J. C. LEACHTENAUER
The Boeing Company
Seattle, Washington

Photo Interpretation Test Development

Psychometric test development procedures are used to develop sensitive measures of P.I. performance.

ABSTRACT: *An interpreter performance test was developed for use in studies of interpretation equipment and methods. The test was designed to measure primarily the visual rather than the cognitive aspects of interpreter performance. Forty inexperienced and ten experienced interpreters examined 85 frames of photography and were required to search for and identify five types of targets. An item analysis was conducted to develop two equivalent forms of a performance measure. Each form consisted of 23 frames of imagery containing 32 targets. These matched sets of items allow the same group of subjects to be tested under two experimental conditions with much greater sensitivity than is normally achieved. With only five subjects, 7% performance differences can be detected at the 95% confidence level. Chi square analyses were performed to determine the effects of target, scene and subject differences. Target type and target background complexity had the greatest effect on performance; subject experience had relatively little effect.*

INTRODUCTION

INTRUSION OF UNWANTED interpreter and imagery differences into test results is a classic problem in interpreter performance research. Differences caused by these two factors often overshadow the effects of the variables of interest. As a result, performance differences of 15 to 20% may be found to be nonsignificant statistically. Increasing research costs make it increasingly difficult to justify such a lack of experimental sensitivity.

What is the solution? One possible approach is the use of larger samples. More imagery or more test subjects can provide greater sensitivity. Unfortunately, the experimenter seldom knows beforehand just how large a sample is needed. Too large a sample is wasteful.

A second possible approach involves matching. Subject groups of equivalent ability or imagery sets of like difficulty may be established before testing. Background knowledge or results from a pretest may be used to match subjects. In both cases, there is no guarantee that the criterion used for matching will correlate highly with actual test performance. If there is low correlation, the subjects will not be matched and again, low sensitivity may result.

Matching of imagery, on the other hand, offers much greater potential for success. Using conventional psychometric test development procedures, two or more interpretation tests of equivalent difficulty can be produced. The same subject can thus be tested under two or more test conditions with tests measuring the same parameter but unaffected by memory of test details. This approach provides the experimenter with a precise measurement tool which can be used for a variety of experiments.

Knowing the characteristics of the tests or performance measure prior to experimentation enables the experimenter to specify the number of test subjects needed to detect a given performance difference at a specified confidence level. This capability considerably reduces the resources required to conduct a given experiment and increases the sensitivity of each experiment.

Objective

A series of experiments on interpretation equipment and methods were planned. Rather than develop a new set of imagery and test questions for each study, a decision was made to produce a single performance measure.

Goals and methodology for developing the performance measure were established after review of the planned study objectives and the relevant literature. Results of this review were presented in a previous report.[1] The primary goal was to develop three equivalent forms of a search and identification task performance measure, each form containing 30 to 35 questions or items. Desired reliability was 0.70. This would provide sensitivity sufficient to detect 10% performance differences with 20 test subjects. The availability of three equivalent forms would enable the same subjects to be tested under each of three experimental conditions.

A secondary objective was to investigate the effects of image and subject differences. A comparison of naive and experienced subjects was of major interest. If naive subjects performed as well as experienced subjects, the subject population available for subsequent use would be greatly enlarged. Image and target differences were of interest because of their effect on item difficulty.

Method

Development of the photo interpretation performance measures followed conventional psychometric test development procedures. These procedures involved three major steps:

1) development of an initial item pool containing several times as many items as were desired in the final versions of the tests;
2) administration of the item pool to a large group of subjects; and
3) statistical analysis of the performance data to select two or more sets of items which would provide equivalent performance.

Final sets of items were selected using two criteria. Mean performance achieved on each set of items was to be approximately 50%, and the performance of an individual on any given item was to be consistent with his overall performance. In other words, the correlation between average score on an item and the overall score of each subject receiving the item was high. Sets of items selected on these two criteria provide tests having maximum sensitivity with a minimum number of test items and subjects. Further explanation of test development procedures is given in Reference 2.

The first step in developing the item pool was to determine the types of targets to be used. Imagery was reviewed and a sample selected showing a wide variety of target types. Using this sample, a pretest was conducted to determine relative difficulty of target detection and identification. Pretest results were used to select the following target types:

1) NIKE missile sites;
2) transformer yards;
3) electronic sites;
4) air strips; and
5) construction sites.

These target types provided a desired range of interpretation difficulty, were sufficiently numerous on the available imagery, and were considered relevant.

A large number of imagery examples of the five target types were then acquired and analyzed. Obtained from the U.S. Air Force Reconnaissance Data Base at the Rome Air Development Center, the imagery consisted of 900 frames of 9 × 9 inch panchromatic transparencies covering various eastern U.S. locales. Scale ranged from 1/7,000 to 1/21,000. Experienced interpreters located all of the target types on this imagery and classified them in terms of size, contrast, location on the image, and scene complexity. Each frame was also classified in terms of target density and general scene type.

Target size and contrast were determined separately for each target type. In the case of transformer yards, for example, facilities having four or more banks of transformers were classified as large, all others were classified as small. In the case of contrast, a subjective determination of target to background contrast was first made on a ten point scale. Examples of each target type were then divided into high and low contrast groups, each group containing approximately the same number of items.

Target location on the frame was classified as "inner" or "outer," inner being the central 5 × 5 inch area of the frame. High scene complexity indicated the presence of other buildings or background clutter within a 0.5 inch radius of the target. Target density measured the number of targets on a frame, high target density indicated the presence of three or more targets on a single frame. Scene type was classified as rural, urban, or mixed. A rural scene was defined as a frame show-

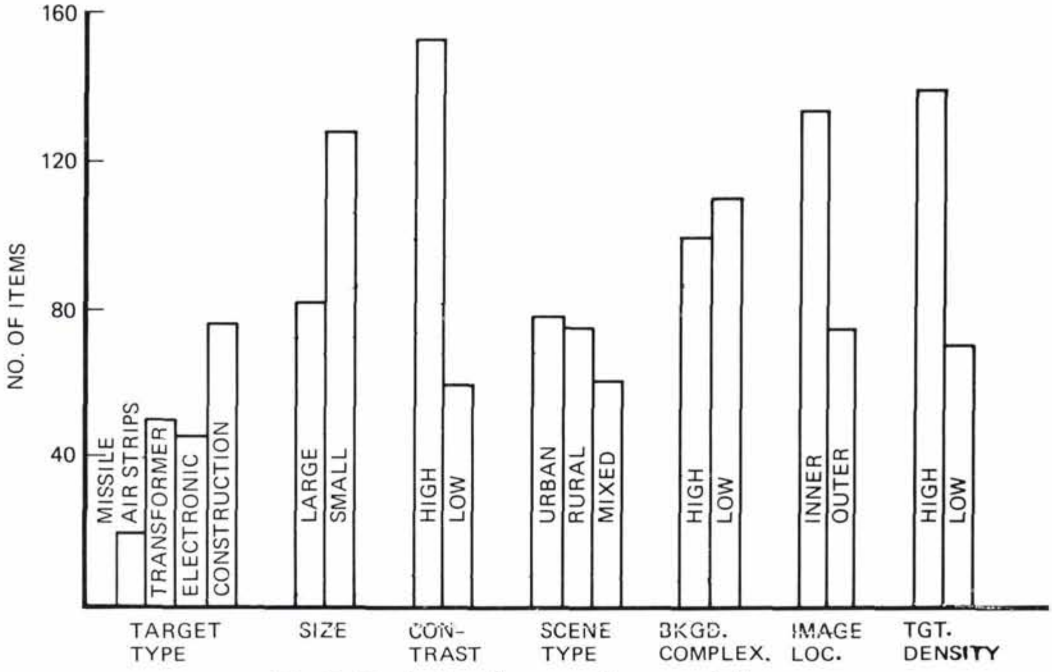

FIG. 1. Distribution of target and scene variables.

ing greater than 80% open land; an urban scene showed less than 20% open land. Intermediate scenes were classified as mixed.

From the original sample of 900 frames, 85 frames containing 212 targets were selected. Each target constituted a test item. To the extent possible, items were selected to provide an even distribution of target types, and for each target type, an even distribution of the remaining variables. Figure 1 shows the actual distributions obtained.

The 85 selected frames ranged in scale from 1/12,600 to 1/18,000. Test subjects viewed fourth generation positive transparencies. These were roughly equivalent in image quality to USFS or USDA imagery of comparable scales. Ground resolution varied from two feet to five feet.

All subjects were employees of the Boeing Company. A group of 40 naive subjects were selected as the primary sample. They were college graduates with no previous experience in image interpretation or allied fields (flying, navigating, radar scope operation). For purposes of comparison, a group of 10 experienced interpreters was also utilized. Members of this group had previous image interpretation experience, ranging from 18 months to 23 years with a median of 5 years. All subjects had near vision correctable to 14/14 or better and passed a simple test of tonal acuity, requiring discrimination of 16 or more steps of a 21-step density wedge.

Immediately prior to taking the test, each subject was given one hour of training on characteristics of the targets and practice on the test task. Photos and graphic abstractions of the targets were used as training aids. An experimental session lasted approximately eight hours and consisted of one hour of training and seven of testing. One session was required for each subject.

Subjects viewed the imagery on light tables which had 10 × 10 inch horizontal viewing surfaces. Illumination intensity was set at 375 footcandles measured at 5.25 inches. A digital counter, mounted behind each light table, displayed elapsed time in tenths of seconds and was coupled with an on/off switch for control of the light table.

Subjects were trained and tested in groups of three. Each of the three light tables was placed in a separate enclosure. The experimenter could view all three subjects but a subject could not see the other subjects or the experimenter.

The test required each subject to search each of the 85 frames for targets. When a target was located, the identification and location of the target were recorded on a separate response sheet. Target location was reported using a 9 × 9 alphanumeric grid located around the periphery of the viewing surface. Time data were obtained by having the subjects start the digital counter each time they began searching a frame and record elapsed time when they finished the frame.

Subjects were told that their performance would be scored on the basis of speed, completeness, and accuracy. They were provided with a 7× tube magnifier, sketches of the targets (Figure 2), and a list of identification cues. The order of presentation of frames was partially counterbalanced by use of the Latin square.

Scoring consisted of comparing each subject's response with a previously developed scoring key. Some latitude was allowed in position accuracy, the amount depending on the particular scene. In complex scenes, where some doubt existed as to the object that the subject had seen, a maximum error of 0.25 inches was allowed. In rural scenes, 0.5 inches of error was allowed.

Each subject's response was scored as correct, a misidentification, targets omitted or a false alarm (incorrect response to a non-target). Prior to scoring a response as a false alarm, an independent determination was made by three experienced interpreters that the object in question was a non-target rather than a target. If it was a target that had been missed in preparing the scoring key, it was entered in the key and data from other subjects were adjusted.

For each item, data were available to show whether or not each subject had correctly responded to the target. For each frame, data were available to show total time spent, correct responses, misidentifications, targets omitted, false alarms, and the order in which the responses were made.

Results

Performance data were subjected to two types of statistical treatment. First, an item analysis designed to select items for inclusion in final forms of the test was conducted. Second, Chi square analyses were performed to determine the effects of target and scene variables.

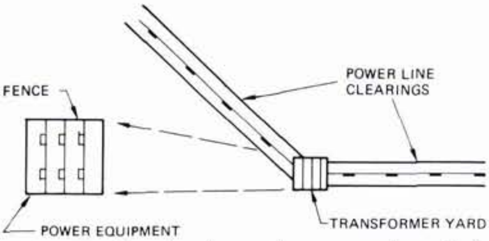

FIG. 2. Sketch of transformer yard used for training.

Item Analysis

The completeness scores for each item were analyzed to select items for equivalent forms of the search and identification test. The first step in the process involved calculation of the point biserial correlations between the subjects' total test scores and their responses for each item (r_{IT}).

Using the calculated correlation coefficients and item completeness scores, a Gulliksen internal consistency coefficient[4] and a Spearman-Brown split-half reliability value[2] were calculated. Results were in general agreement and thus the simpler split-half method was used in the remainder of the analysis.

Data on the proportion of subjects passing each item and point biserial correlations (r_{IT}) were next reviewed to eliminate items to which all or almost all subjects responded correctly or incorrectly or which had low correlation values. The reduced list contained 88 items. The goal of the study had been to produce three parallel 32 item tests with a reliability of 0.70. With the large number of items eliminated, this goal could not be met. A decision was thus made to aim for development of only two parallel forms of the test.

Using the 88 item test data, further item analysis was performed to reduce the number of items to 64. This item analysis used Gulliksens' internal consistency reliability[4] as a criterion and eliminated items which did not increase reliability.

Using the technique suggested by Guilford[2] for making equivalent test forms, the 64 items were plotted to show the distribution of r_{IT} versus the proportion passing. Items which plotted close together were selected in pairs to provide the basis for assignment to one of the two parallel test forms. Item assignment was accomplished so as to balance, for each form of the test, the mean r_{IT} and proportion passing. In addition, it was also necesary to assign items such that they appeared on different frames of imagery. Table 1 summarizes the test statistics.

An independent or hold out group was not used for validation and consequently the calculated test reliability value of 0.84 shown in Table 1 is biased upward. To obtain a

TABLE 1. TEST STATISTICS SUMMARY

Characteristic	*Form I*	*Form II*	*Total*
No. of Frames	23	23	46
No. of Items	32	32	64
No. of Missile Sites	4	4	8
No. of Airfields	6	5	11
No. of Transformer Yards	7	9	16
No. of Construction Sites	8	7	15
No. of Electronics Sites	7	7	14
Mean Difficulty Level (proportion passing)	0.480	0.470	0.475
Mean Point Biserial Correlation (r_{IT})	0.274	0.280	0.277
Corrected Split Half Using Spearman Brown	0.85	0.79	0.84*

* Upper estimate due to item selection process; final estimated reliability is 0.81.

better estimate of reliability without the expense of using another group of subjects, a reliability value based on random selection of the 64 test items was calculated. The value obtained, using the Spearman-Brown prophecy formula, was 0.78. Since the item selection process actually used would be expected to produce closer correlation than that obtained by random assignment, an estimated reliability value of 0.81 was assigned to the test.

COMPLETENESS

Completeness, defined as the percentage of items present which were correctly identified, was 25.3% for the original pool of 212 test items. The remainder of the items were either not found or were misidentified.

Seven Chi Square analyses were performed on completeness scores (Reference 3, p. 629). Because these analyses were not the primary objective of the effort, unequal and sometimes small numbers of observations occurred in the different experimental conditions. Therefore, in order to have sufficient data for reliable analysis of a variable, it was necessary to analyze the data in different combinations and arrangements. Table 2 lists the analyses performed.

This approach inflates the chance of finding significance and, with the procedures used, tends to produce significant interactions which may have little practical meaning. For statistically significant interactions, it was sometimes necessary to refer back to the actual target images used in order to determine whether the interaction had any practical significance or occurred simply because of the specific target images utilized or the manner in which data were combined. The reader is therefore cautioned to keep in mind the limitation of the data analysis procedures, and to recognize that the analysis was not an end in itself but was simply a supplement

TABLE 2. CHI SQUARE ANALYSES OF COMPLETENESS SCORES

Analysis	*Primary Variables*	*Secondary Variables*
1	Complexity and Size	Subject Type, Target Type (Transformer, Construction, Electronic Data Only)
2	Size and Contrast	Subject Type, Target Type (Transformer, Construction, Electronic Data Only)
3	Location and Size	Subject Type, Target Type (Transformer, Construction, Electronics, Missile Site Data Only)
4	Location and Complexity	Subject Type, Target Type (Transformer, Construction, Electronic Data Only)
5	Density and Location	Subject Type, Target Type (Transformer, Construction, Electronic, Air Strip Data Only)
6	Density, Location and Scene Type	Subject Type
7	Complexity, Location, Size and Contrast	Subject Type

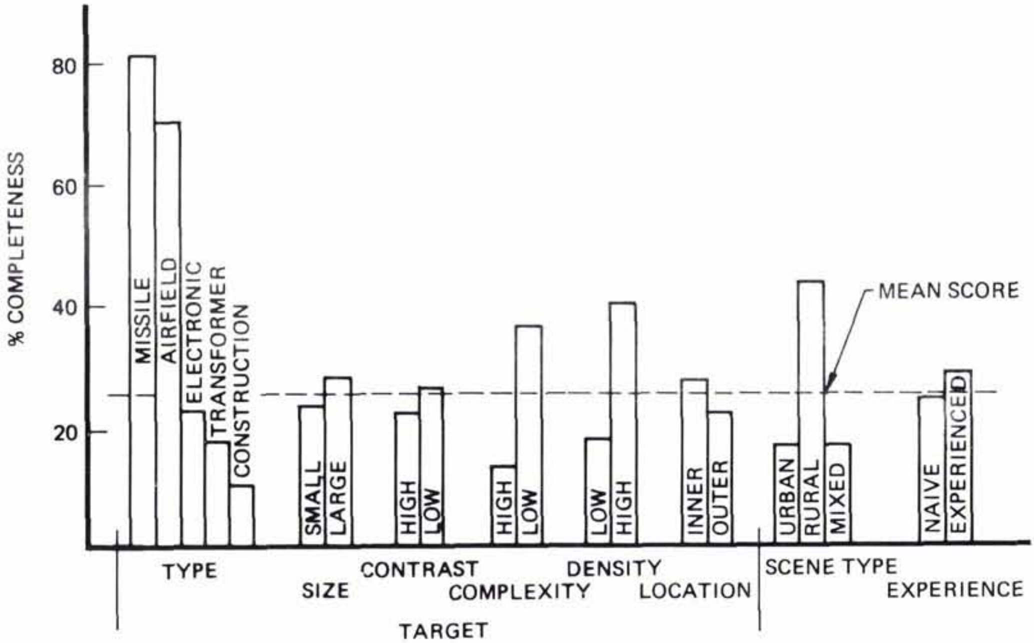

FIG. 3. Effects of experimental variables on completeness scores.

to the development of a test. Chi Square analysis indicated that the main effects of all independent variables on completeness were statistically significant at the .01 level. Had a .005 level of confidence been applied, all but the subject difference effect would have shown significance. A summary of completeness score differences produced by each of the independent variables is shown in Figure 3.

A number of significant ($p < .01$) interactions among the independent variables were found and these are summarized in Table 3. In all of the analyses in which target type was a variable, a significant three factor interaction occurred with two other variables

TABLE 3. SUMMARY OF SIGNIFICANT COMPLETENESS INTERACTIONS*

Two-Factor Interaction	*Three-Factor Interaction*	*Four-Factor Interaction*
Target Type × Size	Type × Size × Contrast	
Target Type × Contrast	Type × Size × Complexity	
Target Type × Location	Type × Size × Location Type × Location × Complexity Type × Density × Location	
Target Size × Complexity	Size × Complexity × Contrast Size × Complexity × Location Size × Contrast × Location	Size × Complexity × Contrast × Location
Target Location × Complexity	Location × Contrast × Complexity	
Target Location × Density Target Location × Scene Type		

* All interactions significant at 0.01 level.

describing target or scene characteristics. Target size, location, contrast, complexity, and density were included in these three factor interactions. The existence of these interactions suggests that the variables used to describe target characteristics were not uniformly effective.

In several instances, a particular target type interacted with another variable, producing results contrary to those found in the tests of main effects. For example, performance was better on small rather than large construction sites, and better on low rather than high contrast transformer yards. For large construction sites, performance was better on low contrast than high, the same reversal occurred for small transformer yards.

A review of these data led to the following observations regarding the effects of the target descriptor variables. Target size was an effective means of classifying target difficulty for transformer yards and electronic sites. It was not effective for missile sites and no data were available on airfields. The failure on missile sites was most likely due to a restricted size range. For construction sites, results were apparently confounded by contrast. Large construction sites tended to be of relatively uniform contrast, this contrast was somewhat lower than that found for many small construction sites.

Target contrast was an effective predictor of difficulty only for electronic sites. Confounding with size occurred for construction sites. Transformer yards apparently had an insufficient range of contrast. No data were analyzed for missile sites and airfields because of insufficient examples.

On an overall basis, target complexity proved to be an effective predictor of target difficulty. Large transformer yards were one exception. Again, no data were analyzed for missile sites or airfields.

The interaction of target density and location, which showed higher performance for construction sites on the outer portion of the imagery, could not be explained from the data. Target location was generally ineffective as a predictor. At least one significant interaction was found for every target type and every target descriptor variable.

Scene type was not analyzed in conjunction with target type. Performance was significantly better on rural than urban or mixed urban and rural scenes. Part of this effect was due to the fact that 10 of 14 missile sites and 15 of 20 airfields were found on rural scenes and part was probably due to the fact that 45 of 54 targets found on rural scenes were classified as low complexity. The question of whether scene type or complexity is the primary factor could not be resolved from the data.

Performance differences between the naive subjects and the experienced were statistically significant (.01 level) but small. None of the interactions of subject experience with other variables in the Chi Square analyses were statistically significant. This indicates that naive and experienced subjects generally reacted the same way to the variables which affected interpretation difficulty.

Accuracy

Accuracy, the percentage of responses given which were correct, averaged 75.4% for the pool of 212 items. False alarms were the primary source of error, 23.2% of all responses were false alarms. Only 1.4% of the responses made were scored as misidentifications.

Of a total of 863 wrong responses, 47 were misidentifications. The proportion of misidentification to right responses for each target type ranged from .002 to .032.

A Chi Square test showed these proportions were significantly different at the .01 level. The lack of similarity among target types suggests that many of these errors resulted from faulty recording rather than identification.

The proportion of false alarms to correct responses for each target were:

1) Transformer Yards–	.142
2) Construction Sites–	.224
3) Electronic Sites–	.415
4) Airfields–	.189
5) Missile Sites–	.031

Target type had a significant effect on false alarm rate.

Experienced subjects were more accurate than naive subjects. Mean accuracy was 80.5% for experienced subjects, 74.1% for naive subjects. Differences in misidentification rates were slight; 1.1% for experienced and 1.5% for naive subjects. Naive subjects did show a significantly higher false alarm rate: 24.5% of their responses were false alarms versus 18.4% for experienced subjects.

False alarms rates were also examined to determine the effects of scene type and density. Scene type/density had a significant (.01) effect on false alarm rates as shown in Figure 4. Greater proportions of responses were false alarms on low density urban and mixed scene types. These results indicate that subjects provided responses even when no, or fewer, targets were present. Results also suggest that there were a greater number of false alarm-producing objects in urban and mixed urban/rural scenes than in rural scenes.

Time and Order

A record was maintained of the time required by each subject to complete each frame of imagery. These time data were analyzed to determine if any relationship existed between time spent on a frame and performance (completeness and accuracy). No such relationship was found.

Data on the order in which imagery was viewed were also analyzed to determine if performance varied over the six to eight hour testing session. The only effect of order was a general reduction in the amount of time spent per image. There was no corresponding variation in completeness or accuracy scores.

Finally, data concerning the order in which targets on a given frame were reported were reviewed. The objective was to determine if some targets consistently "stood out" and thus were reported first. It was found that, because of the low overall performance, insufficient data were available to perform an analysis.

Conclusions

Results of the test development effort provided only two parallel forms of a search and identification performance measure. The goal of three equivalent forms was not met. This failure was due, at least in part, to unanticipated low scores for three of the target types tested. Pretesting had shown higher scores for these targets; lowering of scores in the actual test may have been due to subject, image quality or target difficulty differences. Reduced image quality, resulting from two contact printing operations, appears the most likely cause.

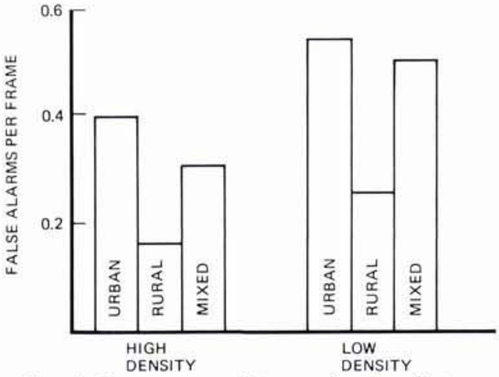

Fig. 4. Scene type and target density effects on false alarm rate.

Reliability of the two parallel forms was found sufficient to discriminate small performance differences. Figure 5 shows test sensitivity as a function of number of subjects. Results were considerably better than the design goal of the study (10% difference with 20 subjects). Improvement in test sensitivity begins to level off with about 10 subjects. Even with 5 subjects, performance differences of 10% can be detected at the 95% confidence level.

The effort also produced a considerable amount of data on factors affecting test item difficulty. Aside from target type, measures of scene complexity had the greatest effect on performance. Scene type (urban, rural, or mixed) and background complexity tended to measure the same condition (or situation); low complexity backgrounds were seldom found in urban scenes. In either case, the results indicate that the subjects had difficulty in separating targets from a complex background. The lowering of performance found with high density target scenes suggests that subjects discontinued their search after finding one or two targets, even though more might have been present.

Target size, location and contrast produced less of an effect than the complexity variables. The location variable was confounded with a resolution difference and thus its effect cannot be attributed to a search pattern deficiency, although such a deficiency may have existed. Resolution in the center of the image was approximately 20% higher than at the edges. The target size and contrast variables produced major performance effects only on certain targets. The lack of effect for other targets appears to have been largely due to the lack of a strong distinction along the dimensions of the variable. In any case, the small performance effect exerted by these variables (4%) indicates that it is not necessary to limit application of the performance measure.

One objective of the program was to determine whether or not naive instead of experienced subjects could be safely used in subsequent studies. Although there was a small and significant experience effect, none of the experience interactions with target type and image difficulty were significant. It thus appears safe to conclude that, in tasks where subjects are required to search for relatively distinctive targets, briefly trained,

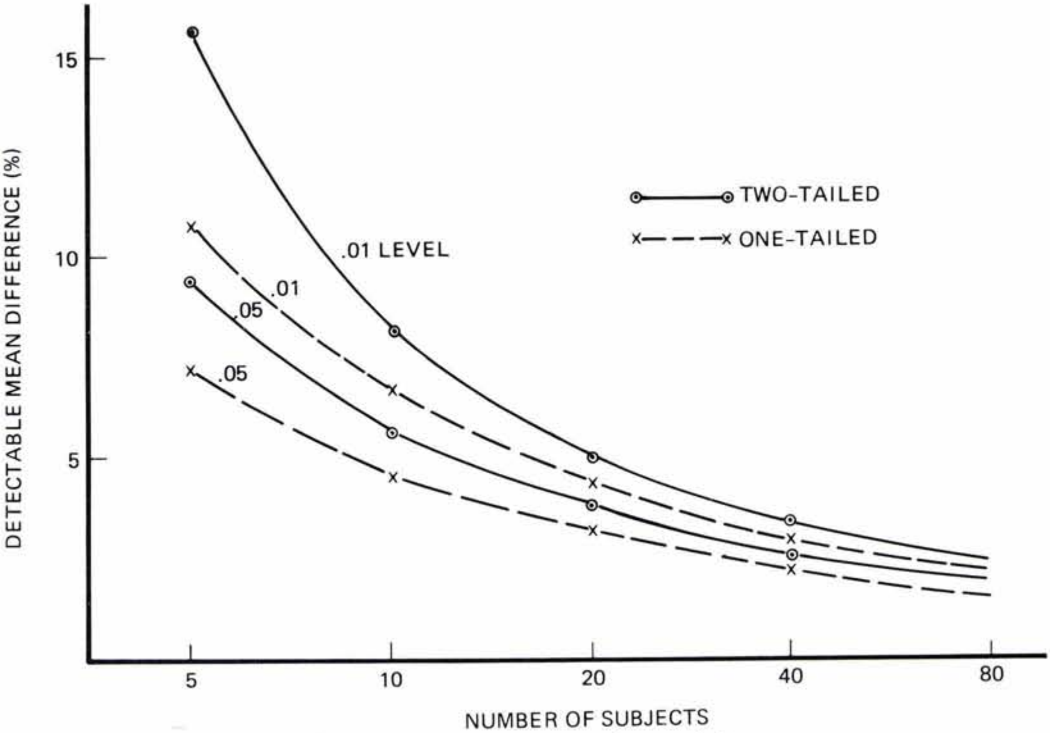

FIG. 5. Test sensitivity using one and two-tailed t test.

naive subjects can do about as well as experienced subjects. This conclusion, however, must be qualified in three ways:

1) Experienced subjects used in the study were not all working P.I.'s. They were probably representative of an experienced group to be found in a large aerospace firm but not in a working P.I. organization.
2) The experienced subjects did not necessarily have any experience in searching for the targets used in the study. Most probably did not.
3) The naive subjects were volunteers, their motivation was thus presumed to be relatively high.

Under these conditions, results indicate the feasibility of developing parallel forms of a search and identification test for use with naive subjects.

The test which has been developed will be utilized in forthcoming studies of displays and image quality, and will provide much greater sensitivity than that found using other experimental approaches. It is estimated that the availability of the test will directly save 500-700 man hours each time it is used. Preparation costs will be recovered after three applications.

ACKNOWLEDGEMENTS

This study was supported by the USAF Independent Research and Development program. The author wishes to acknowledge the helpful guidance and assistance provided by Dr. G. D. Greer, Jr., Dr. S. J. Briggs, and Mr. G. P. Boucek.

REFERENCES

1. Leachtenauer, J. and Bewley, W., *Development of a Photo Interpretation Measure—Test Plan,* D2-114290-1, Boeing Company, Seattle, Washington, 1968.
2. Guilford, J. P., *Psychometric Methods,* McGraw-Hill Book Company, New York, N.Y. 1954.
3. Weiner, B. J., *Statistical Principles in Experimental Design,* McGraw-Hill Book Company, New York, N.Y., 1962.
4. Gulliksen, H., *Theory of Mental Tests,* Wiley, New York, N.Y., 1950.

Meetings Schedule

Annual Conventions

March 1974* Chase-Park Plaza, St. Louis, Mo.

March 7–12, 1975* Washington Hilton, Washington, D.C.

Fall Technical Meetings

Sept. 8–13, 1974,† ISP Comm. V Symposium Washington Hilton, Washington, D.C.

1975,* (open), Phoenix, Arizona.

Sept. 28–Oct. 1, 1976,* Olympic Hotel, Seattle, Wash.; C. E. Buckner, 803 Seattle Municipal Bldg., Seattle, Wash. 98104.

Oct. 18–21, 1977, Little Rock, Arkansas.

International Meetings

Sept. 9–16, 1974, Washington Hilton, Washington, D.C., *14th Congress of the International Federation of Surveyors,* (FIG); Jeter P. Battley, Jr., P.O. Box 14262, Washington, D.C. 20044.

* Jointly with the American Congress of Surveying and mapping.

† To be held as part of the international Congress of FIG.