U.S. Department
of Transportation

**Federal Aviation
Administration**

Safe Passage Data Analysis: Interim Report

DOT-VNTSC-FA3T5-PM-93-3
April 1993

Eric D. Nadler
Peter Mengert

OPERATOR PERFORMANCE AND ANALYSIS DIVISION

Research and
Special Programs
Administration
John A. Volpe National
Transportation Systems Center
Cambridge, MA 02142-1093

Approved for Distribution:

Division Chief

U.S. Department of Transportation
Research and Special Programs
Administration
Transportation Systems Center
Cambridge MA 02142-1093

Preface

The purpose of this report is to describe quantitatively the costs and benefits of screener proficiency evaluation and reporting systems (SPEARS) equipment, particularly computer-based instruction (CBI) systems, compared to current methods of training, selection, operational certification, operational evaluation and recurrent training.

Part of this effort involves statistical analysis of data derived from actual screeners performing simulated screening on one CBI system, the Public Computer Systems, Inc. (PCSI) Safe Passage System (TM). PCSI agreed to cooperate with the Federal Aviation Administration (FAA) on this study following an FAA announcement in the Commerce Business Daily.

All data was analyzed from each of five major domestic airline sites equipped with Safe Passage. In all, these data represent results from 1465 screeners on more than 1.5 million decisions to pass, inspect or hold X-rayed baggage images.

Preliminary results not only suggest a low percentage of critical errors on Safe Passage images, but also suggest that there is room for improvement in several image categories including: Explosives, Knives, Other Sharp Objects, and Suspicious Innocent. These categories showed less than 80% accuracy. The results also suggest that CBI can be effective in accomplishing the needed improvements, establishing performance near or above 90% for all image categories.

Table of Contents

1. INTRODUCTION

The Safe Passage System presents screeners with X-ray baggage images stored in an approximately 2000-image videodisc library. The images are obtained from actual threat items hidden in actual baggage, which is then X-rayed and stored on videodisc. Thus the screening decisions made on these images are likely to represent actual screening capability.

The images represent nine categories: Innocent, Suspicious Innocent, Electronic Innocent, Explosive, Gun, Knife, Other Sharp Objects, and Combined/Other Weapons. The system's selection algorithm takes more images from screeners' weakest category to provide more practice where it is most needed.

When Safe Passage presents an image, the screener decides whether to inspect, pass or hold the imaged baggage. If the decision is to inspect, the system presents a second image showing the baggage contents arrayed outside of the bag. The screener must then decide to pass or hold the baggage.

Each "test" requires screeners to make self-paced decisions on 12 different X-rayed and imaged pieces of baggage. Feedback on decisions for each bag consists of whether full credit, partial credit or no credit. An overall test score and opportunity to review the images and decisions complete the test. Passing any item that should have been held (a "critical error") results in automatic failure on the test.

There are three proficiency levels in the Safe Passage System.[1] Screeners can attain the higher proficiency levels by maintaining high scores and holding all baggage containing threat items. After maintaining an average of at least 85% on 8 consecutive Low tests, screeners attain the Medium Proficiency Level. They must maintain at least a 75% average and avoid any critical errors on all tests to avoid returning to the Low Level. The same rule applies to transition to and remaining at the High Level. More difficult images are shown to the higher proficiency levels.

2. GOALS OF THE ANALYSIS

Describe CBI <u>benefits</u> as exemplified by Safe Passage performance.

- Describe changes in accuracy, consistency, and speed as screeners transition from the Low to the Medium and High Proficiency Levels.
- Describe variation among sites in the above variables.

Describe CBI <u>costs</u> as exemplified by Safe Passage performance.

- Describe the amount of practice time and calendar days needed to attain the higher proficiency levels, as well as the amount of practice time and calendar days for those screeners who remain at the Low Level.
- Describe variation among sites in the above variables.

1 The proficiency levels are numbered 3, 2, and 1 in Safe Passage. They are designated Low, Medium, and High in this memorandum for clarity.

3. PRELIMINARY RESULTS AND DISCUSSION

3.1 The Sample

All data was analyzed from each of five domestic sites equipped with Safe Passage. The sites are: AA/JFK (American Airlines at Kennedy Airport, New York); AS/JFK (Aviation Safeguards at Kennedy Airport, New York); AA/MIA (American Airlines at Miami Airport); AA/DFW (American Airlines at Dallas - Ft. Worth); and NW/DTW (Northwest Airlines at Detroit Airport). A 500 - screener "Sampled Data Set" was constructed to represent all of the sites.

While the final report will provide all results of interest, for brevity, this memorandum will focus primarily on Sampled Data Set results. The reader may assume that analyses involve the Sampled Data Set unless use of the complete site data is indicated.

3.2 Benefits

Accuracy

Accuracy is determined in several ways:

- Percent of total possible points on a 12-image test
- Percent of decisions resulting in a critical error
- Points on a 200-point "Performance Index" which combines the preceding two measures in a single overall index
- Percent of total possible points on images in particular image categories
- Probability of decisions to hold innocent images.

Performance Index

This accuracy measure was constructed by assigning 100 points if no critical error was made on the 12-image test or zero if one or more critical errors were made, plus the percentage of points (100 maximum) assigned for accurate decisions within Safe Passage.

Site

The Performance Index results for the five sites are provided in the two following tables.

Performance Index by Site

Site	Tests	Mean	Standard Deviation
AA/DFW	4,273	167.44	47.78
AA/JFK	9,813	169.15	47.11
AA/MIA	6,615	160.84	50.48
AS/JFK	11,185	176.04	43.14
NW/DTW	11,829	171.30	45.58

An analysis of variance (general linear model) procedure found the differences among these sites to be significant, $F(4, 43{,}714) = 118.72$, $p < .0001$. A Tukey HSD test found all pairwise comparisons significant ($p < .05$) except the AA/DFW and AA/JFK comparison. Figure 1 displays the relationships among the five sites' mean Performance Index scores.

The above results represent screeners' performance proportionate to the amount they use Safe Passage. The amount of safe Passage use and scores are likely to be positively correlated, so

that the above results may disproportionately represent higher-scoring screeners who use Safe Passage more, instead of the average Safe Passage user.

Therefore, we calculated the mean Performance Index for each screener and then again obtained the Performance Index results for each site:

Performance Index by Site

Site	Screeners	Mean	Standard Deviation
AA/DFW	100	157.34	28.99
AA/JFK	100	157.35	24.83
AA/MIA	100	156.32	23.78
AS/JFK	100	165.15	26.17
NW/DTW	100	155.92	25.45

These results provide equal weight to each screener, without regard to the number of tests taken. An analysis of variance performed on these data did not find significant differences, $F(4, 499) = 2.17$, $p. > .05$. The best estimate of typical Performance Index at each site lies between the values in the above two tables.

Proficiency Level

The horizontal distance between the lines on Figure 2 shows an approximately 15-point difference between the Low and Medium performance levels and an approximately 5-point difference between the Medium and High levels, for about 90% of the 500 screeners in the sample. The percentages of screeners advancing to the Medium and High performance levels were 69.2% (to Medium) and 55.4% (to High).

Percent of Decisions Resulting in a Critical Error

Few screeners committed critical errors on Safe Passage images. Figure 3 shows that in our analysis, about 50% committed none at each proficiency level. The effectiveness of CBI is shown by results for the 50% who did commit critical errors: About 75% Low, 86% Medium, and 92% High Proficiency Level screeners made a critical error in *less than one percent* of their decisions. Also, about five percent more Low than Medium and High Proficiency Level screeners made a critical error in *more than one percent* of their decisions.

Image Category

The last 20% of each screener's scores was analyzed at each proficiency level. This was done to characterize performance after the learning curve had stabilized. The results show that all image categories rise to about 90% accuracy, or better. Those that showed the lowest Low Proficiency Level accuracy thus show the most dramatic gains.

Percent Accurate by Image Category

Image Category	Proficiency		
	Low	Medium	High
Innocent	88.05	91.51	95.15
Suspicious Innocent	75.13	76.17	87.40
Electronic Innocent	89.46	86.48	89.40
Explosive	77.40	79.30	93.13
Gun	93.00	93.78	94.69
Knife	76.83	83.57	93.04
Other Sharp Objects	77.32	80.53	89.56
Combined/Other Weapons	90.20	93.90	93.88

Probability of Decisions to Hold Innocent Images

This aspect of simulated screening accuracy was included because improved recognition skills should afford a reduced incidence of incorrect decisions to hold bags in the operational X-ray. The analysis was limited to images that should have been passed, thus excluding those that should have been inspected and then passed. Figure 4 shows an approximately 50% decrease in the probability of decisions to hold innocent images as screeners transition from the Low Proficiency Level to the Medium and High Proficiency Levels. This decrease is more pronounced for decisions to hold the imaged baggage than for decisions to inspect and then hold the baggage.

Consistency

Two main measures of screener consistency were used: within- subjects standard deviation and between-subjects standard deviation. The within-subjects standard deviation is based on the variability of each screener's test scores (Performance Index). In contrast, the between-subjects standard deviation is based on the variability between screeners of the screeners' test score averages. The following standard deviations were calculated from the complete data sets.

Within-Subjects Standard Deviation

Site	Tests	Proficiency Level		
		Low	Medium	High
AA/DFW	11,689	54.63	45.32	41.15
AA/JFK	24,904	54.40	46.07	40.99
AA/MIA	17,446	54.60	46.81	43.80
AS/JFK	21,951	55.00	46.30	34.46
NW/DTW	48,998	54.87	47.99	39.41

Between-Subjects Standard Deviation

Site	Screeners	Proficiency Level		
		Low	Medium	High
AA/DFW	258	18.84	16.50	14.64
AA/JFK	226	16.42	12.99	11.59
AA/MIA	302	16.67	10.87	12.14
AS/JFK	263	17.72	12.28	8.73
NW/DTW	416	17.06	14.64	12.27

The decreased within and between-subject standard deviations that were found with increased proficiency level may have resulted from test scores approaching ceiling performance. Further analysis is planned.

Bartlett's test for homogeneity of variance was used to determine whether there were significant differences among the sites' overall within-subject and between-subject variances. The results showed that both the sites' within-subject variances (Chi Square (4) = 586.06, and between-subject variances (Chi Square (4) = 23.736) differed significantly, $p < .005$.

Speed (Test Duration)

Simulated screening speed was estimated by obtaining the average duration for each 12-image test, using the complete data sets. We assumed that screeners viewed images only as long as necessary. This is likely because Safe Passage is self-paced.

Test durations less than 0 sec or greater than 300 sec were filtered out after it became apparent that negative durations were recorded when test-taking crossed midnight, and that some extremely lengthy durations occurred, perhaps when a screener left the test without completing it. This filter removed 465 of 43,725 records.

Test Duration

Site	Tests	Proficiency Level		
		Low	Medium	High
AA/DFW	4,195	139.70	125.19	106.86
AA/JFK	9,752	113.64	109.36	90.75
AA/MIA	6,465	130.13	137.09	122.00
AS/JFK	11,115	123.90	123.52	97.29
NW/DTW	11,723	121.07	113.95	87.49

For reasons analogous to those presented with the Performance Index results, the above results may reflect faster screeners who use Safe Passage more, instead of the average Safe Passage user. Therefore, we calculated the mean test duration for each screener at each proficiency level and then again obtained the test duration results for each site. As with the Performance Index results, the best estimate of typical test duration at each site lies between the values in the two tables.

Test Duration

Site	Screeners	Proficiency Level		
		Low	Medium	High
AA/DFW	189	145.01	126.97	114.26
AA/JFK	249	127.93	109.94	101.88
AA/MIA	217	142.05	133.23	123.59
AS/JFK	222	147.58	124.59	108.26
NW/DTW	233	122.71	106.93	90.19

The above results indicate that screeners view each image, on average, less time as their proficiency level increases. This result suggests that screeners do not trade off speed for accuracy in order to advance.

3.3 Costs

Practice Time Required to Achieve Safe Passage Proficiency Levels

This analysis determined the total practice needed to reach the Medium and High Proficiency Levels. In the Safe Passage system, screeners must maintain 75% and commit no critical errors to remain at either of these higher levels. As a result, screeners did not always advance directly from Low to Medium to High. Instead, sometimes screeners returned to a lower proficiency level. Thus, the duration of all tests prior to the first test taken at the Medium or High Proficiency Levels were included, without regard to any lower level tests taken after those levels were reached.

The results are shown in a cumulative distribution (Figure 4) to indicate the maximum practice time any proportion of screeners required to reach either of the two advanced proficiency levels. For example, 90% of the screeners who reached High Proficiency took 120 min or less to reach this level.

Figure 4 also provides data related to screeners who remain at the Low Proficiency Level. The majority of these screeners (60% to 65%) practice less than the time needed for any screener

to reach a higher proficiency level. Only 20% practice longer than the average (median) screener takes to attain Medium Proficiency.

Calendar Days Required to Achieve Safe Passage Proficiency Levels

This analysis determined the total number of days needed to reach the Medium and High Proficiency Levels. This calculation was made by determining the number of days from the first Low Proficiency Level test date to the first test date at the Medium or High Proficiency Level, and then adding one. Thus a screener who began at the Low Proficiency Level and who also advanced from Low to Medium Proficiency on the same day is said to have required one day.

The results are shown in a cumulative distribution (Figure 5) to indicate the maximum number of days any proportion of screeners required to reach either of the two advanced proficiency levels. For example, the average (median) screener requires at least 15 days to attain Medium Proficiency, and at least 50 days to attain High Proficiency. It required less than about 100 days for 90% of the screeners who attained High Proficiency to do so.

4. Preliminary Conclusions

The image category results clearly indicate room for improvement in screeners' ability to distinguish threat from innocent items in X-ray images of baggage. Four of eight image categories show accuracy less than 80% at the Low Proficiency Level. These are the Suspicious Innocent, Explosive, Knife, and Other Sharp Object categories. It is unlikely that these findings result from inexperience with Safe Passage because only the last 20% of tests were included in this analysis.

The image category results also clearly demonstrate the potential for computer-based instruction (CBI) to increase image recognition capabilities. The High Proficiency screeners' accuracy was about 90% or better for all image categories. The approximately 50% reduction in decisions to hold innocent images, as well as the substantial reduction in critical error rates, represent additional compelling illustrations of CBI effects.

Nearly all screeners who reached the High Proficiency Level required less than 120 min. of practice. During this time they made decisions regarding 600-800 images. Yet, in current applications, this practice is acquired sporadically: One fourth of these screeners reached High Proficiency in 12 days, and half within 44 days.

The results indicated statistically significant differences among sites in accuracy (Performance Index) and consistency. It is important to understand the reasons underlying these findings.

Two additional analyses are currently underway. One will determine the extent to which High Proficiency Level performance results from CBI, as opposed to lower-scoring screeners leaving the sample. This involves obtaining the learning curves of screeners who reach High Proficiency, compared to those who remain at lower proficiency levels. The second will determine the proportion of tests taken at High Proficiency after that level is reached. This information is needed to more fully describe the effects of reaching High Proficiency.

In addition to these analyses, a field study should be conducted to determine the operational effects of CBI. Forthcoming recommendations will be based on additional analysis of CBI data, field study results, and interviews with industry experts.

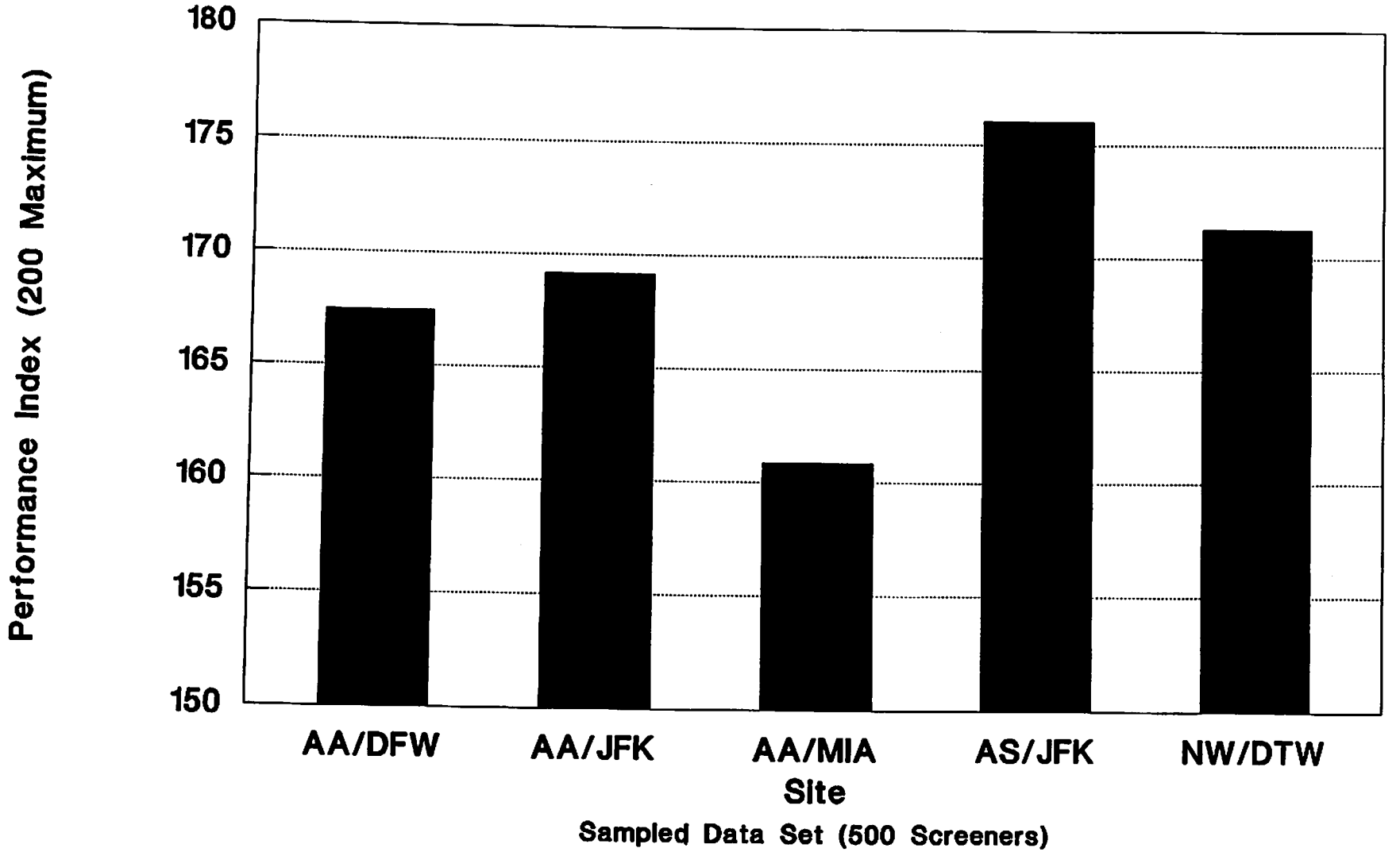

FIGURE 1. PERFORMANCE INDEX DIFFERENCES AMONG SITES

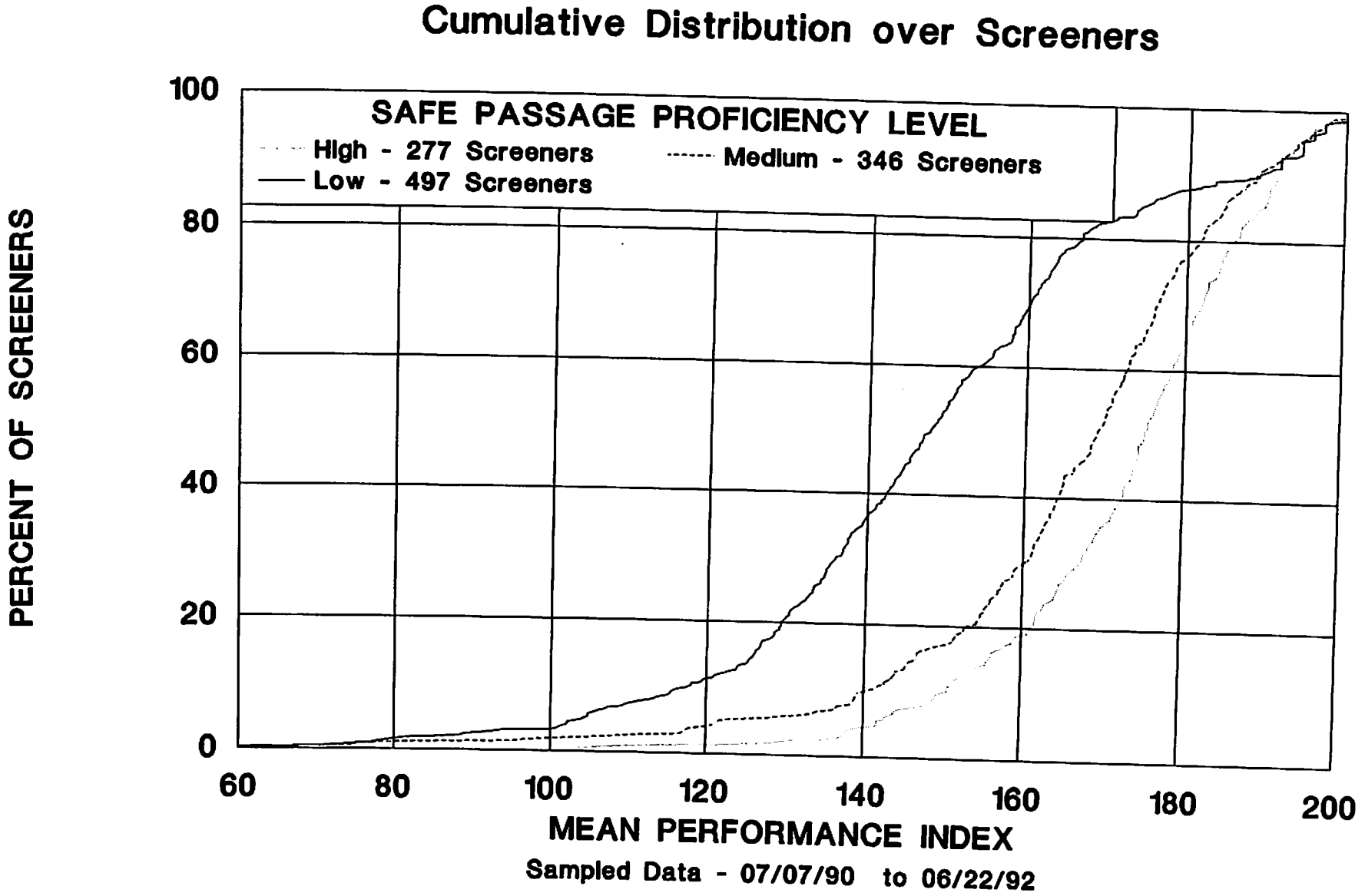

FIGURE 2. AVERAGE SAFE PASSAGE PERFORMANCE BY SAFE PASSAGE PROFICIENCY LEVEL

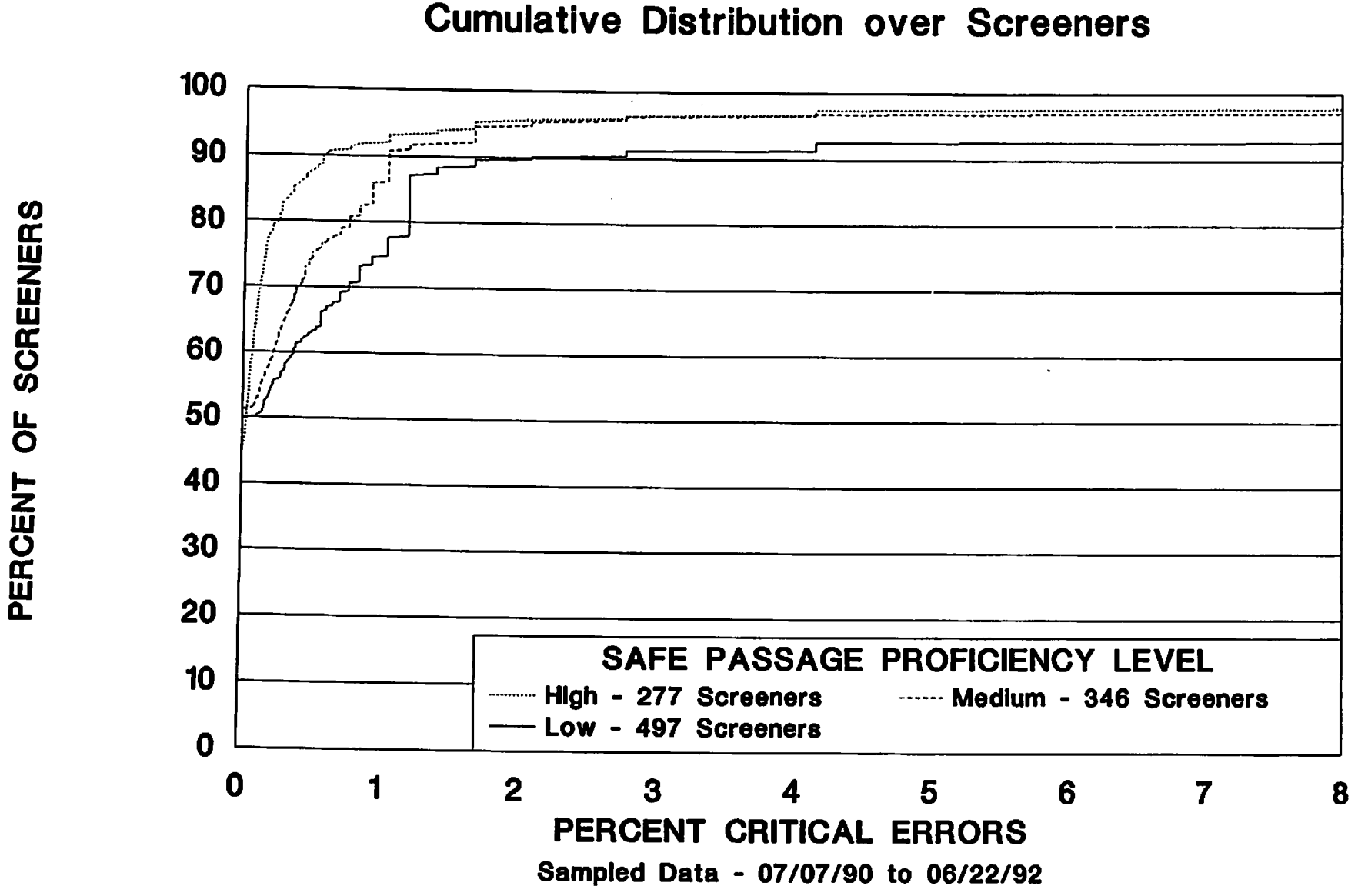

FIGURE 3. SAFE PASSAGE CRITICAL ERRORS BY PROFICIENCY LEVEL

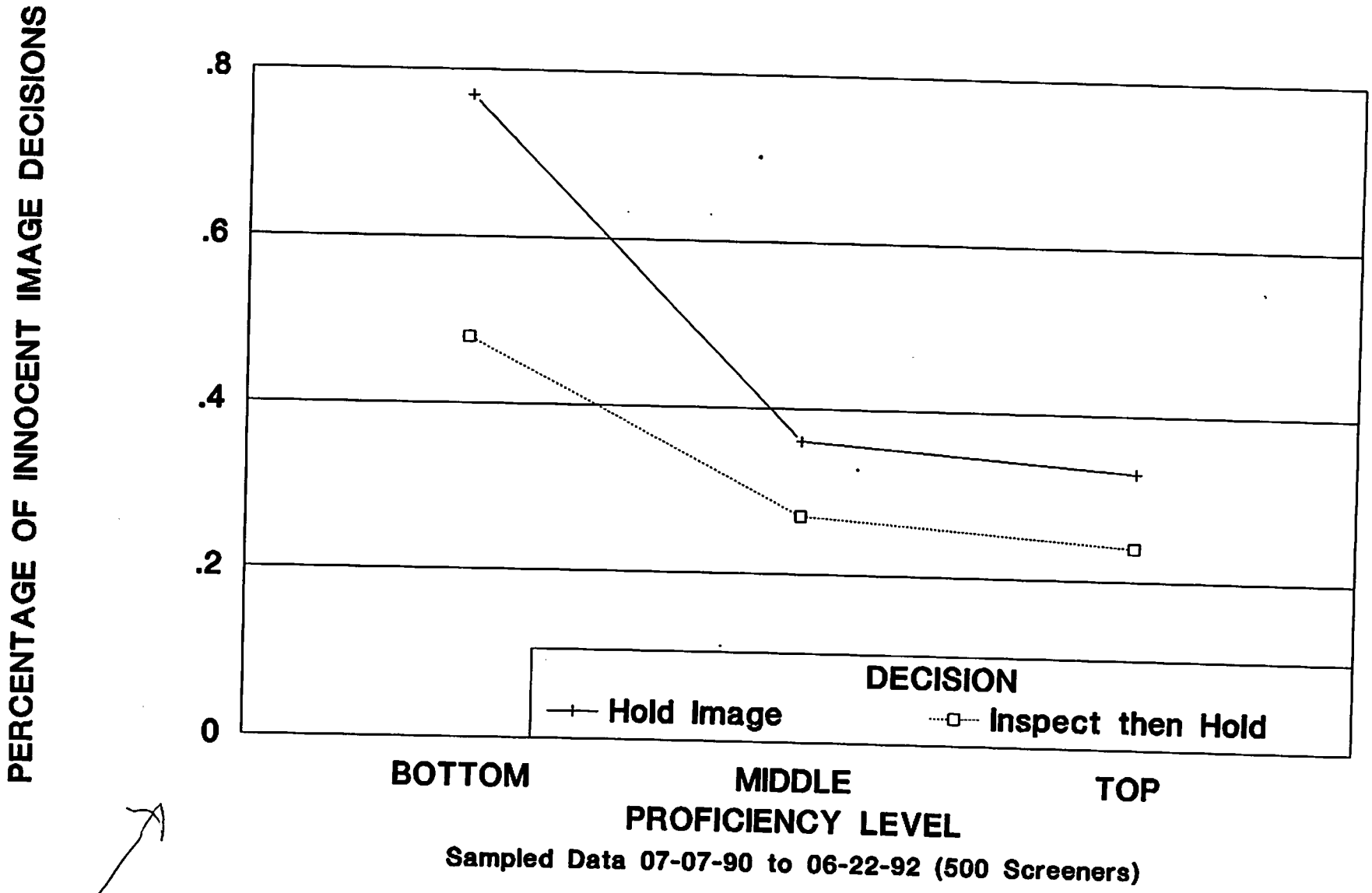

FIGURE 4. DECISIONS TO HOLD INNOCENT IMAGES

fraction

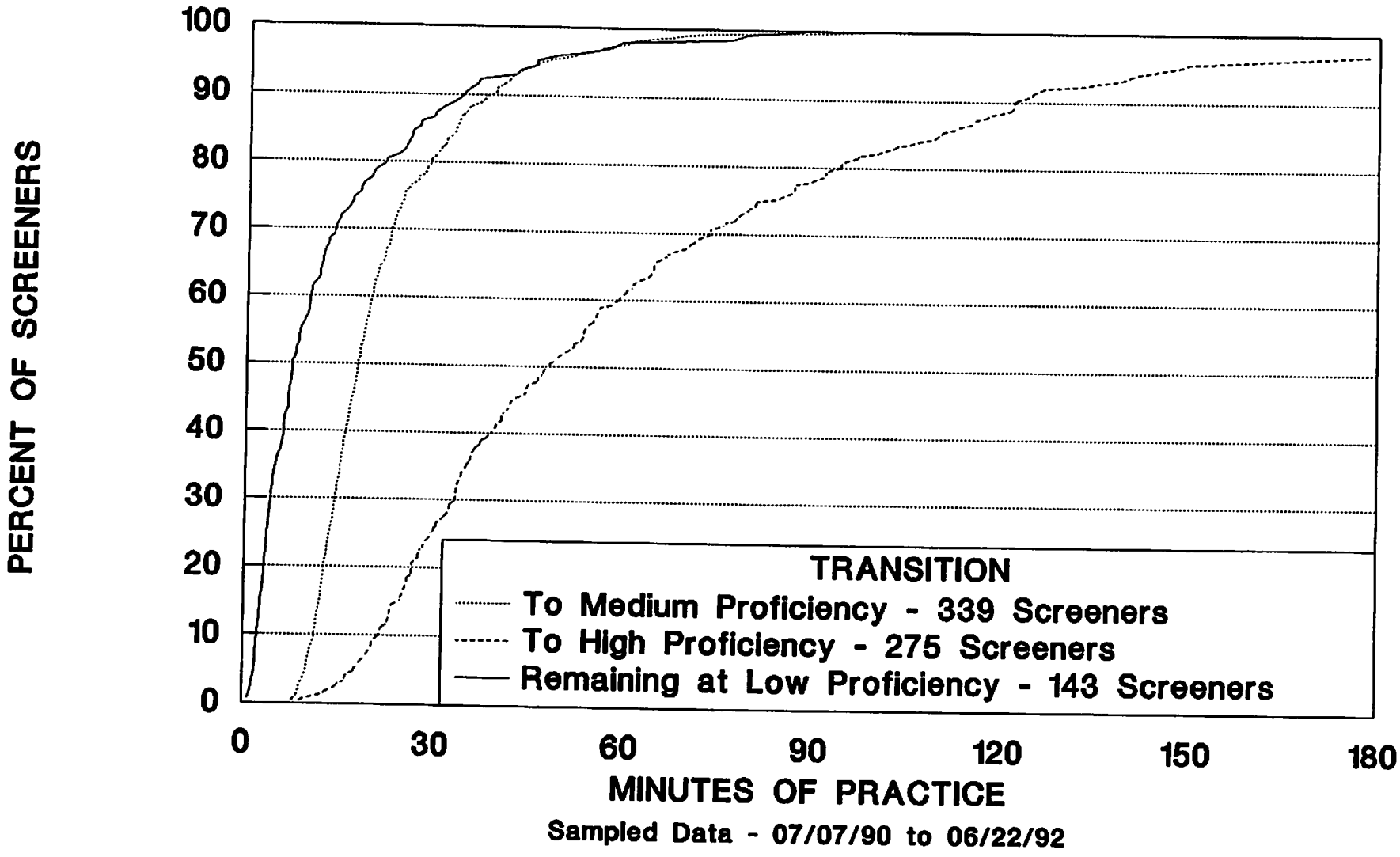

FIGURE 5. CUMULATIVE DISTRIBUTION OF TIME TO REACH SAFE PASSAGE PROFICIENCY LEVELS

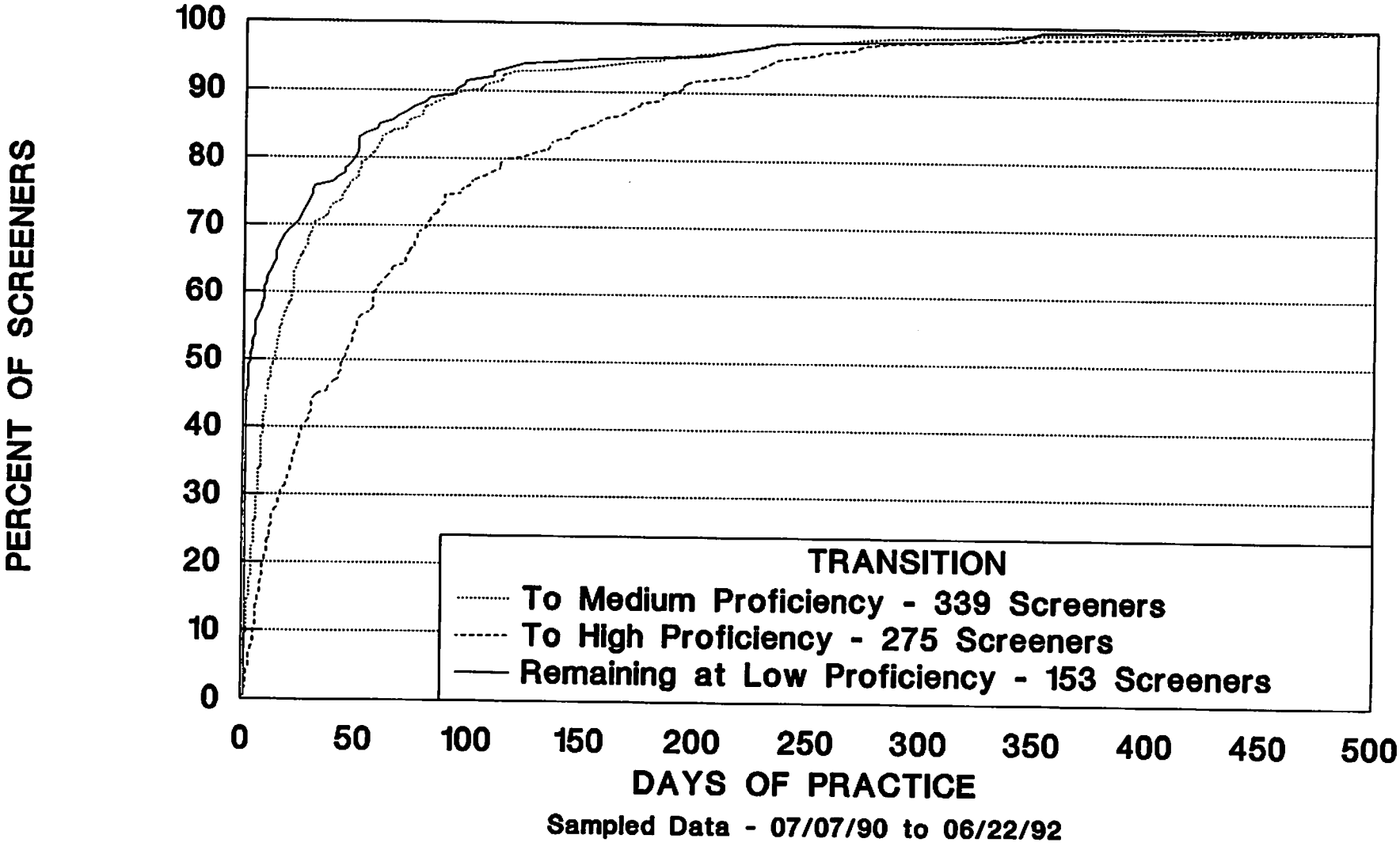

FIGURE 6. CUMULATIVE DISTRIBUTION OF DAYS TO REACH SAFE PASSAGE PROFICIENCY LEVELS