

AD-A053 871

NAVAL POSTGRADUATE SCHOOL MONTEREY CALIF
A CULTURE-FREE PERFORMANCE TEST OF LEARNING APTITUDE. (U)
FEB 78 J K ARIMA
NPS54-78-2

F/G 5/10

UNCLASSIFIED

NL

OF |
AD
A053871



2

NPS 54-78-2

NAVAL POSTGRADUATE SCHOOL

Monterey, California

AD A 053871



DDC
RECEIVED
MAY 12 1978
B

AD No. _____
DDC FILE COPY

A CULTURE-FREE PERFORMANCE TEST
OF LEARNING APTITUDE

by

James K. Arima

February 1978

Approved for public release; distribution unlimited.

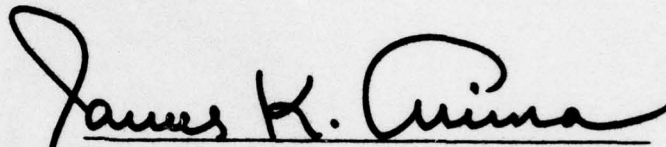
NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral Tyler Dedman
Superintendent

Jack R. Borsting
Provost


Reproduction of all or part of this report is authorized.

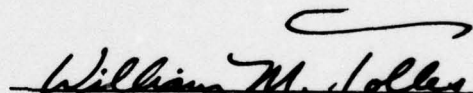
This report was prepared by:


James K. Arima, Associate Professor
Department of Administrative Sciences

Reviewed by:

Released by:


C. R. Jones, Chairman
Department of Administrative
Sciences


W. M. Tolles, Acting Dean of Research

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS 54-78-2	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A CULTURE-FREE PERFORMANCE TEST OF LEARNING APTITUDE	5. TYPE OF REPORT & PERIOD COVERED Technical Report - Final <i>rept log</i>	
6. AUTHOR(s) James K./Arima	7. PERFORMING ORG. REPORT NUMBER	
8. CONTRACT OR GRANT NUMBER(s)	9. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS	
10. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940	11. REPORT DATE Feb 1978	
12. CONTROLLING OFFICE NAME AND ADDRESS None	13. NUMBER OF PAGES 38 P.	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) None	15. SECURITY CLASS. (of this report) Unclassified	
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE		
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Intelligence Performance test Culture-fair Self-paced Learning Aptitude Navy recruits Testing Culture-free Minorities		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) A prototype performance test of intellectual capability (learning ability) was created and given a field trial. Test materials and procedures were designed to be culture-free as possible. Six pairs of random polygons were used as stimuli in a two-choice, multiple discrimination learning paradigm. Variables were racial group (white, nonwhite) and pacing mode (self-paced, machine-paced). Subjects were 121 white and 39 nonwhite male Navy recruits. Over 10 trials (approximately 6 min. of testing time), a learning effect with internal		

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

251450

hc

cont
L
Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

(split-half) reliability of .84 was obtained. White performance was superior to nonwhite only in the machine-paced mode. Significant correlation between the learning rate and Navy General Classification Test scores occurred only for the white group when the sample was divided by race. These results provide considerable encouragement toward the development of a reliable, culture-free test of general learning ability that is very practical and time-efficient.

ACCESSION for	
NTIS	White Section <input checked="" type="checkbox"/>
DDC	Butt Section <input type="checkbox"/>
UNANNOUNCED	<input type="checkbox"/>
JUSTIFICATION _____	
BY _____	
DISTRIBUTION/AVAILABILITY CODES	
Dist. Avail. and/or SPECIAL	
A	

unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

TABLE OF CONTENTS

PREFACE AND ACKNOWLEDGMENTS	iii
INTRODUCTION	1
TEST DEVELOPMENT	3
THE MODEL	3
STIMULUS MATERIALS	5
CONSTRUCTION OF TEST LISTS	7
TEST APPARATUS	12
TRIAL ADMINISTRATION	14
METHOD	14
RESULTS	16
DISCUSSION	25
CONSTRUCT VALIDITY	25
CULTURAL IMPLICATIONS	26
TEST AND TESTING CONSIDERATIONS	27
SUMMARY AND CONCLUSIONS	28
REFERENCES	30

FIGURES

FIGURE 1. Shapes selected for use in assembling stimulus lists	6
FIGURE 2. Stimulus List I.	9
FIGURE 3. Stimulus List II.	10
FIGURE 4. Stimulus List III.	11
FIGURE 5. Layout of Test Equipment	13
FIGURE 6. Information Processing Rate by Test Group and Blocks of Trials	18
FIGURE 7. Information Processing Rate by Racial Group, Pacing Mode, and Blocks of Trials	19

TABLES

TABLE 1. Test Design	15
TABLE 2. Information Processing Rate in Multiple Discrimination Learning by Test Group, Blocks of Trials, and Racial Group	17
TABLE 3. Analysis of Variance of Overall Performance by Test Group, Racial Group, and Blocks of Trials	20
TABLE 4. Analysis of Variance of Overall Performance by Racial Group and Pacing Method	22
TABLE 5. Analysis of Variance of Overall Performance by Racial Group and Stimulus Set (Machine - Paced Only)	22
TABLE 6. Split-Half Reliability Coefficients	23
TABLE 7. Correlations of Test Performance (IPR) with Navy General Classification Test (GCT) Score	24

PREFACE AND ACKNOWLEDGMENTS

This was an independent research project of the author that was not supported financially by sources outside of the Naval Postgraduate School. I am indebted to LT Peter A. Young, USN, for his extensive help on the project as a part of his master's thesis in Operations Research and to Paul Sparks of the Man-Machine Systems Design Laboratory for creating the instrumentation. The Navy Personnel Research and Development Center (NPRDC) was most helpful in permitting us to satellite this study on one of their ongoing projects at the Navy Recruit Training Center, San Diego, in order to obtain the necessary subjects. It would seem, however, that the NPRDC was adequately rewarded for its help with the assignment of LT Young to the Center subsequent to his matriculation at the Naval Postgraduate School. This, of course, is just another example of the close relationship that exists between the course work at the Naval Postgraduate School and the needs of the Navy for professionally qualified officers.

I am also grateful to Dr. Malcolm D. Arnoult, Texas Christian University, for providing me the original prints of the random forms that were used in this research.

Portions of this paper were presented at the 19th Annual Conference of the Military Testing Association held in San Antonio, Texas, 17-21 October 1977. This paper is scheduled for presentation at the 1978 Annual Meeting of the Western Psychological Association, San Francisco, California, 19-22 April 1978.

A CULTURE-FREE PERFORMANCE TEST OF LEARNING APTITUDE¹

James K. Arima
Naval Postgraduate School

From World War I to the late 1950s, standardized mental tests with nationally based norms became widely used for selection, placement, and classification decisions. Their great acceptance was due, in large part, to their role in furthering the American concept of an egalitarian society (Holzman, 1971). That is, decisions of considerable importance to individuals could be made on the basis of merit, given a person's score on an objective test of ability with the requisite reliability and validity.

The Armed Services were leaders in the testing movement, and the use of the Army Alpha and Beta tests in World War I has been identified with the beginning of the testing movement in which large numbers of persons are routinely tested for selection and placement. Nearly two million people were given the tests during the course of the war, and the results provided much of the information for later studies of demographic, socioeconomic, and cultural differences in intelligence and ability (Matarazzo, 1972). World War II saw a similar emphasis on mass testing and the development of the Army General Classification Test (Melton, 1957). Again, the results of the testing program provided large amounts of valuable information for scientific study that went far beyond the limited purposes for which tests were originally administered. Eventually, the AGCT was made available in commercial form for sale to qualified users in the general public.

In the post-World War II years, the Armed Forces Qualification Test (AFQT) with a scoring in readily understandable percentiles became the standard, general test of mental ability for the services. The AFQT designation of mental categories is still in use today. Throughout these developments, special-purpose tests were also being created by the individual services until a common entrance test was no longer the rule with the advent of the All Volunteer Force (Melton, 1957; Windle and Vallance, 1964). More recently, however, an emphasis on efficiency in the testing program on the part of Congress and the Defense Secretariat has seen the emergence of the Armed Services Vocational Aptitude Battery (ASVAB) as a common test of general aptitude for military service. A form of the ASVAB is also used in civilian, secondary schools in the High School Testing Program managed by the Armed Forces Vocational Testing Group (AFVTG).

¹I am indebted to Peter A. Young for running the subjects and collecting and analyzing the data as a part of his master's thesis (Young, 1975). Paul Sparks created the instrumentation for the experimental administration of the test. The terms culture-free and culture-fair will be used to mean the same thing indiscriminately.

The growth and apparent success of the testing movement has not been without its critics and detractors. The criticism did not reach social significance until the middle and late sixties when many of our institutions were put to severe test with a reexamination of our value systems and the emergence of new concepts for improving the quality of life in America. The routine testing of job applicants took a severe setback in the Griggs et al. vs. Duke Power Company decision of the U.S. Supreme Court when it ruled that a test could not be used as a selection device unless the measured abilities represented by the scores on the test were shown to be required for acceptable performance on the job. This decision had at least two implications for testing. One, obviously, related to the traditional concept of the predictive validity of tests, and the other was with respect to the use made of tests.

Regarding the predictive validity of tests, the court's decision was quite telling, since most tests predict intermediate criteria well--such as normatively scored achievement tests--but not more distant, more ultimate criteria, such as occupational success (Goslin, 1968). This situation is particularly prevalent in such large institutions as the military (Thomas, 1972a, 1972b) and the nation's educational systems. The question of the use, or misuse, of tests focuses on the results that testing programs produce. The argument has been that differential prediction or classification of individuals results when they are categorized on the basis of ethnic and socioeconomic backgrounds. Broadly stated, differential prediction means that the proportion of individuals who, for example, pass a selection cutoff score is not the same for the different categorical groups. Such differential prediction has been labeled bias because culturally deprived persons have not had the opportunity to master the material content of the tests nor to develop the test-taking motivation, experience, and specific skills of other groups of persons (Goslin, 1968). The bias is usually attributed to the test, rather than to the uses made of the test, but the argument is not entirely convincing (Green, 1975). Even on a strictly psychometric basis, several different definitions of bias are possible (Hunter, Schmidt, and Rauschenberger, 1977).

While the Armed Services have managed to escape severe criticism in the past, there are signs that the situation is changing. The use of the ASVAB in the High School Testing Program recently received very sharp criticism from Lee J. Cronbach, and the Office of Management and Budget (OMB) has instituted a series of inquiries into the management of their testing programs on the part of the several services.

Complicating the issues of test validity and test usage as sources of bias is the argument with respect to the roles of heredity and environment in the determination of a measured, mental ability--such as intelligence. If, as argued by Jensen (1968a), heredity plays the predominant role by a margin of as much as 2-to-1, then the cultural deprivation argument loses considerable weight. That is, the important

differences exist, more or less, independent of environmental factors. On the other hand, if it is argued that the range of performance capabilities at a fixed hereditary level is broad and essentially unpredictable due to the influence of many environmental factors (Feldman and Lewontin, 1975), then the role of cultural and socioeconomic factors in causing the differential prediction of testing programs must be acknowledged and corrected. A deceptively simple solution would be to create tests that are culture free. Presumably, a culture-free test would be measuring the "real" or hereditary potential--the genotype--of the person being tested. But, if an operational definition of an unbiased, culture-free test is that all categories of cultural groups have the same mean and distribution function on the test, the use of such a test for selection is highly likely to result in differential outcomes on some criterion measure, such as the ability to complete a course of training within a prescribed or reasonable period of time. The test has been made culture free, but it has little or no predictive validity. The argument could be made that the fault lies in the criterion, and not the test. In this case, a third fundamental question regarding the testing movement arises, and that is the construct validity of a test or what is the test supposed to be measuring? (Goslin, 1968).

As explained in the preceding argument, the creation of a culture-free test places a greater burden on the construct validity of the test rather than its predictive validity, since it may not be possible to determine the latter in the traditional manner. In addition to escaping criticism for being biased, a culture-free test of mental ability with high construct validity would be of great value to the military services and other large institutions that face increasingly difficult problems in personnel procurement owing to the shrinking of the pool from which new recruits must be obtained (Congressional Budget Office, 1977). Under these circumstances, if standards are not to be lowered, means must be found to identify individuals with high native ability who do not score well on traditional tests. It was the purpose of this project to explore the possibility of developing such a test that was relatively culture-free, had high construct validity with respect to identifying individuals of high native ability, and would be feasible and practical to administer in the military testing environment.

TEST DEVELOPMENT

THE MODEL

The first problem in developing the test was to find a model upon which to build the test. A model, in this usage, is a procedure or paradigm that reliably elicits for quantitative measurement a behavior that is the result of a cognitive process that is frequently involved in many situations in real life. Models of this sort would be available in such traditional experimental areas as learning and memory, information processing, problem solving, and decision making. It was felt that most of the paradigms for information processing placed an overly high

emphasis on verbal behavior and materials and that this feature would make it difficult to achieve a culture-free test. The problem-solving paradigm was thought to be inappropriate for test construction from a reliability and measurement standpoint, since an attempt to control and standardize the set or approach an individual takes would tend to destroy the objectives of the paradigm, itself, which encourages experimentation by the subject. Also, the frequency of chance or "aha" solutions would tend to make test scoring difficult, categorical, and unreliable. The decision-making paradigm was not considered appropriate because of the paradigm's reliance on value systems in the elicited behavior--value systems developed through life experiences and very much the product of an individual's culture.

This left the area of learning as a logical choice for the model. Learning paradigms have been the traditional vehicle of the majority of research in the behavioristic tradition, and learning ability is generally recognized as an important ingredient in an individual's adaptation to a job. In the industrial engineer's armamentarium, the "learning curve" is an important ingredient for an entire production process. There are many reliable measures of the learning process--at least in the aggregate. And the law of effect, in its empirical form, is without precedence among the many, so-called "laws" in psychology. As quoted and discussed by Estes (1974), Thorndike believed that intellect is the ability to learn and that estimates of intellect should be estimates of the ability to learn. In another sense, Thorndike believed that intellect is the ability to learn more things or to learn the same things more quickly. Typical intelligence tests that sample the products an individual is able to produce seem to be assessing intelligence with respect to the amount of stored information, knowledge, and intellectual skills, whereas the typical experimental learning paradigm would seem to consider the rate of learning as a measure of intellectual performance.

Within the field of learning, visual discrimination learning was selected as the general paradigm in which to build the test because it has been widely used at many phylogenetic levels to study the evolution of intelligence (Bitterman, 1965, 1975). There is also an extensive literature in the visual discrimination learning of human subjects as well (Green and O'Connell, 1969). The typical paradigm for visual discrimination learning involves two or more dissimilar, visual stimuli of which one has been arbitrarily designated as correct. The organism learns to respond to the correct alternative--e.g., peck the middle disc--by being reinforced for making the correct choice.

Examination of the Green and O'Connell (1969) bibliography will show that most of the experimental tasks in visual discrimination learning have been relatively simple owing to the design of such tasks for animals, children, and retardates. The visual discrimination learning situation has been made more complex by manipulating reinforcement contingencies or the quality of reinforcements. In their altered form, emphasis has been on such phenomena as reversal learning, probability learning, and the effects of partial reinforcement and incentive contrasts. Bitterman has shown that the acquisition (learning) curve may

be very similar for all organisms, but the switch to one of the other conditions following original learning has led to qualitatively different behaviors by different species. Thus, it would be highly desirable to adhere to the basic learning paradigm but make the task more demanding for the human subject. This could be done by having an individual learn several discriminations simultaneously, which shall be called multiple discrimination learning. Except for the fact that pictorial materials would be used, the situation would be very similar to verbal discrimination learning (Eckert and Kanak, 1972). In a typical verbal discrimination learning experiment, a list of several word pairs is created in which one member of each pair has been designated as the correct alternative. The pairs, referred to as items, are presented individually and a complete presentation of the list is a trial. The subject instrumentally learns the correct alternatives by being reinforced when the correct member of the word pair is vocalized. Arima (1974) has shown that the paradigm is very robust in the sense that the learning rate is constant regardless of the number of alternatives (up to four) presented in a stimulus (item) as long as the information presentation rate is also constant. The key to determining this relationship was the measurement of information content in terms of Shannon bits and learning in terms of the information transmission rate.

To recapitulate, the model for the test was a visual discrimination paradigm presented in the manner of verbal discrimination learning experiments. That is, the model calls for the subject to learn several visual discriminations simultaneously, a process that will be referred to as multiple discrimination learning.

STIMULUS MATERIALS

Construction of a multiple discrimination learning test required a relatively large set of stimuli that were homogeneous, yet discriminable, and which were as free of cultural influence or implications as possible. Homogeneity of stimulus materials was desired so that each of the stimulus pairs within a "list" could be of comparable difficulty and so that any stimulus pair would be representative of the test task. Geometric shapes were eliminated because of their limited numbers and the possibility that their familiarity and association values might be linked with cultural variables. Color, hue, and brightness were also rejected because of the difficulty in production and replication and because difficulties in sensory discrimination might result when a large number of items was required. Additionally, there would be the problem of using the test with colorblind individuals. For these reasons, two-dimensional, black-and-white patterns of uniform size were investigated. The set of 30, two-dimensional, random-shaped, metric polygons used by Arnoult (1956) were found to fit the requirements admirably. They are shown in Figure 1. Moreover, they had already been categorized, as a group, as figures having high discriminability.

Prior to constructing pairs and lists of items using the forms, it was necessary to obtain measures of the pairwise similarity of the

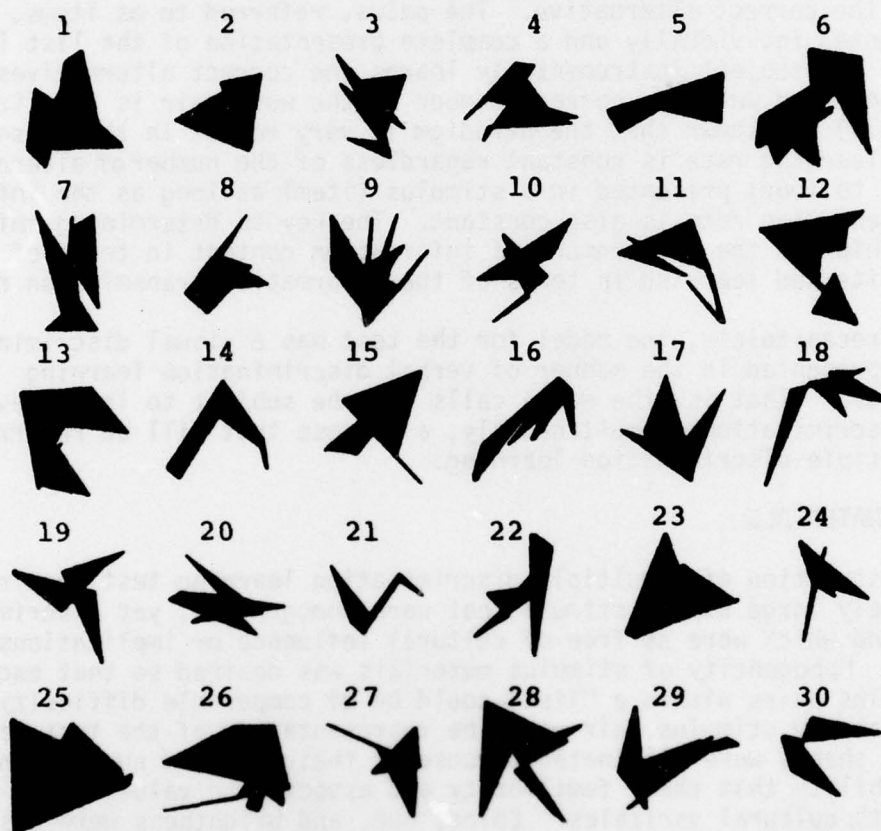


FIGURE 1. Shapes selected for use in assembling stimulus lists.

(I am indebted to Dr. Malcolm D. Arnoult of Texas Christian University for providing me the original prints for this application.)

forms and to develop a set of pairs for which there would be assurance that either member would be likely to be chosen as a correct alternative on a first (guess) trial. It was particularly necessary to develop pairs with an a priori choice of 50-50 for either member so that the information content (uncertainty) of each item would be at a maximum (1 bit) and constant within all lists. The similarity measure was desired because similarity had been found to be a significant variable affecting learning rate in verbal learning under some conditions. Accordingly, it was assumed that similarity among and between the stimuli should be controlled in constructing the test items.

In order to obtain empirical values for these relationships among the forms, a small, data-gathering experiment was conducted. The 30 stimulus polygons were arranged in pairs. All possible pairs were constructed under the constraint that an item would not be paired with itself. Left-right order within a given pair was not considered. This resulted in the assembly of $(30 \times 29)/2 = 435$ different pairings. These pairs were then arranged in three columns on sheets. Three separate booklets, each containing 145 pairs, were constructed and distributed to 60 graduate students at the Naval Postgraduate School. Each subject received a single booklet selected at random from the three, and was asked to perform two separate tasks--selection of one item from each pair and rating of the degree of similarity seen between the items of each pair. Subjects were told that one item in each pair had been arbitrarily designated as "correct," i.e., the desired response, and were asked to designate that item which they thought to be the "correct" response. This selection was to be made with the knowledge that designation of the "correct" response was made completely arbitrarily.

Subjects were cautioned to make their choices solely on the basis of a given pair alone, and without regard to previous selections. This exercise was intended to simulate as closely as possible the condition of facing a stimulus pair in a forced-choice situation with no prior knowledge of the correct item in the pair.

Subjects then went through the list a second time, rating each pair as to whether the two items in each appeared to be very similar, slightly similar, or dissimilar. Each pair was then assigned a similarity factor of one, two, or three, respectively.

The choice preferences of the 60 subjects (20 for each set of 145 pairs) were translated into percentages and cast into a matrix. In addition, averages of similarity ratings given for each pair were computed and cast into the same matrix format. Thus pairwise estimates of choice preference and item similarity were obtained and placed in usable form.

CONSTRUCTION OF TEST LISTS

A subgroup of pairs was selected from the original 435 that had been rated. These pairs were singled out on the basis of choice preference. Subjects making choices within these pairs had displayed no

significant preference, on the average, for either item in each pair (selections were distributed either 50%-50% or 45%-55% between each). This subgroup was then used to construct the test lists. Since no marked preference for a given item in a pair had been demonstrated, it was felt that the choice probabilities associated with each could be considered to be "equally likely" for the purposes of evaluating the information content of the choice associated with each pair.

Three stimulus lists of six pairs each were constructed from the "equally likely" subgroup of pairs. These lists were assembled under the following constraints with respect to the similarity variable:

List I. Figures in each pair were as dissimilar as possible. In addition, all figures in the entire list were as dissimilar as possible. (Within-pair similarity factors were at least 2.50, averaging 2.60, while between-pair factors were not less than 1.75, averaging 1.98.)

List II. Figures in each pair were as similar as possible, but dissimilarity between pairs was maintained. (Within-pair similarity factors were no greater than 1.95, averaging 1.58; the between-pair factors were no less than 1.90, averaging 2.20.)

List III. Figures were as similar as possible, both within each pair and between other figures in the list. (Within pair similarity factor was no more than 1.90, averaging 1.73; between-pair factor was no greater than 2.30, averaging 1.92.)

These lists are presented in Figures 2, 3, and 4, respectively. As can be seen, the lists were constructed in order to present discrimination tasks of increasing difficulty. Stimulus items in List I were chosen to be as distinguishable as possible, minimizing intra- and interpair confusion. Similarity within pairs was added in List II, but each pair was kept as distinguishable as possible from other pairs in the list. Similarity was extended to cover all items in List III. List III, of course, is the most homogeneous.

When lists of six pairs each had been completed, three test lists of 60 pairs were assembled. Each test list consisted of 10 repetitions of each of the six pairs of Lists I, II, and III. Order within these replicates was random. Left-right order within pairs was varied in a random fashion as well with the restriction that a given form was seen on the right five times and on the left five times. At least one different pair was presented before a given pair was repeated. The polygons were not rotated or reversed, but were presented "upright" at all times.

Thus each test subject could be presented a total of 60 pairs of stimuli. Pairs appeared in no apparent order, and the correct response was not always on either the right or left side; subjects were forced to learn the correct response in each pair solely on the basis of recognition of the items within that pair alone.

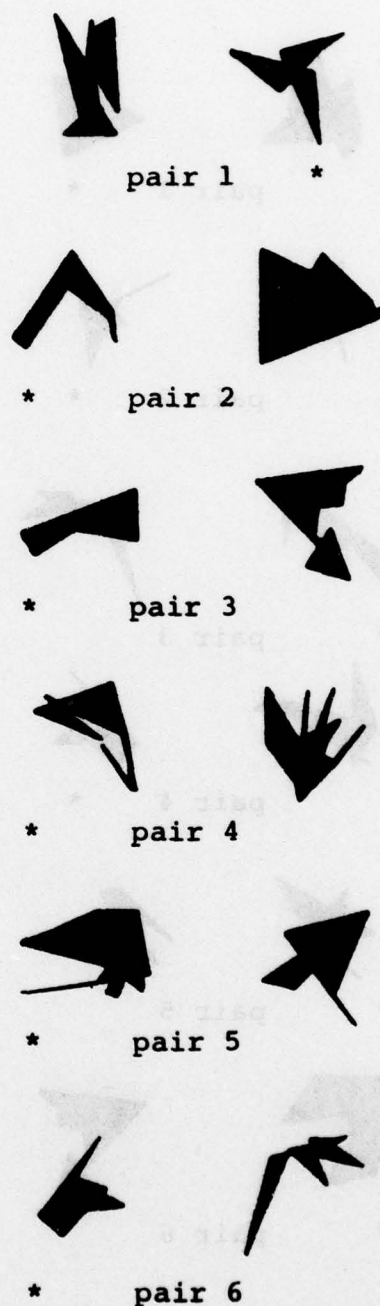


FIGURE 2. Stimulus List I.
(Least similarity within and
between pairs)
*Indicates "correct" shape.

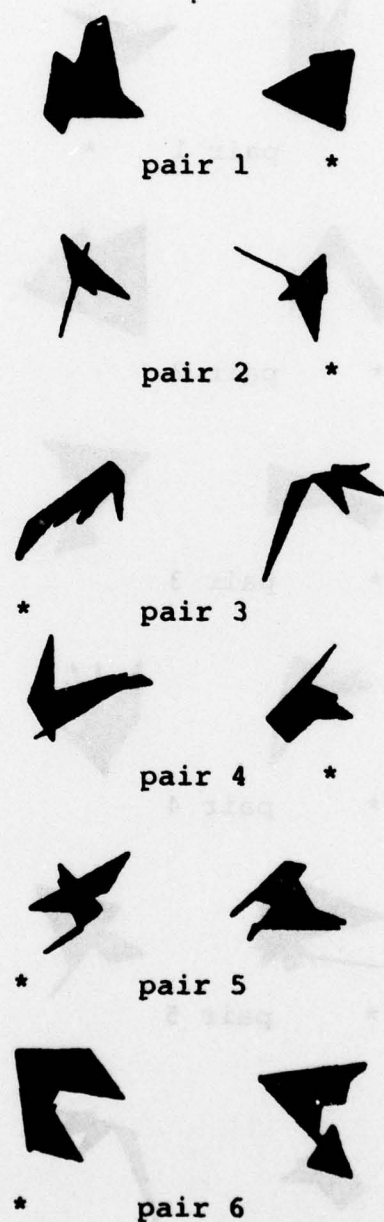


FIGURE 3. Stimulus List II.

(Maximum similarity within pairs; minimum similarity between pairs.)

*Indicates "correct" shape.

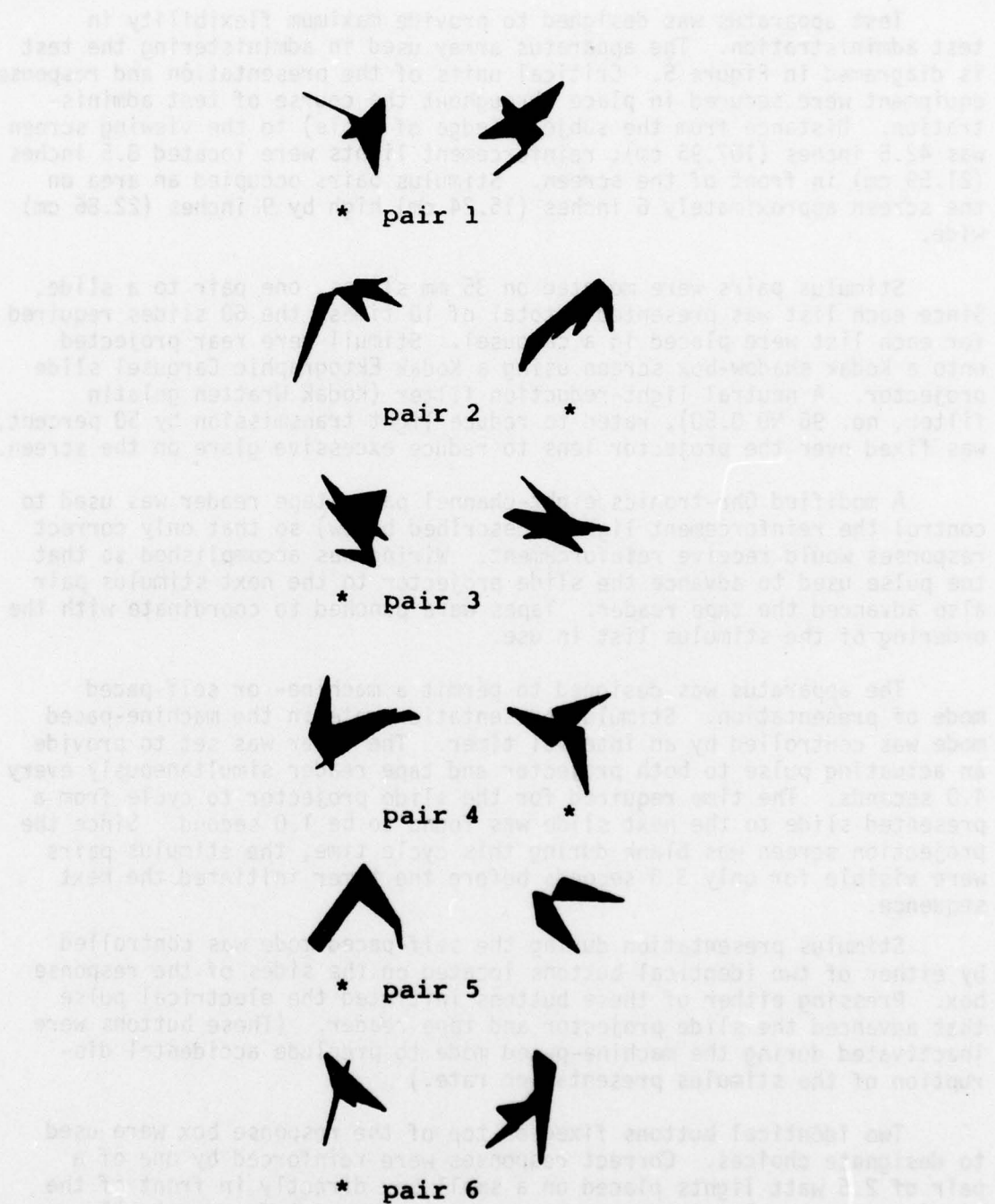


FIGURE 4. Stimulus List III.

(Maximum similarity both within and between pairs.)

***Indicates "correct" shape.**

TEST APPARATUS

Test apparatus was designed to provide maximum flexibility in test administration. The apparatus array used in administering the test is diagramed in Figure 5. Critical units of the presentation and response equipment were secured in place throughout the course of test administration. Distance from the subject (edge of table) to the viewing screen was 42.5 inches (107.95 cm); reinforcement lights were located 8.5 inches (21.59 cm) in front of the screen. Stimulus pairs occupied an area on the screen approximately 6 inches (15.24 cm) high by 9 inches (22.86 cm) wide.

Stimulus pairs were mounted on 35 mm slides, one pair to a slide. Since each list was presented a total of 10 times, the 60 slides required for each list were placed in a carousel. Stimuli were rear projected onto a Kodak shadow-box screen using a Kodak Ektographic Carousel slide projector. A neutral light-reduction filter (Kodak Wratten gelatin filter, no. 96 ND 0.50), rated to reduce light transmission by 50 percent, was fixed over the projector lens to reduce excessive glare on the screen.

A modified Ohr-tronics eight-channel paper-tape reader was used to control the reinforcement lights (described below) so that only correct responses would receive reinforcement. Wiring was accomplished so that the pulse used to advance the slide projector to the next stimulus pair also advanced the tape reader. Tapes were punched to coordinate with the ordering of the stimulus list in use.

The apparatus was designed to permit a machine- or self-paced mode of presentation. Stimulus presentation rate in the machine-paced mode was controlled by an interval timer. The timer was set to provide an actuating pulse to both projector and tape reader simultaneously every 4.0 seconds. The time required for the slide projector to cycle from a presented slide to the next slide was found to be 1.0 second. Since the projection screen was blank during this cycle time, the stimulus pairs were visible for only 3.0 seconds before the timer initiated the next sequence.

Stimulus presentation during the self-paced mode was controlled by either of two identical buttons located on the sides of the response box. Pressing either of these buttons initiated the electrical pulse that advanced the slide projector and tape reader. (These buttons were inactivated during the machine-paced mode to preclude accidental disruption of the stimulus presentation rate.)

Two identical buttons fixed on top of the response box were used to designate choices. Correct responses were reinforced by one of a pair of 2.5 watt lights placed on a small box directly in front of the viewing screen. Incorrect responses received no reinforcement. Responses, regardless of reinforcement, were recorded on a two-channel Clevite brush recorder. The tapes thus obtained could be used to confirm observed responses, and in the self-paced mode to measure inter-response time and total test time.

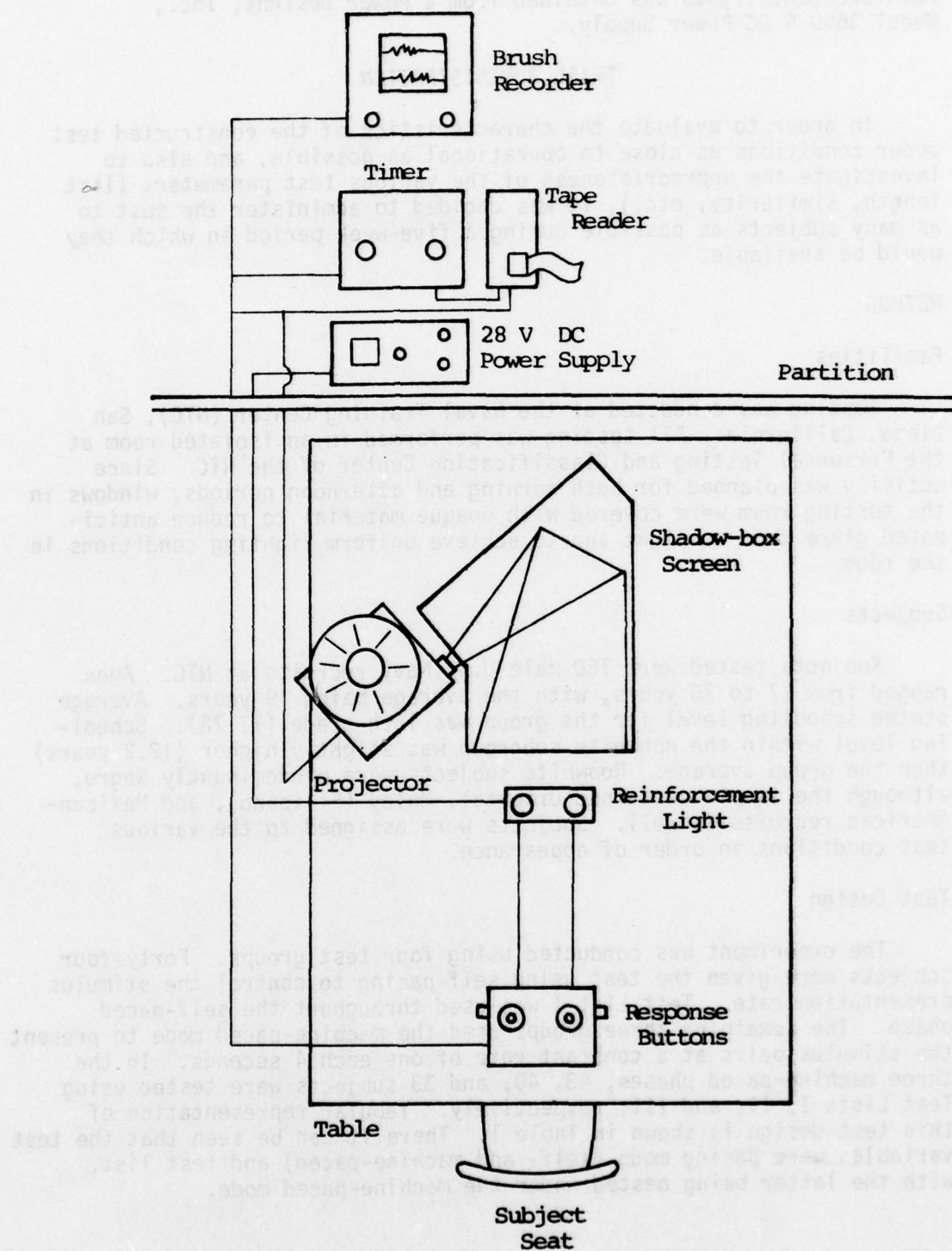


FIGURE 5. Layout of Test Equipment

Twenty-eight volt DC current to power the tape reader and reinforcement lights was obtained from a Power Designs, Inc., Model 3650-S DC Power Supply.

TRIAL ADMINISTRATION

In order to evaluate the characteristics of the constructed test under conditions as close to operational as possible, and also to investigate the appropriateness of the various test parameters (list length, similarity, etc.), it was decided to administer the test to as many subjects as possible during a five-week period in which they would be available.

METHOD

Facilities

Testing was conducted at the Naval Training Center (NTC), San Diego, California. All testing was performed in an isolated room at the Personnel Testing and Classification Center of the NTC. Since activity was planned for both morning and afternoon periods, windows in the testing room were covered with opaque material to reduce anticipated glare from sunlight and to achieve uniform lighting conditions in the room.

Subjects

Subjects tested were 160 male U.S. Navy recruits at NTC. Ages ranged from 17 to 26 years, with the average being 19 years. Average stated schooling level for the group was 12th grade (11.78). Schooling level within the nonwhite subgroup was slightly higher (12.2 years) than the group average. Nonwhite subjects were predominantly Negro, although the sample contained Oriental, Malay (Filipino), and Mexican-American recruits as well. Subjects were assigned to the various test conditions in order of appearance.

Test Design

The experiment was conducted using four test groups. Forty-four subjects were given the test using self-pacing to control the stimulus presentation rate. Test List I was used throughout the self-paced phase. The remaining three groups used the machine-paced mode to present the stimulus pairs at a constant rate of one each 4 seconds. In the three machine-paced phases, 43, 40, and 33 subjects were tested using Test Lists I, II, and III, respectively. Tabular representation of this test design is shown in Table 1. There it can be seen that the test variables were pacing mode (self- and machine-paced) and test list, with the latter being nested under the machine-paced mode.

Table 1.

Test Design

Test Group	Subjects (White; Nonwhite)	Pacing	Stimulus List
1	44 (31; 13)	Self	I
2	43 (30; 13)	Machine	I
3	40 (31; 9)	Machine	II
4	33 (29; 4)	Machine	III

Procedure

Subjects were brought into the testing room in groups of not more than six. The apparatus was displayed, and the experimental nature of the testing explained briefly prior to issuing the verbal instructions. Instructions emphasized the nature of the stimuli, what was required of the subject in the way of response, and the operation of the apparatus itself. Subjects were then given the opportunity to ask questions about the test and procedure, and to decline participation if they so desired. They were then asked to wait outside the room and were brought in for testing one by one. The instructions for the test were then reviewed with each individual as he was seated at the response box prior to commencement of the experiment.

Stimulus pairs were then presented one by one on the viewing screen for his test condition. Each group of six pairs was presented in 10 consecutive trials with no break between groups. As a subject selected the figure in each pair that he thought was correct, he pressed the corresponding (right or left) response button in front of him. Correct responses were reinforced by a small light in front of the view screen, while incorrect responses received no reinforcement.

As testing was in progress, the experimenter stood behind the subject and recorded his responses on an answer sheet. Responses were also recorded electrically on a two-channel Brush recorder. Upon completion of the test, the subject was cautioned not to discuss anything he had seen or done in the test with those who had not yet been tested. This request was repeated to the entire group after all had been through the test.

Performances by six of the original 160 subjects were discarded. Improper operation of the self-pacing buttons that put the tape reader out of phase with the projector was cause for rejection of three performances. Another subject in the first (self-paced) group was unable to follow instructions. Timer malfunction caused two performances in the first machine-paced group to be eliminated.

Seventeen other subjects' performances were not used in the data analysis because of their Navy Basic Test Battery (BTB) scores and/or demographic data could not be retrieved from computerized records. As a result of these subject losses, the 137 remaining subjects (white and nonwhite) were distributed as follows: Group 1 (24, 11); Group 2 (25, 12); Group 3 (28, 8); and Group 4 (30, 3).

RESULTS

Individual performances in the test, in the form of number of correct choices made per trial per unit of time, were computed to arrive at the test measure of effectiveness, Information Processing Rate (IPR). Specifically, IPR was defined as bits of information correctly processed per second. Performances in the first trial were not used, since responses in the initial trial were dependent wholly upon chance, and as such were not indicative of learning ability.

The number correct in each trial was divided by the amount of time the stimuli were presented to the subject. (In the machine-paced mode, this was a constant 3 seconds per pair. Scores for the self-paced group were scaled to individual rates.) In both situations, the 1-sec. cycle time (inter-stimulus time) of the slide projector was not included in computing IPR. The resultant trial IPR scores were grouped into three blocks of three consecutive trials each. These figures are listed in Table 2. Rates of processing information are seen to generally increase over blocks of trials for all groups. (The single exception is the nonwhite subset of Test Group 4, where performance declines very slightly over trials. This group contained three subjects.) Overall performances by all groups were quite similar, despite differences in pacing mode and stimulus similarity between groups. Overall performance by the nonwhites in Test Group 1 (self-paced) exceeded that of the whites; the reverse was true for the three machine-paced groups. Figures 5 and 6 depict aspects of these situations.

The results listed in Table 2 were subjected to an analysis of variance using a three-way design compensating for unequal cell populations by test group, racial group, and blocks of trials as described by Kirk (1968). The results of this analysis are presented in Table 3. Significant effects were noted between racial groups and among blocks of trials. The blocks effect is important from the construct validity standpoint in demonstrating that learning did occur over all conditions of the experiment. It should also be noted that pacing mode and

Table 2

Information Processing Rate in Multiple Discrimination Learning
by Test Group, Blocks of Trials, and Racial Group

Test Group	Block 1		Block 2		Block 3		Totals		
	White	Nonwhite	White	Nonwhite	White	Nonwhite	White	Nonwhite	Combined
1	181.416	198.000	210.666	227.818	251.958	235.636	214.291	220.484	216.580
2	202.520	156.916	261.640	193.916	284.000	237.333	249.386	196.055	232.089
3	190.071	205.500	240.892	207.750	266.607	212.500	233.713	228.129	232.472
4	187.418	222.000	247.074	215.666	262.814	214.333	231.962	230.533	231.819
<u>Totals</u>									
White	190.009	—	240.509	—	267.461	—	232.659	—	—
Nonwhite	—	187.382	—	210.058	—	228.999	—	208.813	—
Combined	189.361		233.000		257.984				226.783

Note. Entries are bits/sec x 10³.

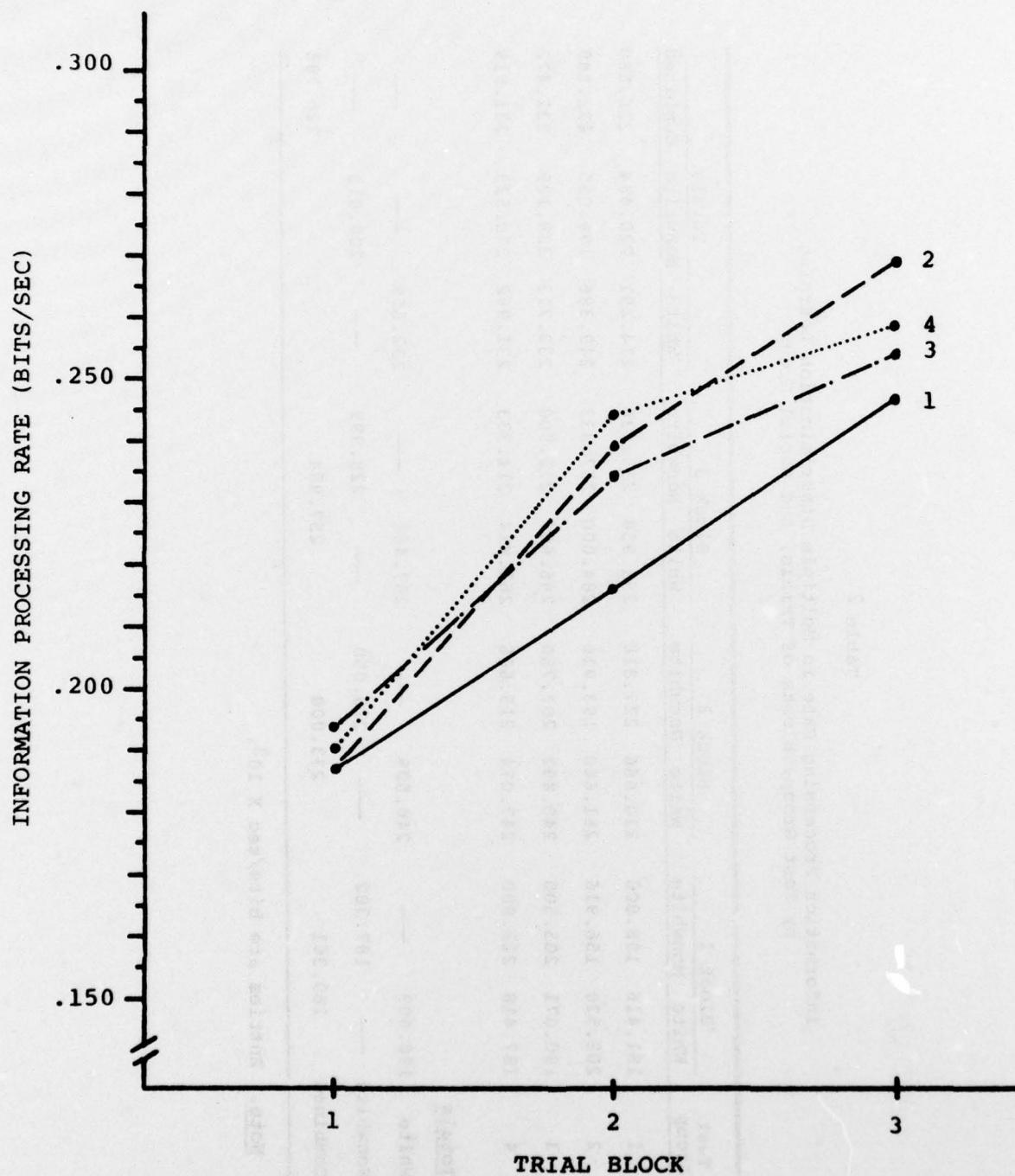


FIGURE 6. Information Processing Rate by Test Group and Blocks of Trials.

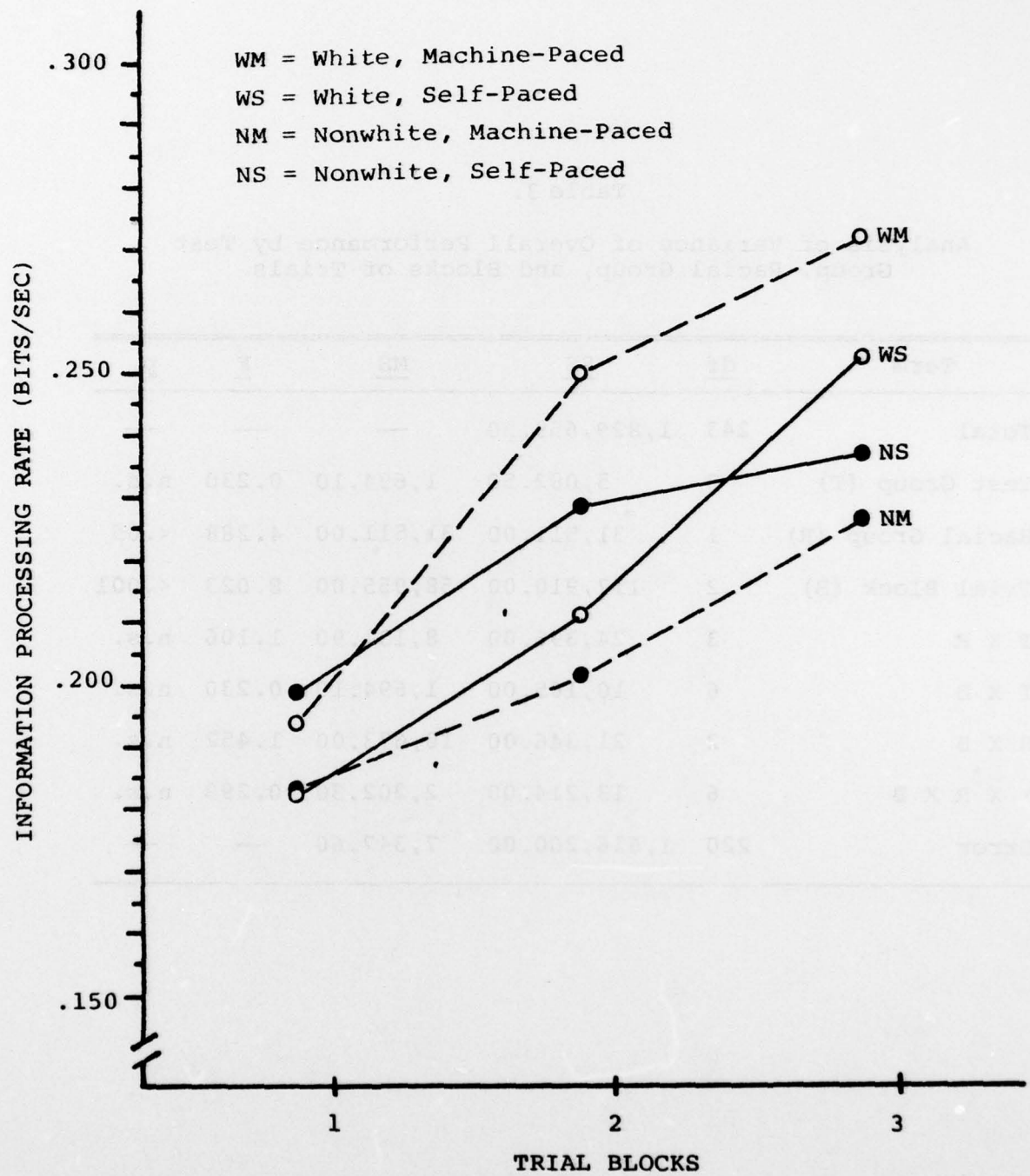


FIGURE 7. Information Processing Rate by Racial Group, Pacing Mode, and Blocks of Trials.

Table 3.

Analysis of Variance of Overall Performance by Test
Group, Racial Group, and Blocks of Trials

Term	<u>df</u>	<u>SS</u>	<u>MS</u>	<u>F</u>	<u>p</u>
Total	243	1,829,659.50	—	—	—
Test Group (T)	3	5,082.50	1,694.10	0.230	n.s.
Racial Group (R)	1	31,511.00	31,511.00	4.288	<.05
Trial Block (B)	2	117,910.00	58,955.00	8.023	<.001
T X R	3	24,396.00	8,131.90	1.106	n.s.
T X B	6	10,165.00	1,694.10	0.230	n.s.
R X B	2	21,346.00	10,673.00	1.452	n.s.
T X R X B	6	13,214.00	2,202.30	0.299	n.s.
Error	220	1,616,200.00	7,347.60	—	—

similarity were confounded in the test group variable in this analysis, but had the primary effects of either of these variables been substantial, the analysis would have resulted in a significant F for the test group variable. On the other hand, if the effects of both variables had been substantial, the effects on the test group variable would have been indeterminate because of the possibility that the effects of one might cancel the effects of the other.

In order to assess the effects of pacing mode, an analysis of variance was conducted using the total IPR as the dependent measure and racial group and pacing as the independent variables. Racial group was included in the analysis because of the possible interactive effect with the pacing variable, as suggested in Figure 6. With the data collapsed over blocks of trials, the racial variable was not significant (Table 4). The pacing effect was not significant and the hypothesized interactive effect attained a F value that was between the .10 and .20 levels of probability.

In order to assess a possible similarity effect, an analysis of variance was conducted using the total IPR as the dependent measure and racial group and similarity (stimulus set) as the independent variables. Only the machine-paced test groups were used for this analysis. The results, shown in Table 5, found racial group to be significant at less than the 2 percent level of probability, while similarity and the interaction term were not statistically significant. In addition to the implications for the similarity variable, the comparative analysis provided by tables 4 and 5 with respect to race indicate that race did have a significant effect when the subjects were machine-paced but not when they were allowed to pace themselves.

Finally, in order to confirm that subjects showed a significant difference in their learning rates, as one would expect from the sizable error terms in all of the preceding analyses, several analysis of variance tests were conducted using a repeated measures design with subjects and blocks of trials as the independent variables and the interaction of these two effects as the error term. The dependent variable was the IPR per subject per block. Four such tests were conducted by partitioning the total sample by race and pacing mode. The F ratios were all highly significant for subjects and blocks of trials with most of them at the .001 level of probability.

Internal reliability of the test itself was investigated using a split-half design for each test group and each racial group as well as for overall performances. Processing rates were compared for trials 4, 6, and 8 against those of trials 5, 7, and 9. In addition, scores on the latter group of trials were compared with those obtained on trials 6, 8, and 10. The former comparison will be referred to as "low trials" and the latter, as "high trials."

Correlation coefficients thus obtained were used in the Spearman-Brown formula for split-half correlations. Both the raw coefficients

Table 4

Analysis of Variance of Overall Performance
by Racial Group and Pacing Method

Term	df	SS	MS	F	p
Total	137	465,094.994	—	—	—
Racial Grp (R)	1	4,417.475	4,417.475	1.310	n.s.
Pacing Mode (P)	1	242.501	242.501	0.071	n.s.
R X P	1	8,772.961	8,772.961	2.602	n.s.
Error	134	451,662.057	3,370.612	—	—

Table 5

Analysis of Variance of Overall Performance
by Racial Group and Stimulus Set
(Machine - Paced Only)

Term	df	SS	MS	F	p
Total	102	5,316.928	—	—	—
Racial Grp (R)	1	342.169	342.169	6.810	.020
Stimulus Set (S)	2	4.758	2.379	0.047	n.s.
R X S	2	96.117	48.058	0.956	n.s.
Error	97	4,873.884	50.246	—	—

Table 6
Split-Half Reliability Coefficients

Group	Low Trials (468 vs 579)		High Trials (579 vs 6810)		Totals		
	<u>r</u> (raw)	<u>r</u> (S-B)	<u>r</u> (raw)	<u>r</u> (S-B)	Low	High	
1	White	.767	.868**	.713	.832**	.865**	.872**
	Nonwhite	.756	.861**	.864	.927**		
2	White	.800	.889**	.865	.928**	.871**	.921**
	Nonwhite	.700	.824**	.826	.905**		
3	White	.615	.762**	.632	.775**	.722**	.759**
	Nonwhite	.367	.537	.535	.697		
4	White	.674	.805**	.664	.798**	.802**	.794**
	Nonwhite	.637	.778	.610	.758		
Totals							
	White		.835**		.843**		
	Nonwhite		.788**		.873**		
	Combined		.824**		.851**	.838**	

*Significant at $p < .05$.

**Significant at $p < .01$.

Note: Significance is based on the raw correlations.

and the Spearman-Brown coefficients are listed in Table 6. A majority of the coefficients are seen to be statistically significant.

The relationship between scores on the experimental test and the traditional methods of measuring Navy recruit potential was investigated using the test subjects' scores on the Navy General Classification Test (GCT), a major portion of the standard Basic Test Battery (BTB). The basis for the GCT lies in verbal ability, since the test consists of sentence completions and verbal analogies. Test scores are scaled on a normalized distribution with a mean of 50 and a standard deviation of 10. Performance on the Arithmetic Reasoning Test (ARI) is often combined with GCT scores to obtain a rough "multiple" used in determining Navy technical school eligibility and aptitude.

Pearson product-moment correlations were computed between test scores and GCT scores obtained from individual service files. (One nonwhite subject was dropped from this analysis because his GCT score was not available.) These correlations were determined for racial subgroups of subjects falling below and above the GCT mean score of 50, for both racial groups in toto, and for the entire sample. These figures are seen in Table 7. Significant values of the correlation coefficient are noted only in the white group as a whole and for the entire sample. Nonwhite test scores did not correlate significantly with GCT performance.

Table 7

Correlations of Test Performance (IPR) with Navy
General Classification Test (GCT) Score

Group	Group Averages		Correlation Coefficient			
			GCT	IPR	GCT GRP	Race GRP Total
Nonwhite N=33	Low (<50) N=24	42.67	.208	.316		
	High (≥50) N=9	56.89	.207	.601		
						.213
White N=104	Low (<50) N=17	42.18	.207	.253		
	High (≥50) N=87	59.63	.238	.050		
						.270**
						.223*

*Significant at $p < .05$.

**Significant at $p < .01$.

DISCUSSION

CONSTRUCT VALIDITY

The test was constructed to be a measure of learning ability with the implication that learning ability is a manifestation of the intellectual capacity of a person. Differences in this intellectual capacity between individuals was assumed to be measurable by the rate with which new material is learned. Using IPR as the rate measure, the results of the trial administration of the test showed that learning took place and that the rate was different among individuals. Moreover, the results were found to be highly reliable--especially for a 4-minute test--using an internal (split-half) criterion of reliability. Thus, the basic essential requirements for the construct validity of the test would seem to have been adequately demonstrated. Additional experimentation would be required to show that it is, indeed, a differential measure of intellectual capacity. Probably the best way to demonstrate this essential requirement would be to give the test to different age groups. The fact that the items had been standardized for information content (1 bit per item) would make it possible to administer shorter forms of the test--e.g., four instead of six items--to different age groups and yet have the IPR mean the same when corrected for total information content of the stimulus lists.

Earlier in this paper, it was stated that the construct validity of a test required an answer to the question, What does the test measure? The answer given here is learning ability. But, as Estes (1974) has argued, a product-defined measure of intelligence or ability does not provide an understanding of what intelligence is. Rather, the process should be defined and the relationship between the process and the product measure should be determined. The design of this trial administration of the test does not provide opportunities to answer the process question. Since similarity, however, was not a significant variable, visual discrimination of the stimuli would not seem to have been involved in the learning process. Based on a great deal of research in recent years in the area of human learning and information processing, it would be safe to say that some form of coding of the individual forms and, probably, the stimulus pairs as an entity was required. Additionally, short-term memory was required to hold the information pertaining to one item in working memory while processing a new item. Here, some sort of mnemonic device might be involved, and in both cases verbal fluency and image formation might be the basic skills underlying these processes. With respect to verbal ability playing a role, the small, significant correlation between IPR scores and the GCT scores for the white group would support this contention. Taken in conjunction with this finding, the absence of a significant correlation for the nonwhite group could also be seen as not disconfirming the trend, if it is assumed that the GCT score is not as good a measure of verbal ability for subjects in the nonwhite group. These results, however, only emphasize that the measure of verbal fluency or the capacity to generate useful images must be appropriate to the cultural background of the individual subject.

CULTURAL IMPLICATIONS

If the subjects--white and nonwhite--had comparable learning abilities, no racial group differences would be found on the IPR. The study found no significant differences among the self-paced subjects, but a significant difference was found for racial groups in the machine-paced mode. A problem in attempting to determine from the experiment data whether the white and nonwhite group differed in learning ability lies in the fact that the subjects were a selected group that was not representative of America's youth in general. As noted, the average education level was at the 12th grade. The information in Table 7 shows that 60 percent of the sample was above the median in GCT scores. There was a considerable difference in racial groups, however, with 84 percent of the white group being above the 50th percentile, whereas only 27 percent of the nonwhite subjects were in that category. There was a small but significant correlation of GCT scores with the IPR, but only for the white group and the entire sample. How can these data be related to the cultural implications of the test?

With respect to the differences noted in the paced and self-paced groups, it may be that the machine-paced format placed greater pressure on the subjects and generated greater test anxiety. Where short-term memory and the learning of discriminations involving very similar items constitute the task, the effects of anxiety could be disruptive as shown by Taylor and Spence (1952) and Ramond (1953) in serial, verbal learning tasks. For anxiety to have a differential effect in the racial groups, the anxiety induced by the test conditions would have to be greater for the nonwhite group. This could be true as a part of the larger picture of differences in test-taking motivation, attitudes, experience, and skill that have been attributed to different cultural backgrounds. If these contentions are valid, then the self-paced mode would be more culture-free in its assessment of the test subject. If the finding in this trial administration of the test for the self-paced condition should hold up in subsequent administrations, then this would be strong evidence for the culture-fair nature of this test.

The pattern of correlations between the IPR and the subjects' GCT scores takes the form that Jensen (1968b) found with children of high and low socioeconomic (SES) groups. Noting that children from low SES backgrounds with IQs in the range of 60 to 80 appear to be much brighter in social and nonscholastic behavior than their middle- or upper-middle SES counterparts, he gave groups of such children learning tasks in the laboratory and compared their learning performance with standard intelligence test scores for the children. There was a substantial correlation of IQ and learning scores for middle-class children, but the correlation was negligible for children from low SES backgrounds. Jensen attributed the difference to the fact that the learning tasks and the intelligence tests measured two different levels of intelligence with the lower level, measured by the learning tasks, being common to both groups and the other being better represented within the high SES group. In the present instance, it would seem more parsimonious to conjecture that the IPR was a measure of intellectual capability for both

groups, whereas the GCT, which has been found to be culturally biased (Stephan, 1973; Thomas, 1972c), was a fair measure only for the white group. In addition, the significant correlations accounted for only a very small portion of the variance in IPR scores. Accordingly, it would appear that the multiple discrimination test is indeed culture fair and provides an unbiased measure of learning ability, at least in the self-paced form. Larger and more numerically balanced samples from an unselected population would be necessary to confirm these conclusions.

TEST AND TESTING CONSIDERATIONS

Discussion in this section will deal with the psychometric and physical aspects of the multiple discrimination learning test. Specifically, the length of the test, additional matters pertaining to the pacing mode, and the physical packaging of the test will be considered.

Test Length

The decision to stop the test after 10 trials was arbitrary. Several subjects showed errorless performance within this limitation. In the machine-paced mode where there was a theoretical limit to the IPR of .333 bits/sec., examination of the third block of trials showed that the white subjects attained a maximum of 80 percent of this perfect learning rate, while nonwhites reached 69 percent of this quantity. While it is not possible to tell how many trials are required for perfect learning, since a trials-to-criterion design was not used, it would be advisable from a psychometric standpoint to stop short of perfect learning when the difference in learning rate among subjects is more variable. There would also be a tradeoff between a test length of maximum discriminability among subjects and one of highest reliability, which might not be the same. Thus, the optimum test length is not a simple question that yet remains to be determined.

Pacing Mode

It has been previously shown that pacing mode appeared to have a difference on test results with the self-paced mode being more culture-fair. From a psychometric standpoint, the difference between the two methods is that the self-paced mode places no limit on the IPR that a subject might attain. This would lead to greater variability among subjects and, presumably, a more reliable differentiation among test takers. Since many more variables are free to exert their effects with the self-paced mode, it may be, however, that less reliable performance may result. The self-paced mode, though, should be more representative of the manner in which a subject approaches and deals with a problem, and the results of the testing, as a consequence, would be more generalizable to real-life situations where learning is required. That is, it should permit greater predictive validity.

The self-pacing mode would also be desirable on the basis of the discussion on the construct validity of the test. There it was stated that the rate of learning would be the measure of learning ability, and the self-paced mode is the only one that permits an assessment of this measure. The highest rate in this study was .503 bits/sec., which occurred in the nonwhite subgroup of the self-paced condition. Accordingly, the self-paced mode would appear to be the better procedure for this test.

Physical Packaging of the Test

The type of stimulus materials, their presentation method, and scoring make it relatively simple to institutionalize the test using teaching machines with true-false or multiple-choice response provisions. Scoring counters could be readily integrated with the machine. With the ever-expanding use of computer terminals at remote locations, the test could easily be set up to be administered from a central location. This would permit the ready selection of a test "form" from among several that could be accessed, and scoring and performance analysis would be almost instantaneously provided upon completion of testing.

A specific item that requires improvement over the set-up used in this trial administration of the test is the advance procedure in the self-paced mode. In this trial, the subject had to call for the next stimulus after responding by pressing a button on the side of the response unit. As a result, learning times for the self-paced group might have been slightly biased upwards.

Another feature that requires investigation is whether the reinforcement should be given by a signal only for correct choices. That was the procedure in this trial administration. The learning literature has a large number of studies that have investigated positive reinforcement, negative reinforcement, both positive and negative reinforcement, and correction vs. noncorrection methods--e.g. Arima (1965). There is a good likelihood that the correction method might be best for this test. That is, the next stimulus item will not appear until the subject presses the correct button. If the subject has initially chosen the incorrect alternative, he or she must press the correct button. The best mode should be determined by experimentation.

SUMMARY AND CONCLUSIONS

The purpose of this study was to develop a test of learning ability that would not be affected by the cultural background of the individual being tested. A test was created using randomly shaped, 2-dimensional polygons presented in pairs in a discrimination learning paradigm. Three different lists of six such pairs were created so that multiple discrimination learning was involved. The lists were presented individually in a manner similar to verbal discrimination learning in both a self-paced and machine-paced mode.

In a trial administration of the test using Navy recruits as subjects, significant learning took place over 10 trials. Nonwhite and white racial groups, which differed significantly on their Navy General Classification Test Scores, performed at a comparable level in the self-paced mode. The adjusted reliability of the test (split-half) was .85. The correlation of the test scores with the GCT scores was marginally significant for the white group and the total sample, but not for the nonwhite group. There was no difference in performance among the three lists, which differed considerably in the similarity of the stimulus materials. This suggested that any combination of the forms could be used to create equivalent alternate forms.

It was concluded that a practical test of learning ability that was culture fair to both the white and nonwhite groups had been demonstrated. Refinement of the test would be desirable with respect to optimal length, reinforcement procedure (correction vs. noncorrection), and the physical packaging of the test.

REFERENCES

- Arima, J. K. Human probability learning with forced training trials and certain and uncertain outcome choice trials. Journal of Experimental Psychology, 1965, 70, 43-50.
- Arima, J. K. Verbal discrimination learning: An analysis of randomly presented 2-, 3-, and 4-word items. JSAS Catalogue of Selected Documents in Psychology, 1974, 4, 116.
- Arnoult, M. D. Familiarity and recognition of nonsense shapes. Journal of Experimental Psychology, 1956, 51(4), 269-276.
- Bitterman, M. E. The evolution of intelligence. Scientific American, January 1965, 92-100.
- Bitterman, M. E. The comparative analysis of learning. Science, 1975, 188, 699-710.
- Congressional Budget Office. The costs of defense manpower: issues for 1977. Washington: Superintendent of Documents, January 1977.
- Eckert, E. and Kanak, H. J. Verbal discrimination learning: a review of the acquisition, transfer, and retention literature through 1972. Psychological Bulletin, 1974, 81, 582-607.
- Estes, W. K. Learning theory and intelligence. American Psychologist, 1974, 29, 740-749.
- Feldman, M. W., and Lewontin, R. C. The heritability hang-up. Science, 1975, 190, 1163-1168.
- Goslin, D. A. Standardized ability tests and testing. Science, 1968, 159, 851-855.
- Green, D. R. What does it mean to say a test is biased? Education and Urban Society, 1975, 8, 33-52.
- Green, E. J., and O'Connell, J. A. An annotated bibliography of visual discrimination learning. Teachers College Press, Columbia University, New York, N. Y., 1969.
- Holtzman, W. H. The changing world of mental measurement and its social significance. American Psychologist, 1971, 26, 546-553.
- Hunter, J. E., Schmidt, F. L., and Rauschenberger, J. M. Fairness of psychological tests: implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 1977, 62, 245-260.
- Jensen, A. R. Social class, race, and genetics: Implications for education. American Educational Research Journal, 1968, 5, 1-42. (a)

- Jensen, A. R. Patterns of mental ability and socioeconomic status. Proceedings of the National Academy of Sciences, 1968, 60, 1330-1337. (Reprinted in American Psychologist, 1968, 26, 1330-1337.) (b)
- Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, Calif.: Wadsworth, 1968.
- Matarazzo, J. D. Wechsler's measurement and appraisal of adult intelligence. Baltimore: Williams and Wilkins, 1972.
- Melton, A. W. Military psychology in the United States of America. American Psychologist, 1957, 12, 740-746.
- Ramond, C. K. Anxiety and task as determinants of verbal performance. Journal of Experimental Psychology, 1953, 46, 120-124.
- Stephan, R. A. Evidence of racial bias in military testing. Paper presented at 31st meeting of Military Operations Research Society, Monterey, California, 1973.
- Taylor, J. A., and Spence, K. W. The relationship of anxiety level to performance in serial learning. Journal of Experimental Psychology, 1952, 44, 61-64.
- Thomas, P. J. The relationship between Navy classification test scores and final school grade in 104 Class "A" schools (Res. Rep. SRR 72-15). San Diego: U.S. Navy Personnel and Training Research Laboratory, 1972. (a)
- Thomas, P. J. The relationship between Navy classification test scores and final school grades in 98 Class "A" schools (Res. Rep. SRR 72-22). San Diego: U.S. Navy Personnel and Training Research Laboratory, 1972. (b)
- Thomas, P. J. An investigation of possible test bias in the Navy Basic Test battery (Tech. Bull. STB 73-1). San Diego: U.S. Navy Personnel and Training Research Laboratory, 1972. (c)
- Windle, C. and Vallance, T. R. The future of military psychology: paramilitary psychology. American Psychologist, 1964, 19, 119-129.
- Young, P. A. A culture-free performance test of general learning ability. M.S. Thesis, Naval Postgraduate School, Monterey, Calif., December 1975.