on the Visual Closure Test was 6.76, *SD* = 2.61 with test scores ranging from – 2.33 to 10. The M number correct was 7.28, and the M number incorrect was 1.55. The average amount of time spent on an item was 9.56 seconds. The distribution of scores was not normal, K-S $z = 2.18, p = .0001$, but was skewed to the right.

The proportion correct for the Visual Closure Test ranged from 0.47 to 0.97 (see Table 12). As can be seen in the table, the proportion correct for the first two test items were considerably lower than for the remaining eight. Item to test correlations (item discrimination values) ranged from 0.24 to 0.47. These values are in line with those of the Spatial Relations Test.

For the Visual Closure Test, the proportion of items skipped ranged from 0.09 to 0.21. These values are essentially identical to those of the Spatial Memory Test and somewhat lower than those of the Spatial Relations Test

Table 12. Proportion Correct and Item Discrimination Values for the Visual Closure Test

Item Number	Proportion Correct	Item Discrimination	Proportion Skipped
Item 1	.47	.24	.21
Item 2	.62	.27	.14
Item 3	.81	.36	.11
Item 4	.84	.30	.13
Item 5	.97	.43	.09
Item 6	.93	.47	.09
Item 7	.89	.47	.10
Item 8	.91	.39	.10
Item 9	.89	.38	.12
Item 10	.87	.41	.09

2.5.2 Reliability of the Visual Closure Test

The internal validity of the test (coefficient alpha) was 0.70. Like the reliability of the Spatial Relations Test, this value may be acceptable considering that it is a short test that would be part of a larger test battery.

2.5.3 Adverse Impact of the Visual Closure Test

Analysis shows that there was no effect of race or gender. Therefore, there is no adverse impact of the Visual Closure Test for minority group or females. Mean values for each demographic group are listed in Table 6.

2.5.4 Validity of the Visual Closure Test

There were 46 screeners in the CBT validation group who took the Visual Closure Test. There were no differences in test scores between those who failed the CBT ($M_{\text{failed}} = 7.16$) and those who passed the CBT ($M_{\text{passed}} = 7.81$).

There were 17 screeners in the TIP validation group who took the Visual Closure Test. There were no significant correlations with hit rate, false alarm rate, d', or c.

2.6 Visual Discernment Test

2.6.1 Test Scores and Times

The mean score on the Visual Discernment Test was 7.32 (*SD* = 2.31). Test scores ranged from –1.33 to 10. The M number correct was 7.61, and the M number incorrect was 0.87. The average amount of time spent on an item was 7.32 seconds. The distribution of scores was not normal, K-S $z = 1.84, p = .0001$; it is skewed to the right.

Item difficulty, as measured by proportion correct, ranged from 0.59 to 0.99 (see Table 13). The proportion correct of 6 of the 10 test items was over 0.95 indicating a near-ceiling effect.

Table 13. Proportion Correct and Item Discrimination Values for the Visual Discernment Test

Item Number	Proportion Correct	Item Discrimination	Proportion Skipped
Item 1	.97	.45	.09
Item 2	.96	.36	.11
Item 3	.96	.48	.09
Item 4	.75	.37	.24
Item 5	.59	.16	.27
Item 6	.99	.47	.06
Item 7	.83	.34	.12
Item 8	.99	.52	.13
Item 9	.96	.42	.11
Item 10	.85	.32	.29

Item discrimination correlations for the Visual Discernment Test had a larger range than did any of the previous three tests of the FAA-VST. These values ranged from 0.16 to 0.52.. The proportion of test items skipped ranged from 0.06 for item 6 to 0.29 for item 10. With the exception of the highest and lowest values, the percent skipped values are similar to those of the other tests.

2.6.2 Reliability of the Visual Discernment Test

The internal validity of the test (coefficient alpha) was 0.70. Again, this value may be acceptable considering that it contains only 10 questions which would be part of a larger test battery.

2.6.3 Adverse Impact of the Visual Discernment Test

There was no effect of race or gender. Therefore, there is no adverse impact of the Visual Closure Test for minority groups or females. Mean values for each demographic group are listed in Table 6.

2.6.4 Validity of the Visual Discernment Test

For the 46 screeners who took both the CBT and the Visual Discernment Test, there was no relationship found between scores on the Visual Discernment Test and the CBT pass rate, ($M_{\text{fail}} = M_{\text{pass}} = 7.6$).

There were 17 screeners in the TIP validation group who took the Visual Discernment Test. Again, there were no significant correlations with hit rate, false alarm rate, d', or *c*.

2.7 Combined FAA Visual Skills Tests

2.7.1 Test Scores

The combined test score was constructed by adding up the scores for the four Visual Skills Tests. The average score was 23.99 (*SD* = 7.51).

2.7.2 Reliability of the FAA Visual Skills Tests

From the data obtained for each of the four components of the FAA-VST, it is possible to estimate the reliability of the tests considered as a linear combination of the four test scores.

For the overall test, the reliability (r_{tt}) can be calculated based upon the reliability of the component tests (Nunnally, 1978).

Specifically:

$$r_{tt} = 1 - \left(\frac{\sum \sigma_i^2 - \sum r_{nn} * \sigma_i^2}{\sigma_s^2} \right)$$

where σ_i is the SD of test i and σ_s is the SD of the sum of all tests.

Using the above formula the current battery has a reliability (r_{tt}) of 0.87.

2.7.3 Adverse Impact of the FAA Visual Skills Tests

There was a significant effect of race on test scores [$F(4,178) = 3.62, p = .01$] but no effect of gender [$F(1,178) = .65, p = .42$]. Post-hoc tests (Tukey HSD) of the significant racial difference in scores showed that only the difference between whites and the self-described other group was significant. The other group only had three members. Despite significant adverse impact associated with race and gender on some sub-tests, specifically the HFT-2 and the Spatial Relations Test, the full battery does appear to be fair as constituted.

2.7.4 Validity of the FAA Visual Skills Tests

There were 45 screeners who took all four sub-tests of the FAA-VST. The M combined test score of those who did not successfully complete the CBT was 23.5. The M combined test score of those who did successfully complete the CBT was 26.2. This difference was not significant.

Fifteen screeners, who took all four sub-tests of the FAA-VST, were in the TIP validity group. There were no significant correlations of test battery scores with any of the TIP performance measures.

3. DISCUSSION

3.1 Hidden Figures Test

In its current form, the HFT is very difficult and scores of very few individual test items were above chance and none are well above chance. In addition, the reliability of the test, as measured by the split-half coefficient, is low. Also, the HFT-2 shows adverse impact for females.

The strength of the HFT is that scores correlate with both CBT and TIP performance. These correlations were found despite the relatively small sample size in the validity groups and the poor reliability of the test.

It would be very useful if the reliability of this test could be increased without sacrificing test validity. This may be possible if the best items, as measured by the item analysis, were used in a new configuration of the test.

3.2 Hidden Patterns Test

The data do not support any reasonable method to obtain reliability measures. It would be useful in another selection test deployment if some reliability information could be obtained about the HPT.

In addition, the adverse impact data from the HPT are ambiguous. No main effect of race or gender was found, but there was an interaction of race and gender for HPT-2.

Like the HFT, the HPT correlates with both CBT and TIP performance. Since nobody took both the HFT and the HPT, estimating whether combining theo would enhance test validity is an empirical question. Also, it would be useful to design the next selection test deployment to enable evaluation of the correlation between HPT and HFT scores.

3.3 FAA-VST Battery

The test battery, considered as a whole, has good reliability and does not appear to have adverse impact overall, despite adverse impact associated with the Spatial Relations Test. An unfortunate weakness of the battery lies in the failure to demonstrate criterion-related validity for any of the individual tests (with the exception of the Spatial Relations Test). Since the other components are generally easy, it is possible that the restricted range of scores mitigates validity. Perhaps more difficult versions of these tests may be more valid. The individual components of the FAA-VST are discussed below.

3.3.1 Spatial Memory Test

Spatial Memory Test scores did not demonstrate criterion-related validity. Also, the tests do not vary greatly in difficulty of individual items. It can be made more difficult by increasing the number of black squares from two to three on individual test items.

The strength of the test was that it showed good reliability. Furthermore, the test did not show adverse impact for any minority group or women.

3.3.2 Spatial Relations Test

This test is difficult and shows adverse impact for blacks and females. Relatively simple changes may increase scoring on this test. The most difficult items are items 1, 4, 7, and 8. Items 4 and 7 each have three black squares along with item 10, while all the others only have two. Changes in the number of black squares may reduce the difficulty level of these items.

For items 4 and 8, the pattern of black squares for the correct answer and the distractors is highly similar. Using this criterion, item 1 should be easy. This points to the possibility that the test is not fully understood until item 1 has been completed. There are three ways to make the Spatial Relations Test less difficult:

1. reconfigure items 4, 7, and 8 so that only two squares are blackened,
2. reduce distracter similarity, and
3. rewrite the instructions to increase clarity.

Another problem with the test is that is does not correlate with TIP data. There were no significant correlations between test scores and hit rate, false alarm rate, d', or c.

On a positive note, the reliability of the test may be acceptable considering that it is a short test and that it would be part of a larger test battery. In addition, test scores correlated with CBT scores.

3.3.3 Visual Closure Test

While this test had good reliability, scores did tend to be skewed toward the upper range. This is evident from the M score of 6.7, the median score of 7.3, and the modal score of 8.7. A number of items are characterized by low difficulty (>90% answered the question correctly). This test can be revised to eliminate the least difficult items (specifically items 5, 6, and 8) replacing them with items that have more of the figure removed. This should make those items more difficult and reduce the M, median, and modal scores.

Another downside is that the test did not correlate with either CBT scores nor with TIP performance. On the other hand, there was no adverse impact of the Visual Closure Test for any minority group or for females.

3.3.4 Visual Discernment Test

Like the Visual Closure Test, the Visual Discernment Test had good reliability but scores did tend to be skewed into the upper range. This is evident from the M score of 7.3, the median score of 7.7, and the modal score of 10.0. A number of items are characterized by low difficulty (>90% answer the question correctly). The test should be revised, eliminating the least difficult items; specifically items 5, 6, and 8; and replacing them with more complex alternatives. This should make those items more difficult and move the mean, median, and modal scores down.

The weakness of the test is that it did not correlate with CBT scores. In addition, there was no correlation between test score and hit rate, false alarm rate, d', or c.

Besides reliability, the strength of the test was that it did not show adverse impact for any minority group or females.

4. CONCLUSIONS

As described in the test plan (Fobes & Neiderman, 1998), the strategy for achieving a valid, reliable, and fair selection test battery is by an iterative process of fielding the test, revising the test, and fielding the revision until a battery with desirable properties is achieved. The data analysis indicates some of these tests have some desirable properties although none of them are without deficiencies. The following general revisions are proposed for testing at the next stage.

- Some tests are too difficult and others too easy. Item composition of tests should be changed as discussed above to increase the range of scores of all tests (except the HPT).
- Individual test items that have been identified as undesirable by the item analysis should be eliminated.
- The test structure should be redesigned so that individuals take more than one test and a combination of HPT, HFT, and the Spatial Relations Test is the most promising to pursue. Since different tests seem to have different strengths and weaknesses, the interrelationship of these tests is important.

5. REFERENCES

Fobes, J. L., & Neiderman, E. (1997). *A cognitive model of X-ray security screening: Selection tests to identify applicants possessing core aptitudes* (Technical Report No. DOT/FAA/AR-97/63). Atlantic City International Airport, NJ: DOT/FAA Technical Center.

Fobes, J. L., & Neiderman, E. (1998). *Project plan for the development and validation of a test for initial X-ray screener training* (Technical Report No. DOT/FAA/AR-98/31). Atlantic City International Airport, NJ: DOT/FAA Technical Center.

Neiderman, E. (1997). *Test and evaluation plan for airport demonstration of selection tests for X-ray operators* (Technical Report No. DOT/FAA/AR-97/29). Atlantic City International Airport, NJ: DOT/FAA Technical Center.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw Hill.

See, J. E., Warm, J. S., Dember, W. N., & Howe, S. R. (1995). *Vigilance and signal detection theory: an empirical evaluation of five measures of response bias.* University of Cincinnati, Cincinnati, OH.