PRS REPORT NO. 958

VALIDATION OF THREE OBJECTIVELY SCORED PICTORIAL TESTS OF PERSONALITY FOR THE ASSESSMENT OF LEADERSHIP

By

Raymond A. Katzell, Ph.D.
Principal Investigator

James J. Kirkpatrick
Project Supervisor

Frederick O. Carleton
Research Assistant

Nathan Rosenberg
Research Assistant

PSYCHOLOGICAL SERVICES CENTER
SYRACUSE UNIVERSITY

A CONTRACT RESEARCH REPORT TO

PERSONNEL
RESEARCH
SECTION
PR & P BR - PERS BUR - AGO

PRS REPORT NO. 958

VALIDATION OF THREE OBJECTIVELY SCORED PICTORIAL TESTS OF PERSONALITY FOR THE ASSESSMENT OF LEADERSHIP

by

Raymond A. Katzell
Principal Technician
Psychological Services Center
Syracuse University
Syracuse, New York

James J. Kirkpatrick
Project Supervisor

Frederick O. Carleton
Research Assistant

Nathan Rosenberg
Research Assistant

FINAL REPORT 31 May 1952

DEPARTMENT OF THE ARMY
Project No. 29534100 Subtask 76
RESEARCH CONTRACT NO. DA-49-083 OSA-64

PERSONNEL RESEARCH SECTION PR 3476

PRS reports are primarily technical. While conclusions affecting military policy or operations may appear in them, they are not intended as a basis for official action. Findings and conclusions contained in PRS reports are intended to guide the conduct of further research. When research findings suggest recommendations for administrative action, such recommendations are made separately to the appropriate military agency.

BRIEF

THE EFFECTIVENESS OF PICTORIAL TESTS OF PERSONALITY IN THE ASSESSMENT OF LEADERSHIP
(Based on PRS Report 958)

STATEMENT OF THE PROBLEM

One of the most important problems of personnel management in the Army is identifying (1) men with qualities of leadership, and (2) men who can readily be trained as officers and noncommissioned leaders. [illegible] ful than others. The purpose of the present study was to evaluate three new pictorial tests of personality as predictors of leadership ability at the U. S. Military Academy and in Leader's Schools.

RESULTS

1. For a sample of privates enrolled at Leader's Schools, two out of three of the tests gave a better than chance differentiation between men rated high by their associates and men rated low.

2. The power of these tests to distinguish between high rated privates and low rated privates is about the same as a test already in use: "the Leaders Self-Description Blank." The new tests are not closely related to the old one.

3. However, for noncommissioned officers at Leader's Schools, no one of the tests differentiated between high rated men and low rated men.

4. None of the tests gave scores related to Aptitude-for-Service Ratings for cadets at the Military Academy.

CONCLUSIONS

1. The validity of pictorial tests used in this experiment was insufficient to add significantly to the validity attainable with Self-Description Blanks previously developed by the Personnel Research Section, Personnel Research and Procedures Branch, The Adjutant General's Office.

2. In order for pictorial tests to become effective leadership predictors, it appears necessary to effect improvement in item content and format. Whether such improvement would be sufficient to warrant the cost is debatable.

WORK SUMMARY

Three new tests, Picture Interpretation Test, Army Picture Story Test, and the Picture Fill-in Test, were administered to 216 cadets at the Military Academy and 958 enlisted men in Leader's Schools. In addition, the West Point Personal Inventory was administered to the cadets, and the Leaders Self-Description Blank to the enlisted men.

Responses to test items and total test scores were compared with an independent measure of leadership, the Aptitude-for-Service Rating for cadets or the associate rating for enlisted men and verified on additional groups of 258 cadets and 269 enlisted men.

TABLE OF CONTENTS

Page

LIST OF TABLES iii

PREFACE iv

SUMMARY v

I. INTRODUCTION 1

II. OBJECTIVES 1

III. METHOD 3

A. Tests 4
1. Picture Interpretation Test 4
2. Army Picture Story Test 5
3. Picture Fill-In Test 5
4. West Point Personal Inventory 6
5. Leaders' Self-Description Blank 6

B. Situational Validity at the Commissioned Officer Level 7
1. Sample 7
2. Tests 7
3. Criterion 7
4. Procedure 8
5. Analysis of Data 9
a. Item Analysis 9
b. Pattern Item Analysis 10
c. Cross-Validation 10

C. Enlisted Personnel Validity Studies 11
1. Samples 11
a. Item Analysis Group 11
b. Cross-Validation Group 11
2. Tests 11
a. Item Analysis Group 11
b. Cross-Validation Group 12
3. Criteria 12
a. Item Analysis Group 12
b. Cross-Validation Group 14
4. Procedure 14
a. Item Analysis Group 14
b. Cross-Validation Group 14
5. Analysis 15
a. Item Analysis Group 15
b. Cross-Validation Group 18

Page

IV. RESULTS 20

A. Results with Cadet Officers 20
1. Picture Interpretation Test 20
a. Item Analysis 20
b. Cross-Validation 20
2. Army Picture Story Test 22
a. Item Analysis 22
b. Pattern Item Analysis 22
c. Cross-Validation 22
3. Picture Fill-In Test 24
a. Item Analysis 24
b. Cross-Validation 24
4. West Point Personal Inventory 24
a. Item Analysis 24
b. Cross-Validation 24

B. Results with Enlisted Personnel 26
1. Item Analysis Group: All Tests 26
a. Item Analyses 26
b. Comparison of Item-Analysis Group and Dropouts 29
2. Cross-Validation 29
a. Reliabilities and Related Statistics 29
b. Validities 29
(1) Item Analysis Keys 29
(2) Item Classification Key for the Picture Interpretation Test 32
(3) Multiple Correlation 32

V. CONCLUSIONS 35

A. Validity of the three objective projective tests for measuring leadership performance of West Point cadets 35

B. Validity of three objective projective tests for measuring leadership performance of Army non-commissioned personnel 35

c. Inferences concerning the general promise of the objective projective type of test 37

VI. IMPLICATIONS AND RECOMMENDATIONS 38

LIST OF TABLES

Table		Page
1.	Intercorrelations among Various Criteria for Two Enlisted Samples	13
2.	Number of Cases in Item-Analysis and Cross-Validation Samples	15
3.	Critical Ratios and Variance Ratios for Testing Differences in Mean Associate Rating Scores	15
4.	Distribution of Item Validities in the Picture Interpretation Test for Two West Point Samples	21
5.	Distribution of Item Validities in the Army Picture Story Test for Two West Point Samples	23
6.	Distribution of Item Validities in the Picture Fill-In Test for Two West Point Samples	25
7.	Number of Significant Items for Three Experimental Personality Tests	26
8.	Distribution of Item Validities in the Picture Interpretation Test for Two Enlisted Samples	27
9.	Distribution of Item Validities in the Army Picture Story Test for Two Enlisted Samples	28
10.	Distribution of Item Validities in the Picture Fill-In Test for Two Enlisted Samples	28
11.	Significance of Mean Differences between Dropouts and Item Analysis Group for Picture Fill-In and Picture Interpretation Tests . .	30
12.	Reliabilities and Related Statistics Estimated from Cross-Validation Samples at Leaders' Schools for the Picture Interpretation and Picture Fill-In Tests	30
13.	Validity Coefficients for the Picture Fill-In and Picture Interpretation Tests in the Privates' Sample.	31
14.	Validity Coefficients of the Item Analysis and Item Classification Keys for the Picture Interpretation Test and the Correlation between Keys.	33
15.	Intercorrelations Among Predictor Variables, Correlations with Associate Rating Criterion, and Multiple Correlations for Cross-Validation Sample.	34

PREFACE

This is a report of a research study initiated in May, 1950, and performed under Contract Number DA-49-083 OSA-64, negotiated under authority of Section 2 (c) (5), Act of February 19, 1948 (Public Law 413 - 80th Congress).

The study was conceived and executed in association with the staff of the Personnel Research Section, The Adjutant General's Office, Department of the Army. Particularly noteworthy were the contributions made by Drs. D. E. Baier, H. E. Brogden, R. Perloff, E. K. Taylor, and the late Dr. C. I. Mosier.

On the staff of the contractor, indispensable collaboration was furnished by a number of individuals in addition to those whose names appear, somewhat arbitrarily, in authorship. Among them are Dr. Ernst G. Beier, Mr. D. K. Hable, Miss Mildred E. Leonard, Mr. Rolland Tougas, and Mrs. Elizabeth R. Coleman and her scoring staff.

The cooperation of the authorities at Forts Belvoir, Dix, Jackson, and Knox, and at the U. S. Military Academy, notably Lt. Col. Raymond Rumpf, Maj. Herman F. Smith, and Dr. Douglas Spencer, was invaluable in the acquisition of data.

To these individuals, and to many others of whom space or memory preclude the mention, the authors express their gratitude.

SUMMARY

A. Problem

1. To identify those items, in each of three new objectively scored projective tests, which discriminate between superior and inferior leaders among West Point cadets and enlisted trainees in Leaders Schools.

2. To determine the validity and reliability of the resulting scoring keys for the assessment of leadership in new samples of personnel.

3. To compare the validity of these keys with that of a biographical inventory currently used by the Army for leadership assessment.

4. To factor analyze the several tests found valid with West Point cadets along with other leadership measures, in order to investigate basic personality factors intrinsic to such measures.

B. Method

1. The Tests --

a. Picture Interpretation Test - involves elective identification with individuals depicted in various roles and activities.

b. Army Picture Story Test - involves the ranking of statements with regard to their appropriateness in describing each of a series of pictures.

c. Picture Fill-In Test - entails the rating of appropriateness of rejoinders in conversational situations depicted in cartoons.

d. West Point Personal Inventory - a series of biographical and self-descriptive questions, used with West Point cadets.

e. Leaders Self-Description Blank - a series of biographical and self-descriptive questions, used with Leaders School trainees.

2. The Criteria --

a. The West Point Aptitude Rating was used as the measure of leadership performance of West Point cadets. This is a composite rating on leadership made by the cadet's peers and tactical officer.

b. The Associate Rating, mainly a nomination rating by peers, was employed as the standard of leadership performance of Leaders School trainees.

3. The three projective tests were administered to 404 West Point cadets and 958 Leaders School trainees. The West Point Personal Inventory was also administered to these cadets. Criterion data were obtained for as many of these individuals as feasible.

4. All items on the three projective tests were biserially correlated with the criterion. This was done for four groups of subjects, as follows: two randomly selected groups of cadets, numbering 213 and 224, respectively; 385 privates enrolled in Leaders Schools; [illegible] non-commissioned officers enrolled at Leaders Schools.

5. Scoring keys were developed from this analysis, those items being keyed which had criterion correlations minimally significant at the 10% level of confidence.

6. The Picture Fill-In and Picture Interpretation Tests, and the Leaders Self-Description Blank, were administered to a new sample of 296 privates enrolled at Leaders Schools. Criterion data were secured for these individuals. Validity and reliability statistics were computed for this group.

C. Results —

1. In the two samples of West Point cadets, there was no better than a chance relationship between responses to the items on all three projective tests and the Aptitude Ratings received by the cadets. The West Point Personal Inventory had a correlation of .35 with the criterion in the two samples combined.

2. Similar negative results were obtained in the item analysis of the tests against Associate Ratings of non-commissioned officers enrolled in Leaders Schools.

3. In a sample of privates enrolled at Leaders Schools, it was possible to identify in two of the three projective tests an appreciably larger-than-chance number of items that distinguished between the higher- and lower-rated men. These two tests were the Picture Interpretation and Picture Fill-In Tests.

4. These tests, when scored for the new sample of privates by the scoring key developed on the first group, yielded validity coefficients of .25 for the Picture Interpretation Test, and .19 for the Picture Fill-In Test. The Leaders Self-Description Blank had a validity coefficient of .30 in this sample. Each of these coefficients differs significantly from zero at the 1% level of confidence.

5. The split-half reliability coefficients, augmented by the Spearman-Brown Prophecy formula, were .85 for the Picture Interpretation Test and .91 for the Picture Fill-In Test in the cross-validation sample.

6. The correlations of the two tests with each other and with the Leaders Self-Description Blank were all [illegible] and positive.

D. Conclusions

1. The three projective tests, as now constituted, are of no value for leadership assessment of West Point cadets.

2. These tests are also of no value for leadership assessment of non-commissioned officers in Leaders Schools.

3. Scoring keys were developed for both the Picture Fill-In and Picture Interpretation Tests on a sample of privates in Leaders Schools. Scoring the two tests for a new sample of privates by means of these keys yielded scores which were significantly correlated with the criterion of leadership in Leaders Schools.

4. The biographical inventories (West Point Personal Inventory and Leaders Self-Description Blank) showed significant criterion correlations in their respective samples.

5. Among the privates, the two valid projective tests did not add appreciably to the predictive power of the biographical inventory when combined with it in a multiple regression equation. Nonetheless, their correlations with the inventory are low (about .35), as is their correlation with one another (.18).

6. It is inferred that the Picture Fill-In Test and the Picture Interpretation Test show considerable promise as techniques for leadership assessment, although improvements are needed to translate this promise into a state of practical utility. Suggestions are made manifest in this study as to how improvements may be effected in regard to: (1) power to discriminate more accurately between superior and inferior leaders, and (2) extending the range of personnel with whom such tests would be useful.

7. In view of the lack of validity of the projective tests among West Point cadets, it was not meaningful to proceed with the factor analysis designed to reveal the basic personality factors common to these and other measures of leadership, so that this objective of the study could not be achieved.

I. INTRODUCTION

The identification of men with high potentialities as leaders is understandably a matter of prime importance to the Army. Accordingly, a considerable amount of research has been done or sponsored by the Army on techniques for accomplishing such identification.

Although it is commonly believed that non-intellective factors are of major importance in determining a man's leadership performance, methods for measuring such factors still leave much to be desired in the way of validity and accuracy. In recent years, the evidence has grown more suggestive that projective tests[1] may have promise along these lines. However, these tests are typically time consuming to administer and score, and typically require trained psychologists for their interpretation. These characteristics are manifestly unsuited for large-scale military classification purposes.

To circumvent these deficiencies, The Personnel Research Section of the Adjutant General's Office undertook the preparation of several tests which are fundamentally projective in nature but which are amenable to group administration and objective (even machine) scoring. When any new test is constructed, the questions of its validity and what it measures immediately arise. These questions become even more urgent when the test represents a radically new departure. Thus, in the case of the new objective projective tests, not only are their particular validities unknown, but also subject to question are the issues of the general fruitfulness of the approach and of the underlying psychological dimensions measured by such techniques.

The research described in this report was undertaken in an effort to shed light on these questions.

II. OBJECTIVES

More specifically, the objectives of this research may be described as follows:

A. To ascertain the validity of each of three objective projective tests for measuring leadership performance of Army commissioned personnel.

1. To determine the correlation of each item with a criterion of leadership performance, on the basis of which to develop a scoring key for each test.

[1] A projective test requires the examinee to interpret or structure a stimulus situation which lends itself to a variety of meanings, and thereby to reveal aspects of his personality.

2. To ascertain for each of these scoring keys its reliability and validity against a criterion of leadership performance at the level of commissioned personnel.

3. To compare the relative validities of these tests with one another, and with a self-description questionnaire.

B. To ascertain the validity of each of these objective projective tests for measuring leadership performance of Army non-commissioned personnel.

1. To determine the correlation of each item with a criterion of leadership performance, on the basis of which to develop a scoring key for each test.

2. To ascertain for each of these scoring keys its reliability and validity against a criterion of leadership performance at the level of non-commissioned personnel.

3. To compare the relative validities of these tests with one another, and with a self-description questionnaire.

C. From these data, to infer the general promise of this type of test, and to deduce indications of which lines of future development seem most fruitful.

It was also hoped originally to factor analyze the relationships among these tests, together with other personality measures including ratings and behavior measures, with the objective of determining basic personality factors gauged by such variables. Since the non-test variables were more appropriate and available in the commissioned personnel situation (West Point), the plan was to perform this analysis in connection with the data obtained from that sample. However, it was discovered in the course of the research that the projective tests were virtually uncorrelated with the leadership criterion in this situation, thus making the planned analysis pointless.

III. METHOD

The general plan of the study consisted of administering the tests to samples of personnel who were representative of the two levels of leadership activities for which such tests might be valuable assessment techniques. An additional requirement for selecting the samples was that the personnel be assigned to situations in which criteria of leadership performance would be available.

In accordance with these standards, cadets in the upper classes of the United States Military Academy at West Point were chosen as the sample whose characteristics and activities were approximately representative of personnel to be assessed for potential leadership at the commissioned level.

Students at Leadership Schools were selected as suitable for representing potential leaders at the non-commissioned level. This group is actually composed of two subgroups, as regards age, background, and previous experience: privates and non-commissioned officers. It was deemed advisable to investigate separately the validity of the tests for each of the two subgroups.

Thus, there were three categories of personnel who were the subjects of the investigation: West Point cadets, privates assigned to Leadership Schools, and non-commissioned officers assigned to Leadership Schools.

The research design involved the following steps for each of the categories of personnel:

1. Administering the three objective projective tests to samples of the personnel.

2. Collection of criterion data for these individuals.

3. Correlation of the test items against the criterion.

4. Development of a scoring key for each test.

5. Application of the key to the test results of new samples of personnel.

6. Correlation of the scores on each test with the criterion of leadership performance.

In the remainder of this chapter, the tests will first be described, followed by a description, for the cadet officers, of the samples, criterion, procedure for collecting data, and methods of analyzing the data. Finally, the same rubrics of information will be presented for the enlisted personnel.

A. Tests (Copies of the tests are included in the Appendix of this report.

The following tests were included in the validation study:

1. Picture Interpretation Test, 1949. (DA AGO PRT - 1775)

This 432 item test consists of a series of 268 pictures, some of which present individuals participating in military activities and others involving individuals in civilian situations. The general directions indicate that the test is a measure of interests, although it may be considered a projective instrument to the extent that the examinee tends to identify with the situations and individuals illustrated in the pictures.

Instructions for the first six parts of the test follow the same general pattern. For the individuals or situations presented in the pictures in each part of the test, the examinee is required to choose between two alternative reactions, as follows:

(1) Part I

(a) "Yes, I would like to do what he is doing," or
(b) "No, I would not like to do what he is doing."

(2) Part II

(a) "Yes, I would like to be that person," or
(b) "No, I would not like to be that person."

(3) Part III

(a) "Yes, this person is like me," or
(b) "No, this person is not like me."

(4) Part IV

(a) "Yes, I would admire this person," or
(b) "No, I would not admire this person."

(5) Part V

(a) "Yes, I am good at doing what this person is doing," or
(b) "No, I am not good at doing what this person is doing."

(6) Part VI

(a) "Yes, I like what is shown in this picture," or
(b) "No, I do not like what is shown in this picture."

Part VII differs from the rest of the test in that pictures of military situations and civilian situations are presented along with one descriptive statement for each picture. The examinee is required to make the following choice in regard to each statement:

"Yes, the picture made me think of this idea," or
"No, the picture did not make me think of this idea."

2. Army Picture Story Test, Series B, 1950, Syracuse University Press.

The Army Picture Story Test is an objective test, based on the general idea of the Thematic Apperception Test, consisting of a series of ten pictures. The pictures included in the Army Picture Story Test involve both military and non-military situations and are not the pictures used in the Thematic Apperception Test. For each picture, there are thirty items presented in groups of three. The items are relatively short statements which are descriptive of the picture. The examinee is instructed to read the statements within each triad and to select two statements: the most descriptive and the least descriptive.

The statements used in this test were obtained by administering the set of ten pictures to a large group of soldiers in a free response situation. The descriptions written by this group were edited and arranged in triads on the basis of their frequency of occurrence and with respect to a number of clinical categories. That is, triads were composed of items which were approximately equal in frequency of occurrence but which dealt with different personality needs.

3. Picture Fill-In Test, Second Form, 1949 (DA AGO PRT-1726)

The Picture Fill-In Test is an adaptation of the Rosenzweig Picture-Frustration Test. It differs from the Rosenzweig test in that the responses are obtained in objective form. A series of 43 cartoon-like pictures is presented, comprising a total of 392 items. In each picture, one individual is represented as saying something to another individual. Some of the pictures deal with military situations, while 24 pictures were taken directly from the Rosenszweig test. In an experimental administration of the Preliminary Form of the Picture Fill-In Test, the examinees wrote responses in the cartoon balloons. Responses made most frequently by this experimental group were selected for each of the pictures. Certain responses which seemed to be particularly revealing or measuring important factors also were included, regardless of their frequency of occurrence. From seven to ten responses were selected, and are presented below each picture in the Second Form of the test. This form, which was used in the present investigation, was developed so that it would be suitable for objective scoring in the following manner: The instructions require that the examinee rate each of the responses presented with the pictures with respect to how likely it is that the person shown would give that response. This rating of each response is accomplished on the following three point scale:

A. "Might say something like this."
B. "Is likely to say something like this."
C. "Is very likely to say something like this."

4. West Point Personal Inventory, 1949 (DA AGO PRT-1756) (Also referred to as ROTC Self-Description Blank, Form II, 1949, DA AGO PRT-1744, and in previous progress reports as Biographical Information Blank, ROTC edition.)

The West Point Personal Inventory used in the present investigation with cadets consists of four sections and a total of 420 items. This test does not make use of pictorial material. The items are in the form of statements concerning various characteristics, as follows: Section I includes pairs of statements dealing with personal characteristics; the individual is instructed to select the statement in each pair that is the best description of him. In Section II, the individual makes a choice between each of two activities as to which he believes he can do better. Statements dealing with likes and dislikes are presented in Section III, and the individual again selects the statement in each pair that he likes the better. Section IV contains statements describing personal characteristics, likes and dislikes, abilities, and beliefs. For each statement, the individual indicates whether the statement applies to him or does not apply.

5. Leaders' Self-Description Blank, Form B, 1951, Syracuse University Press.

The Leaders' Self-Description Blank is a 342-item version of the Biographical Information Blank and was used at the non-commissioned level at the Leaders' Schools in the present investigation. It is similar in composition to the West Point Personal Inventory, but the exact content of the items is different. Like the West Point Personal Inventory, it does not present pictorial material and consists of four sections.

Section I contains pairs of statements dealing with personal characteristics. The examinee chooses the statement from each pair that describes him better. The pairs of statements in Section II describe various activities, and the individual selects the activity which he can do better. Pairs of statements are presented in Section III dealing with likes and dislikes, and the instructions require selecting the statement that you like better. Personal characteristics, likes and dislikes, abilities, and beliefs make up the content of Section IV, and the examinee is instructed to indicate whether each statement applies to him or does not apply.

B. Situational Validity at the Commissioned Officer Level

1. Sample

The first and second classes of cadets at the U. S. Military Academy, West Point, in July, 1950, were the subjects of the investigation. A total of 454 cadets was tested.

Available cases from this sample were later divided into two random-half groups for purposes of performing a double cross-validation analysis. These subgroups comprised, respectively, 213 and 223 cases.

2. Tests[1]

The tests employed with this sample were:

a. Picture Interpretation Test, 1949, (DA AGO PRT-1775)
b. Army Picture Story Test, Series B, 1950, Syracuse University Press
c. Picture Fill-In Test, Second Form, 1949, (DA AGO PRT-1726)
d. West Point Personal Inventory, 1949 (DA AGO PRT-1756)

3. Criterion

The Aptitude for the Service System[2] was ascertained for each cadet for use as a criterion measure of leadership. The Aptitude for the Service System is used at West Point for the purpose of providing an accurate evaluation of the leadership effectiveness of cadets. The Aptitude Rating is a composite measure including the pooled opinion of the cadet's Tactical Officer and a small group of classmates within his Company. The evaluation by his classmates is accomplished through an associate (buddy) rating procedure.

Each cadet is ranked in order of merit by his Tactical Officer and by the cadets in his Company in regard to the following definition of leadership:

"The criterion of my appraisal is each cadet's ability (if or when placed in command of a group) to elicit the group's maximum cooperation; maintain the highest possible standards of administration and

[1] A description of the tests used in the study is presented in Section III, A.

[2] A detailed description of the Aptitude for the Service System may be found in "The Operation and Administration of the Aptitude for the Service System, U.S.M.A.", West Point, New York: United States Military Academy, 1951.

discipline; and at the same time, develop and preserve high morale and group spirit."[1]

From the raw ratings, the _median_ ranking for each cadet is determined and transposed to a standard score (called Army Standard Rating). The Tactical Officer's rating is assigned a weight of one-third in combining it with the associate ratings. It is this final or composite Army Standard Rating (Aptitude Rating) that constituted the criterion of leadership in this investigation.

4. Procedure

a. Test Administration

On June 30 and July 1, 1950, the four tests were administered in group situations to 216 cadets of the first and second classes at the U. S. Military Academy, West Point. The test battery was divided into two sessions, two of the tests being administered in the first session and the other two tests being given in the second session. Each session required about three hours of testing time. A similar procedure was utilized when the second group of West Point cadets was tested on July 27, 28 and 29, 1950. This second group of cadets numbered 238 and were from the first and second classes.

b. Collection of criterion data and constitution of criterion groups.

Criterion data, entered on Hollerith cards, were received from the West Point statistical office. These cards contained the cadet serial number, the _mean_ Aptitude Rating based on the first term and the second term of the second class, Aptitude Ratings for both terms, and year of expected graduation. The criterion of leadership effectiveness utilized in this investigation with the West Point sample was the mean Aptitude Rating which summarizes the cadet's leadership performance during his second class.

The total sample was divided randomly on the basis of serial numbers, group A being composed of those cadets with even serial numbers, and group B having odd serial numbers. As a check on the randomness of this procedure, _t_ and F statistics were computed between the mean Aptitude Index criterion scores of the two groups; this analysis indicated that the two groups may be considered as random samples from the same population in regard to the leadership criterion. The purpose of fractionizing the sample in this manner was to make it possible to perform a double cross-validation on the scoring keys derived in the item analyses.

[1] _ibid._ p. 2.

D. Analysis of Data

a. Item Analysis

The validity of the items in the experimental tests was estimated by computing the biserial correlation coefficient between dichotomized item responses and the Aptitude Rating criterion. The computation of the item validities was facilitated by making use of the Kolbe and Edgerton table for estimating biserial correlation coefficients.[1]

The Aptitude Rating criterion was normalized by dividing the distribution into equal frequency eighths, and assigning the standard score equivalent of the mid-point of each eighth in a normal distribution to each criterion score within that eighth. Thus, all cases falling in a given eighth of the obtained distribution of criterion scores received the same standard score equivalent.

While the same general item analysis procedure was followed for the West Point study, somewhat different techniques of dichotomizing the item responses were necessary for the different tests, as follows:

(1) Picture Interpretation Test, 1949 (DA AGO PRT-1775) The item responses in this test fit a natural dichotomy since the examinee is instructed to indicate either "Yes" or "No" for each item. Thus, there is no problem in dichotomizing the responses for the purposes of the biserial correlation type of item analysis.

(2) Army Picture Story Test, Series B, 1950, Syracuse University Press. As described in Section III, A, Tests, the Army Picture Story Test requires that the individual choose the most descriptive and the least descriptive statements from groups of three items. Within the triad, the item that is considered to be most descriptive is marked A, while the item that seems to be least descriptive is marked B, and the intermediate item is not marked. For purposes of obtaining item frequencies, I.B.M. graphic item counts were made for each item, for the A or B alternatives. The trichotomous alternatives for each item were dichotomized in order to apply the biserial correlation item analysis technique; in doing this, the extreme alternative ("Best" or "Worst") having the larger frequency of response was used as one category of the dichotomy, while the combination of the other extreme with the intermediate alternative constituted the other category. This arrangement was used in order to yield the closest approximation to a 50%-50% dichotomy, thus maximizing the stability of the resulting item validity coefficients.

[1] Kolbe, L. E., and Edgerton, H. A., "A Table for Computing Biserial r", J. Exp. Educ., 1936, 4, 245-251.

(3) Picture Fill-In Test, Second Form, 1949 (DA AGO PRT-1725). This test requires that the individual rate each item on a three-point scale in regard to the degree of likelihood that the item is an appropriate statement, as explained in Section III, B, Tests. The dichotomy required by biserial item analysis was achieved by combining the B and C responses. Thus, the frequencies of responses were obtained for the A category vs. the combined B and C category. The basis for grouping the B and C responses rather than utilizing some other combination in order to dichotomize the responses was both logical and empirical. On a logical basis, it seemed more reasonable to believe that B (Is likely to say something like this) is closer on a continuum to C (Is very likely to say something like this). Moreover, an inspection of the item responses indicated that by using the dichotomy of A vs. B and C, the ideal 50%-50% dichotomy was more closely approximated.

(4) West Point Personal Inventory, 1949 (DA AGO PRT-1756). An item analysis of this test was not necessary since the scoring key had already been developed in a previous study and was made available by the Personnel Research Section, AGO, for the validation phase of the present investigation.

b. Pattern Item Analysis

Since the items of the Army Picture Story Test are grouped in triads, it was hypothesized that the pattern of responses might be significant. In order to investigate this hypothesis, the following pattern analysis was performed: For each triad, six patterns of responses are possible. For Group A of the West Point sample, frequency counts were made of the responses to each of the six patterns for each of the 100 triads. A level of significance test based on X^2 was made among the frequencies of the patterns for each triad, contrasting upper and lower criterion groups. This procedure made it possible to estimate the validities of the pattern responses.

c. Cross-Validation

In general, the validities of the tests were estimated by computing Pearson product-moment correlation coefficients between the scoring keys derived and the Aptitude Rating leadership criterion.

In following this procedure, the two samples, group A (even serial numbers) and group B (odd serial numbers) were treated separately, in order that the scoring key derived on group A could be crossed over and validated on Group B, and the scoring key obtained on group B could be validated on group A. This double cross-validation technique makes use of the principle of replication in determining which items are consistently valid in both samples and permits two minimum estimates of the validity of the test.[1]

[1] For a fuller discussion of the double cross-validation technique, see Katzell, R. A., "Cross-Validation of Item Analyses", Educ. Psychol. Measmt., 1951, 11, 16-22.

C. Enlisted Personnel Validity Studies

1. Samples

a. Item Analysis Group

The enlisted personnel samples in this study were drawn from Army Leader's Schools whose mission is to train personnel for leadership, primarily at the non-commissioned officer level. It is anticipated that the bulk of students will consist of privates who have just completed basic training and who have been recommended by their company officers as evidencing leadership potential. This source of students is called "pipeline". Other sources have included reenlistments and National Guard personnel called up for active duty as a result of the Korean conflict. For these groups, training at the Leader's Schools is considered a refresher course. One other important category includes officer candidates who, at present are required to complete a leadership course before attending Officer Candidate Schools. Leaders Schools at Ft. Dix, N. J., and Ft. Knox, Ky., which train soldiers from ground force units, were visited to gather the data for item analysis. 958 men were tested at these two installations.

The sample was divided into two subgroups: privates (including privates first class) and non-commissioned officers. These groups differ in average age and military background, factors which might affect performance on the tests and criteria. Hence, it was considered desirable to perform separate item analyses and validations for the two subsamples.

b. Cross-Validation Group

Leaders Schools at Ft. Jackson, S. C., and Ft. Belvoir, Va., were visited to secure the data for cross-validation. These schools train personnel from infantry and engineering units, respectively. 368 cases were utilized for the cross-validation results.

2. Tests

a. Item Analysis Group

At Ft. Dix and Ft. Knox, the Picture Interpretation Test, 1949 (DA AGO PRT-1775), Picture Fill-In Test, Second Form, 1949, (DA AGO PRT-1776), and Army Picture Story Test, Series B, 1950, Syracuse University Press were administered.

b. Cross-Validation Group

On the basis of the item analysis performed, it was decided to administer only the Picture Interpretation Test and the Picture Fill-In Test to the cross-validation sample. In addition, at the request of PRS, the Leader's Self-Description Blank, Series E, 1951, Syracuse University Press, was administered to this same group.

Descriptions of the nature of all the above tests may be found in Section III, Part A, of this report.

3. Criteria

a. Item Analysis Group

The training cycle at Leader's Schools is divided into two four-week phases, Phase I and Phase II. Training during Phase I is primarily academic in nature, whereas Phase II consists primarily of practical leadership experience in field situations. During the training program, soldiers are periodically evaluated for their performance on different criteria by various kinds of raters, i.e., both commissioned and non-commissioned cadre as well as by their peers. All soldiers tested in this study were in Phase I of the training cycle at the time of testing.

The following criterion measures of leadership were obtained for the group tested at Ft. Dix and Ft. Knox: Faculty Board Rating, Associate Rating, Leaders' Reaction Test, Rating of Phase II Performance, and Total Rating (a weighted combination of the foregoing).

Intercorrelations among the above criteria were computed for a sample of the soldiers tested at Ft. Dix and Ft. Knox. These statistics are useful for estimating the extent to which the criteria measure different aspects of leadership. The following table shows these results. (See Table 1).

In both samples, it seems evident that the various criteria are somewhat unrelated to each other. Although the Associate Rating also appears to be somewhat different from other ratings of leadership potential, on the basis of its recommendation by Personnel Research Section for use in this study, it would seem to be the most appropriate measure of leadership available. Results from other Personnel Research Section studies[1] had shown associate ratings to be superior measures of leadership. In the present study, furthermore, Associate Ratings were available for more subjects tested than any of the other ratings.

[1] Wherry, Robert H. and Fryer, Douglas H., "Buddy Ratings: Popularity Contest or Leadership Criteria?", Personnel Psychology, 1949, 2, 147-159.

Table 1. Intercorrelations among Various Criteria for Two Enlisted Samples.

A. Fort Dix

N = 173

	F.B.R.	L.R.T.	A.R.	Perf. II	Total
Faculty Board Rating	---	.14	.45	.15	.74
Leaders Reaction Test		---	.34	[illegible]	.57
Associate Rating			---	.06	.44
Performance during Phase II				---	.66
Total					---

Correlation between Phase II and sum of other three variables = .20

B. Fort Knox

N = 112-162

	F.B.R.	L.R.T.	A.R.	Perf. II	Total
Faculty Board Rating	---	.32	.15	.33	.73
Leaders Reaction Test		---	.00	.38	.50
Associate Rating			---	-.12	.20
Performance during Phase II				---	.73
Total					---

Correlation between Phase II and sum of other three variables = .24

All of the above considerations led to the decision to use Associate Ratings as the criterion for the enlisted sample of this study.

Since this criterion was adopted, it will be valuable to describe in somewhat greater detail the operations by which scores on this measure were obtained for the samples. Each student is evaluated by his fellow students at the end of Phase I training. The students are each given a "Student Leadership Evaluation Report-Rating Sheet". On this sheet is a roster of the men in the student's group, customarily numbering from nine to fifteen men. The student, from this roster, chooses those whom he thinks the three best leaders and the three poorest leaders. On the next day, each student is given "Student Leadership Evaluation Report-Description Sheet". On this sheet are printed the names of the men in the group. There are also ten pairs of descriptive statements. For each man on the roster, the student is to choose the description in each pair of statements which most appropriately describes the man being rated. These sheets are then scored by using the keys furnished by Personnel Research Section. One score is based on the nominating technique, weights being given to the number of nominations received. The more nominations a soldier receives which are indicative of better leadership, the higher his Associate Rating score. The other score is derived from weights based on empirical evidence as to which descriptions are more characteristic of better leaders. The scores from the rating sheets and description sheets are averaged. These scores constituted the Associate Rating criterion used for item analysis purposes.

b. Cross-Validation Group

Associate Ratings were also obtained for the soldiers tested at Ft. Jackson and Ft. Belvoir, who had been tested in the period from about January through March, 1952. Test scores were obtained by use of the keys developed from the item analysis group. These scores were correlated with the Associate Ratings.

4. Procedure

a. Item Analysis Group

Trainees in Phase I at Ft. Dix were tested in August and October of 1950 and January 1951. The Picture Interpretation, Picture Fill-In, and Army Picture Story Tests were administered to groups ranging in size from approximately 50 to 100 trainees. Every effort was made to elicit the cooperation of the soldiers tested, including some explanation of the purpose of the study. A total of 480 subjects in Phase I was tested at Ft. Dix. In field trips made during October 1950 and January 1951, 478 subjects in Phase I were tested at Ft. Knox. Thus, the total number of trainees tested for the item analysis group was 958.

b. Cross-Validation Group

Trainees in Phase I at Ft. Jackson, S. C., and Ft. Belvoir, Va., were tested in January, 1952, under conditions similar to those obtaining for the item analysis group. On the basis of the results of the item analysis, it was decided to administer only the Picture Interpretation and Picture Fill-In tests to these groups. An additional test was administered at the request of Personnel Research Section, the Leader's Self-Description Blank. At Ft. Jackson, the number of privates whose test papers were adequately filled out was 166; non-commissioned officers numbered 50. At Ft. Belvoir, test papers from 143 privates were acceptable; the non-commissioned officer sample numbered 9. The total number for all three tests included 309 privates and 59 non-commissioned officers.

Table 2 summarizes the number of cases used.

Table 2. Number of Cases in Item-Analysis and Cross-Validation Samples.

A. Number of Cases in the Item Analysis Group

	Ft. Dix	Ft. Knox	Total
Privates	246	139	385
Non-coms	90	136	226

B. Number of Cases in the Cross-Validation Group

	Ft. Jackson	Ft. Belvoir	Total
Privates	166	143	309
Non-coms	50	9	59

5. **Analysis**

a. **Item Analysis Group**

For the purpose of item analyzing the three tests used, it was desired to combine the installation samples, since the resulting keys would be used irrespective of installation. The following analyses were performed in order to determine the most appropriate statistical method for combining the Associate Rating scores from the two installations.

Critical ratios were computed comparing mean Associate Rating scores obtained by soldiers at Ft. Dix with those at Ft. Knox. The differences between installations were significant at the 1% level of confidence.

Variance ratios were computed for these data to test the significance of differences in variability for Associate Rating scores. Differences in variance were significant at the 2% level of confidence for non-coms at Knox versus non-coms at Dix. This significant difference in variability makes ambiguous the interpretation of tests of significance for mean differences reported, since significant critical ratios between means may arise because of differences in variability.

The following tables present data from which the above interpretations were made.

Table 3. Critical Ratios and Variance Ratios for Testing Differences in Mean Associate Rating Scores

A. Privates

	Fort Dix		Fort Knox
N	[illegible]		[illegible]
Mean	79.16		72.44
Standard Deviation	4.79		4.02
Critical Ratio		14.6***	
Variance Ratio		1.42*	

B. Non-commissioned Officers

	Fort Dix		Fort Knox
N	90		137
Mean	80.44		72.30
Standard Deviation	3.94		5.12
Critical Ratio		13.6***	
Variance Ratio		1.68**	

*** Significant at the 1% level of confidence
** Significant at the 2% level of confidence
* Significant at the 10% level of confidence

On the basis of the preceding analyses, the best procedure for combining the two installations seemed to be conversion of Associate Rating raw scores to standard scores within each installation before pooling the two. Although about 950 men had been tested on each of the three experimental tests administered, attrition in the number of cases had occurred as a result of improperly answered tests and also by the inability to secure criterion measures on some of the subjects. The graphic item counts are based on a sample of 385 privates and 228 non-commissioned officers. From the graphic item counts, biserial r's were computed for each item of each of the three tests administered. The method by which these were computed was analagous to that used for the cadet officer sample.

The Associate Rating criterion was normalized by dividing the distribution into equal frequency eighths, and assigning the standard score equivalent of the midpoint of each eighth in a normal distribution to the criterion scores within that eighth.

Graphic item counts for the three projective tests were obtained separately for the samples of privates and non-commissioned officers. Each of these two samples had been fractionized into eight subsamples of equal frequency, after first arranging the cases in descending order according to their Associate Rating standard scores.

While the same general item analysis procedure was followed for the three tests, somewhat different techniques for dichotomizing the item responses were necessary for the different tests. For the method of dichotomizing responses used for the different tests, see Section B, part 5, under cadet officers.

Scoring keys were developed for those tests with promising validity based on the item analysis results. Items were selected whose biserial r's were significant at the 10% level of confidence. The value of biserial r necessary for significance was determined from the standard error of biserial r computed from the following formula[1]:

$$SE_{r_{bis}} = \frac{\frac{\sqrt{pq}}{z} - r^2_{bis}}{\sqrt{N}}$$

Significant biserial r's are a function of the percentage of cases in each dichotomized group, as well as the confidence level adopted. For a 50% dichotomous split, an r of .105 was required for significance at the 10% level of confidence in the private's sample. For the non-commissioned sample, an r of .137 was required for significance at this level of confidence.

One key was developed for the Picture Interpretation Test in addition to those obtained as above. This test was selected for special study in an effort to discover the nature of those items which yielded significant biserial r's. Two judges classified the significant items from this test into 13 categories suggested by the kinds of pictures which produced significant responses. Those non-significant items whose content fitted the classification scheme adopted were also placed into these categories. It was reasoned that if the classifications used were indicative of real relationships between item content and criterion, non-significant items in the same classification might show correlations with the criterion having the same direction as that found for the statistically significant items classified in the same category.

To check on the extent to which non-significant items were predicted with correct signs for the various categories, the following analysis was performed. The proportions of positive and negative biserial r's among all non-significant items classified were determined. Likewise the proportion of positive or negative items allocated to each category was determined. The differences between proportions in each category and in the total were then tested for significance. If these differences were significant, it was concluded that the classification of items within these categories was meaningful. By this procedure, an item classification key of 89 items was developed. The item-analysis key for this test contained 99 items.

Certain trainees had been eliminated from the item analysis

[1] Kelley, T.L., Fundamentals of Statistics. Cambridge: Harvard, 1947, p. 372.

because they lacked Associate Rating scores as a result of having been dropped, for various reasons, from the Leader's Course. There were 25 such cases who had complete sets of tests. Mean scores were obtained for this group on the Picture Interpretation Test and the Picture Fill-In Test by scoring their papers with the item analysis keys developed as described previously.

It was desired to compare the mean score for dropouts to the mean score of the total item analysis group. It was postulated that, if a positive correlation existed between criterion and test scores, dropout mean scores on the tests would be significantly lower than the means of the item analysis group. This assumes that dropouts, had they been rated, would have received relatively low Associate Rating scores.

In order to estimate the mean test scores of the item analysis group, it was necessary to score their papers with the item analysis keys. Rather than scoring test papers for the entire sample of 385 privates, 50 cases were selected at random from this group. A stratified random sampling technique was used since the 385 privates had been fractionated into eighths on the basis of Associate Ratings for item analysis purposes (see page 9). Furthermore, for both this group and the dropouts, a different keying of responses was used to simplify scoring than that used later for the cross-validation sample. Since the keying of items is arbitrary, the two keying methods used result in scores which differ only by a constant.

For testing the significance of the differences between means of the dropout and item analysis group, $\underline{t}$ tests were computed. The standard error for the difference between means was adjusted in the $\underline{t}$ formula to take account of the use of a stratified random sample.[1]

b. Cross-Validation Group

Using the item-analysis keys, the Picture Fill-In and Picture Interpretation Tests were scored for the cross-validation sample. The Picture Interpretation Test was also scored for the item classification key described above. The Leader's Self-Description Blank was scored by using the key furnished by the Personnel Research Section.

Reliability coefficients were computed for the Picture Interpretation and Picture Fill-In tests. The method of computation used was the correlation between scores from odd-and-even numbered items, augmented by the Spearman-Brown prophecy formula to estimate the reliability of the whole test.

[1] McNemar, Q., Psychological Statistics, John Wiley and Sons, Inc.: New York, 1949, pp. 333-336.

Validity coefficients were computed for the cross-validation sample. Scores on each of the three tests, obtained as described above, were correlated with the Associate Rating criterion. The correlations were computed separately for the Ft. Jackson and Ft. Belvoir samples, and for the two combined.

Before combining the two installations into a total sample, it was advisable to test whether mean Associate Rating scores for the two installations differed significantly. A critical ratio was computed testing the significance of this difference. If this ratio were not statistically significant, Associate Rating scores would not be converted to standard scores for computation of the validity coefficients.

To test whether the validity coefficients differed significantly for the two installations, critical ratios were computed for the difference between two sample correlation coefficients. An r to z transformation was made prior to this statistical test.

Two multiple correlation coefficients were computed, by the Wherry-Doolittle method, with Associate Ratings as the criterion variable in both cases, and as the predictor variables (1) Picture Interpretation Test and Picture Fill-In Test, and (2) Picture Interpretation Test, Picture Fill-In Test, and Leaders' Self-Description Blank. Only those trainees who had completed all three tests and had received an Associate Rating were utilized for this analysis.

IV. RESULTS

A. Results with Cadet Officers

1. Picture Interpretation Test, 1949 (DA AGO PRT-1775)

a. Item Analysis[1]

The proportion of items with statistically significant validities failed to exceed chance expectancy, indicating a lack of validity for the test with this sample. A scatterplot was prepared showing the relationship of the obtained validity coefficients of Group A (N = 211) vs. the corresponding coefficients of Group B (N = 223). The correlation between the two sets of validity coefficients was approximately zero, indicating little consistency of item validity from sub-sample to sub-sample. This finding, together with the low proportion of significantly valid items, strongly suggests that the test does not possess sufficient validity for the prediction of leadership with West Point cadets.

A qualitative investigation of those items for which combined validities exceeded chance expectancy did not yield logical categories or trends which were considered to be psychologically meaningful.

Table 4 shows the distribution of the item validities in the Picture Interpretation Test.

b. Cross-Validation

To substantiate the evidence from the item analysis, the validity of the Picture Interpretation Test was estimated by computing the correlation coefficient between the scoring keys derived on the item analysis samples and the Aptitude Rating leadership criterion. For Group A (N = 210) the scoring key yielded a correlation coefficient of .12. For Group B (N = 222) the correlation coefficient was found to be .07, indicating the lack of appreciable validity of this test for these samples.

[1] Item validities have been reported in detail for each of the tests in tables included in regular monthly progress reports submitted to the Department of the Army during the course of the study. Slightly different N's from sample to sample and from test to test are the result of incomplete data on a few cases in the sample tested.

Table 4. Distribution of Item Validities in the Picture Interpretation Test for Two West Point Samples.

Group A N = 211		Group B N = 225	
r	f	r	f
.00-.04	172	.00-.04	152
.05-.09	100	.05-.09	140
.10-.14	87	.10-.14	166
.15-.19	38	.15-.19	45
.20-.24	16	.20-.24	15
.25-.29	11	.25-.29	11
.30-.34	4	.30-.34	2
.35-.39	1	.35-.39	0
Total	429 items*	.40-.44	0
		.45-.49	1
		Total	432 items

* The total number of items for which it was possible to compute validity coefficients was 429 in Group A, since three items (no. 184, 344, and 362) yielded no responses in one category of the dichotomy.

2. Army Picture Story Test, Series B. 1950. Syracuse University Press.

a. Item Analysis

The item analysis of the Army Picture Story Test revealed only a chance proportion of statistically significant items. The scatterplot between the validity coefficients in Group A (N = 213) and Group B (N = 222) for the Army Picture Story Test indicated a near zero relationship, suggesting little consistency of item validity. However, as in the case of the Picture Interpretation Test, the subjects in the two samples, Group A and Group B, did respond similarly to individual items indicating a marked degree of inter-sample consistency.

A qualitative analysis of the significant items of the Army Picture Story Test also failed to disclose meaningful categories or trends in terms of postulated leadership characteristics.

Tabl‘ 5 shows the distribution of item validities in the Army Picture Story Test.

b. Pattern Item Analysis[1]

Of the 600 possible patterns of response in the Army Picture Story Test, 30% of the patterns showed significantly high criterion relationships at the 20% level of confidence in Group A (N = 213). The percentage of significant patterns may not be beyond chance expectations because of inter-pattern correlation. However, the degree to which the relationships among patterns affect the number of patterns appearing to possess significant validities is impossible to determine. Thus, a scoring key was constructed on the basis of the significant patterns by assigning a weight of +1 to patterns with positive validity at the 20% level of confidence, and -1 to patterns with negative validity at this level.

c. Cross-Validation

The validity of the item analysis keys of the Army Picture Story Test was estimated by calculating the Pearson product-moment correlation coefficient between test scores and the Aptitude Rating leadership criterion. Group A (N = 210) and Group B (N = 222) yielded validity coefficients of -.08 and -.05 respectively, indicating essentially zero validity for the test with these samples.

[1] Validities of each pattern have been reported in regular progress reports submitted to the Department of the Army during the course of the study.

Table 5. Distribution of Item Validities in the Army Picture Story Test for Two West Point Samples.

Group A		Group B	
N = 213		N = 222	
r	f	r	f
.00-.04	144	.00-.04	156
.05-.09	79	.05-.09	69
.10-.14	49	.10-.14	45
.15-.19	18	.15-.19	20
.20-.24	7	.20-.24	6
.25-.29	1	.25-.29	1
.30-.34	1	.30-.34	1
.35-.39	1	.35-.39	0
Total	300 items	.40-.44	0
		.45-.49	1
		.50-.54	0
		.55-.59	0
		.60-.64	1
		Total	300 items

Using the key derived by the pattern analysis on Group A, the scores of Group B (N = 222) were calculated. The Pearson product-moment correlation coefficient between the test scores and the Aptitude Rating leadership criterion was .01, indicating the lack of validity in the pattern key for this sample.

3. Picture Fill-In Test, Second Form, 1949 (DA AGO PRT-1726)

a. Item Analysis

The proportion of statistically significant items failed to exceed chance expectancy. The scatterplot between the correlation coefficients for Group A (N = 213) and Group B (N = 223) indicated approximately a zero relationship for the Picture Fill-In Test. Thus, again negative evidence was found in regard to the inter-sample consistency of item validity. As in the case of the two tests mentioned previously, a qualitative analysis of the significant items found in the two samples failed to reveal categories of responses which seemed to be psychologically meaningful.

However, there was evidence of considerable consistency between samples in the proportion of individuals who responded in the same way to the items in the test. This same indication of the inter-sample consistency of the responses was also found for the two tests discussed previously: The Picture Interpretation Test and the Army Picture Story Test.

The following table presents the distribution of item validities: (See Table 6).

b. Cross-Validation

Estimates of the validity of the Picture Fill-In Test were obtained by computing the Pearson product-moment correlation coefficients between the item analysis keys and the Aptitude Rating criterion of leadership. The validity of the test for the West Point samples was found to be approximately zero: in Group A (N = 210) the validity coefficient was .04 and in Group B (N = 222) the validity coefficient was .03.

4. West Point Personal Inventory, 1949 (DA AGO PRT-1756)

a. Item Analysis

As explained in Section III, B. 3. a. (4), the scoring key for this test was provided by the Personnel Research Section, AGO.

b. Cross-Validation

The West Point Personal Inventory was the only one of the [illegible] approached [illegible] validity for

Table 1. Distribution of Item Validities in the Picture Fill In Test for Two West Point Samples.

Group A		Group B	
N = 213		N = 223	
r	f	r	f
.00-.04	164	.00-.04	161
.05-.09	106	.05-.09	119
.10-.14	69	.10-.14	57
.15-.19	32	.15-.19	29
.20-.24	12	.20-.24	20
.25-.29	4	.25-.29	2
.30-.34	3	.30-.34	2
.35-.39	1	.35-.39	0
.40-.44	0	.40-.44	1
.45-.49	0	.45-.49	1
.50-.54	0	Total	392 items
.55-.59	0		
.60-.64	0		
.65-.69	1		
Total	392 items		

the West Point sample. The Pearson product-moment correlation coefficient between the scores on this test and the Aptitude Rating leadership criterion was .351, based on a sample of 426 cadets. This value of .351 is a statistically significant validity coefficient, since a value of .128 is required for significmnce at the 1% level of confidence.

B. Results with Enlisted Personnel

1. Item-Analysis Group: All Tests

a. Item Analyses

Tables 8, 9, and 10 show the distribution of biserial r's obtained for items in each test. The sign of the coefficients was disregarded in tabulating these results, since keying the responses to each item is quite arbitrary.

Table 7 summarizes the number of significant items found for the two samples at the 10% level of confidence in each of the three tests.

Table 7. Number of Significant Items for Three Experimental Personality Tests.

	Total No. of Items	Number of Significant Items
Picture Interpretation Test		
1. Privates (N = 385)	432	99
2. Non-Coms (N = 228)	432	53
Picture Fill-In Test		
1. Privates	392	130
2. Non-Coms	392	38
Picture Story Test		
1. Privates	300	46
2. Non-Coms	300	45

From this table, it is apparent that more significant items are found for the sample of privates than is true for non-commissioned officers. By inspecting the individual items, it is also apparent that those items found significant in the private's sample generally are not found significant in the non-commissioned officer sample. The results indicate, furthermore, that the Picture Fill-In Test and Picture Interpretation Test are functioning in the sample of privates at a level appreciably better than chance expectancy. Since the 10% level of confidence was adopted, chance, on the average, would result in approximately 39 significant items for the Picture Fill-In Test, 42 for the Picture Interpretation Test, and 30 for the Army Picture Story Test, assuming no correlation among the items in each test.

Table 8. Distribution of Item Validities in the Picture Interpretation Test for Two Enlisted Samples

A. Privates

N = 395

r	f
.00-.04	152
.05-.09	131
.10-.14	90
.15-.19	44
.20-.24	13
.25-.29	1
.30-.34	1
Total	432 items

B. Non-Commissioned Officers

N = 225

r	f
.00-.04	168
.05-.09	107
.10-.14	88
.15-.19	45
.20-.24	15
.25-.29	2
.30-.34	2
.35-.39	1
.40-.44	2
.45-.49	0
.50-.54	1
.55-.59	0
.60-.64	0
.65-.69	0
.70-.74	1
Total	432 items

Table 9. Distribution of Item Validities in the Army Picture Story Test for Two Enlisted Samples.

A. Privates
N = 385

r	f
.00-.04	132
.05-.09	105
.10-.14	51
.15-.19	11
.20-.24	1
Total	300 items

B. Non-Commissioned Officers
N = 228

r	f
.00-.04	112
.05-.09	91
.10-.14	59
.15-.19	29
.20-.24	7
.25-.29	2
Total	300 items

Table 10. Distribution of Item Validities in the Picture Fill-In Test for Two Enlisted Samples.

A. Privates
N = 385

r	f
.00-.04	140
.05-.09	99
.10-.14	71
.15-.19	42
.20-.24	30
.25-.29	9
.30-.34	1
Total	392 items

B. Non-Commissioned Officers
N = 228

r	f
.00-.04	176
.05-.09	109
.10-.14	72
.15-.19	27
.20-.24	5
.25-.29	2
.30-.34	1
Total	392 items

For the Army Picture Story Test, the number of significant items was about 1.6 times what would be expected on the basis of chance. This figure is not large enough to warrant concluding that non-chance relationships are involved, particularly since the assumption of non-correlation among items is untenable. It is possible that a pattern analysis might reveal more convincing evidence for the validity of this test, in view of the triad form of item responses. Experience with the pattern analysis performed for the cadet officer sample did not encourage a parallel analysis for the enlisted sample. In view of the relatively low number of significant items found for this test, it was not administered to the cross-validation sample.

For the Picture Interpretation Test, about 2.3 times as many items were found significant for the sample of privates as would be expected on the basis of chance. For the Picture Fill-In Test, this figure was about 3.3. These tests were selected to be administered to the cross-validation sample, since the evidence is suggestive of validity.

b. Comparison of Item-Analysis Group and Dropouts

Table 11 shows *t* tests for the significance of the mean differences between dropouts and the item analysis group. It had been expected that dropouts would show lower mean test scores for the Picture Fill-In Test and for the Picture Interpretation Test. Such is the case, and furthermore, this difference is significant at the 5% level of confidence for the Picture Interpretation Test, and at the 1% level of confidence for the Picture Fill-In Test.

2. Cross-Validation

a. Reliabilities and Related Statistics

Table 12 summarizes reliabilities and related statistics calculated from the cross-validation sample for the Picture Fill-In Test and the Picture Interpretation Test.

These results indicate that scores for these instruments are sufficiently reliable to be useful for large-scale classification purposes.

b. Validities

(1) Item Analysis Keys

Table 13 summarizes validity coefficients for the keys developed from item analysis of the Picture Fill-In and Picture Interpretation Tests.

Table 11. Significance of Mean Differences Between Dropouts and Item Analysis Group for Picture Fill-In and Picture Interpretation Tests

	Picture Fill-In	Picture Interpretation
Mean of Dropout Group	68.0	49.4
S. D. of Dropout Group	21.1	9.6
N of Dropout Group	25	25
Mean of Item Analysis Group	90.6	56.2
S. D. of Item Analysis Group	21.6	11.8
N of Item Analysis Group	50	50
t Ratio between groups	4.26**	2.44*

* Significant at 5% level

** Significant at 1% level

Table 12. Reliabilities and Related Statistics Estimated from Cross-Validation Samples at Leaders' Schools for the Picture Interpretation and Picture Fill-In Tests

Picture Interpretation Test

	N	Mean	S	S^2	S.E. of Meas.	Odd-Even Rel.	Spearman Brown
Ft. Jackson	158	61.34	10.31	106.25	4.50	.81	.90
Ft. Belvoir	138	60.51	9.58	91.80	5.58	.66	.80
Total Sample	296	60.95	9.97	99.47	5.08	.74	.85

Picture Fill-In Test

	N	Mean	S	S^2	S.E. of Meas.	Odd-Even Rel.	Spearman Brown
Ft. Jackson	145	97.39	12.53	157.08	5.46	.81	.90
Ft. Belvoir	133	91.91	18.61	346.25	5.58	.91	.95
Total Sample	278	94.77	15.85	251.28	6.34	.84	.91

Table 13. Validity Coefficients for the Picture Fill-In and Picture Interpretation Tests in the Privates' Sample.

	Ft. Jackson	Ft. Belvoir	Combined
Picture Fill-In Test			
r	.19*	.24**	.19**
N	124	132	256
M	97.0	91.3	94.1
S. D.	13.2	18.8	16.7
Picture Interpretation Test			
r	.36**	.16	.25**
N	133	136	269
M	61.3	60.4	60.8
S. D.	9.9	9.7	9.8

* Significant at the 5% level of confidence.

** Significant at the 1% level of confidence.

All correlations reported are significantly different from zero at the 5% level of confidence with the exception of r = .16 for the Ft. Belvoir sample on the Picture Interpretation Test. This correlation approaches significance at the 5% level of confidence so closely that it, too, is unlikely to be considered a sampling fluctuation from a correlation of zero.

Differences in mean criterion scores for the two installations were not statistically significant at the 5% level of confidence. On the basis of this result, validity coefficients were computed using Associate Ratings in raw scores rather than converting them to standard scores.

Tests of the significance of the difference between sample correlation coefficients were performed in order to compare validity coefficients at the two installations. These results fail to reveal a difference, significant at the 5% level of confidence, between the validity coefficients at the two installations. The tests seem to be about equally valid in both of the cross-validation samples.

(2) Item Classification Key for the Picture Interpretation Test.

Table 14 presents validity coefficients for the item classification key, for the item analysis key, for the combination of the two, and the correlation between the item analysis and item classification keys. The criterion used for the validity coefficients was that of Associate Ratings.

The fact that the item classification key correlates significantly with the item analysis key is interpreted to mean that these categories have an appreciable degree of internal consistency. The failure of the items classified to correlate significantly with the criterion is an indication of their consistent lack of validity for this criterion even in a cross-validation sample.

(3) Multiple Correlation

Table 15 shows the multiple correlations for the total cross-validation sample, and the intercorrelations from which the R's were computed.

The standard errors of these R's are such as to render no combination of the tests appreciably superior in prediction to another, nor, indeed to the Leaders Self-Description Blank alone.

Table 14. Validity Coefficients of the Item Analysis and Item Classification Keys for the Picture Interpretation Test and the Correlation between Keys.

Retention

	Ft. Jackson	Ft. Belvoir	Combined
Item Classification Key Alone	-.01	-.18	-.06
N	134	136	270
Item Classification Key and Item Analysis Key	.25	.06	.16
N	133	135	267
Item Analysis Key Along	.36	.16	.25
N	133	136	269
Item Analysis Key vs. Item Classification Key	.54	.26	.40
N	158	138	296

Table 15. Intercorrelations among Predictor Variables, Correlations with Associate Rating Criterion, and Multiple Correlations for Cross-Validation Sample

(N = 234 cases in cross-validation sample with all 4 measures)

	Assoc. Rating	P.F.I.	P.I.T.	L.S.D.B.
Associate Rating (0)	....			
Picture Fill-In (1)	.17	....		
Picture Interpretation (2)	.23	.18	....	
Leaders' Self Description Blank (3)	.30	.34	.37	....

Multiple Correlations

$R_{0.12} = .27$

$R_{0.123} = .32$

V. CONCLUSIONS

The conclusions to be derived from the results described in the preceding section will be presented below in the sequence corresponding to the objectives set forth in Section II.

A. Validity of the three objective projective tests for measuring leadership performance of West Point cadets.

1. <u>In none of the three tests does the aggregation of items correlate with the leadership criterion (Aptitude Rating) appreciably beyond chance expectation</u>. This follows from the fact that the number of items found statistically significant at a given level of confidence is no greater than the number which would manifest that degree of validity through sampling fluctuations about a true validity of zero.

2. When scoring keys are developed independently for each of two representative samples by keying items whose individual validity coefficients appear to be statistically significant, there is practically no better than chance correspondence in the items comprised within the two keys. Hence, it can be inferred that <u>there is inadequate consistency of the scoring keys from sample to sample</u>. <u>Furthermore, each of the two keys for each test correlates approximately zero with the leadership criterion in its cross-validation sample</u>.

3. Duplicating previous findings of the Personnel Research Section, the <u>West Point Personal Inventory</u> is found to <u>correlate appreciably and significantly with the leadership criterion</u>. In addition to demonstrating the validity of the test, this indicates that <u>the criterion is predictable</u>.

<u>Discussion</u>: This investigation failed to reveal a stable and valid method of keying the responses to three objective projective tests so as to predict leadership among West Point cadets with better-than-chance efficiency. This failure cannot be attributed totally to inadequacy of the leadership criterion, the Aptitude Rating, for it is predictable from scores on the <u>West Point Personal Inventory</u>.

B. Validity of three objective projective tests for measuring leadership performance of Army non-commissioned personnel.

1. a. Among privates enrolled at Leaders' Schools, the <u>Army Picture Story Test</u> failed to yield a proportion of items which correlates with the Associate Rating appreciably beyond chance expectation. However, in both the <u>Picture Interpretation Test</u> and the <u>Picture Fill-In Test</u>, considerably more than 10 per cent of the items were valid at least at the 10% level of confidence. Hence it was possible to develop a scoring key for each of these two tests with some expectation that the resulting scores would correlate appreciably with the leadership criterion in new samples.

c. Among non-commissioned officers enrolled in Leaders' Schools, none of the three tests yields an aggregation of items which correlates with the leadership criterion (Associate Rating) appreciably beyond chance expectation. This follows from the fact that the number of items found to be statistically significant at a given level of confidence is no greater then the number which would manifest that degree of validity through chance fluctuations about a true validity of zero.

2. a. When short keys for the Picture Interpretation and Picture Fill-In Tests are applied in new samples of privates enrolled in Leaders' Schools, the Spearman-Brown reliabilities of the scores are .85 and .91, respectively.

The correlation of these scores with the Associate Rating criterion yields a validity coefficient of .25 for the Picture Interpretation Test and .19 for the Picture Fill-In Test: these validity coefficients are significant at beyond the 1% level of confidence.

These keys also discriminate, on the average, between trainees who, for various reasons (including lack of leadership potential), are separated early in the program and those who are graduated.

b. (No cross-validation was undertaken with non-commissioned officers, in view of the negative results of the item analysis.)

3. a. The Leaders' Self-Description Blank is found to correlate appreciably and significantly with the Associate Rating criterion for privates. The validity coefficients of this and the other two tests do not differ from one another at the 5% level of confidence.

b. The multiple correlation between the criterion and the two projective tests is not appreciably higher than the validity coefficient of the Picture Interpretation Test alone. The multiple correlation between the criterion and the two projective tests plus the Leaders' Self-Description Blank is not appreciably higher than the validity coefficient of the last-named test alone.

c. The Picture Interpretation and Picture Fill-In Tests are virtually uncorrelated with one another. This, in the light of their relatively high reliabilities, suggests considerable independence of the factors measured. Both of the tests correlate somewhat more highly with the Leaders' Self-Description Blank, although still at a level far below their reliability coefficients.

Discussion: The Army Picture Story Test revealed evidence of validity in neither the sample of non-commissioned officers nor of privates. The Picture Interpretation Test and the Picture Fill-In Test evidenced no validity in the non-commissioned officer sample but both showed appreciable and significant validity in the sample of privates.

It should be recognized, in this connection, that the Associate Rating criterion does not necessarily represent a total appraisal of leadership performance. This a priori consideration is substantiated by the fact that this particular criterion was only slightly correlated with other assessments of leadership obtained for these personnel in the Leaders' School. It is, therefore, conceivable that the validity of these tests for a more comprehensive criterion would be considerably higher.

B. Inferences concerning the general promise of the [illegible] projective type of test.

The fact that these tests correlate significantly and appreciably with a limited criterion of leadership performance among privates, and their relative independence of one another and the Leaders' Self-Description Blank, combine to suggest that we have here a type of instrument that may constitute an important contribution to the techniques for identifying potential leaders. However, in the light of the results, this statement must be made with certain limitations. In the first place, the failure of the tests to function with samples of cadets and non-commissioned officers indicates that they may be most effective with personnel of relatively little sophistication and limited experience. Secondly, it should be recalled that the predictive value of the tests has been shown only for the Associate Rating criterion: this criterion probably does not reflect all aspects of leadership performance.

Conversely, the positive results that have been obtained with these measures by no means define the limits of their value. They may, for example, correlate as well or better with other aspects of Army leadership performance, or with similar criteria in other types of leadership situations, such as under field conditions. Their validity for other types of performance involving personality factors, e.g., adaptability to stress situations, is likewise in the realm of possibility. At present, these are, of course, moot points. But the promise evidenced by these tests for the limited criterion and situations involved in this study imply the advisability of exploring their value for other criteria and situations.

Such further work should, however, be done with revisions of the present tests. Many items which failed to show validity in this study could well be eliminated, and replaced by other items constructed along lines which the present study suggests are more fruitful. Suggestions for such future modifications are incorporated in the following section.

VI. IMPLICATIONS AND RECOMMENDATIONS

The major implication of the present study is that there are at least two of these objective projective tests (the Picture Interpretation and Picture Fill-In Tests) which evidence validity for leadership criteria. Yet, the results indicate that there is considerably more work to be done before these instruments will have the utility which practical considerations require.

For one thing, the evidence is that the tests, as presently constituted, are valid only within a limited segment of personnel in connection with whom leadership assessments would be made, namely, the youngest and least experienced group. It behooves us to consider why this should be the case and what may be done to augment the range of utility of these instruments.

For another thing, the evidence does not suggest that these tests should be used to supplant or supplement the Leaders' Self Description Blank, which is currently available for assessing leadership potential of enlisted personnel. The reason for this statement is that neither of the two new tests correlates better with the criterion than the Leaders' Self-Description Blank, nor is the multiple correlation of the three tests appreciably higher than the validity of the Leaders' Self-Description Blank. However, it should be noted that the failure of the new tests to add appreciably to the validity of the Leaders' Self-Description Blank is not a function mainly of the degree of overlap among the measures; rather, the failure is primarily attributable to the relatively low (even though significant) validity manifested by the new tests. It would thus seem that revision of these tests so as to augment their predictive value could well result in useful additions to current selection instruments.

There are therefore two dimensions along which revision should proceed: (1) broadening of the range of personnel to which the tests are applicable, and (2) intensifying the discrimination power of the items within this range. Conceivably, of course, the two objectives may not be attainable with a single form of each test, so that a different form may be required for each of several types of personnel, e.g., commissioned and enlisted.

Examination of the characteristics of the various items serves as a source of hypotheses for effecting these improvements. For example, extending the effective range of these tests to reach more highly educated and experienced personnel seems to require modification in two respects: (1) the situations depicted should be more appropriate to the interests and vital experiences of such personnel; at present, most of the situations appear to be rather simple, socially and motivationally, making it difficult to elicit real identification and projection on the part of more sophisticated subjects; (2) in spite of the projective approach, there may still be too much transparency in regard to what constitutes "right" and "wrong" answers; the use of more subtle [illegible] and responses, and possibly from the use of forced-choice [illegible] may produce marked improvements in the [illegible] of such tests for higher-level personnel.

The latter point, namely forced-choice coupling of response alternatives, may also be a fruitful method of improving the discrimination power of the items for any level of personnel. But more than this should be done to attain this objective: new items should be prepared whose content is likely to generate validity. Clues from the [illegible] for this work.

Among these clues are the following:

A. For Picture Interpretation Test, items depicting high or low prestige activities seem more frequently to be valid. The same is true of items portraying leadership or dominance.

B. For Picture Fill-In Test, the discriminating responses seem to be those which reflect of extra-punitiveness and intra-punitiveness. Social appropriateness or inter-personal skill seems to be another dimension along which a number of the items discriminate.

The elimination of the many non-discriminating items from the present forms of both of these tests, and their replacement by items constructed along the lines suggested above may well be productive of greater validity.

Most of the facts needed to effect modifications along lines suggested above already are at hand. For example, the types of items which seem most productive of validity could readily be classified beyond the point already accomplished. Also, as regards the forced-choice possibility, preference values (defined in terms of frequency of response selection) and validity coefficients are available for all items.

The basic decision that must first be made is whether or not to proceed further with instruments of this type. It is felt that the evidence disclosed in this investigation is that the Army possesses at least two new-type tests which show appreciable validity for leadership criteria, while being, at the same time, relatively independent of existing instruments used for leadership assessment. These considerations cogently denote the promise inherent in further exploration and development along these lines. This is particularly true in view of the critical need for techniques of leadership assessment and the paucity of existing means for meeting this need.

PRS REPORT NO. 958

VALIDATION OF THREE OBJECTIVELY SCORED PICTORIAL TESTS OF PERSONALITY FOR THE ASSESSMENT OF LEADERSHIP

By

Raymond A. Katzell, Ph.D.
Principal Investigator

James J. Kirkpatrick
Project Supervisor

Frederick O. Carleton
Research Assistant

Nathan Rosenberg
Research Assistant

PSYCHOLOGICAL SERVICES CENTER
SYRACUSE UNIVERSITY

A CONTRACT RESEARCH REPORT TO

PERSONNEL
RESEARCH
SECTION
PR & P BR-PERS BUR-AGC

ABSTRACT

U. S. Dept. Army. The Adjutant General's Office. Personnel Research Branch. Validation of three objectively scored pictorial tests of personality for the assessment of leadership. Personnel Research Branch Report 958, 31 May 1952. 39 pp. Washington: American Documentation Institute c/o Library of Congress, Document No. 3975, microfilm, $2.50; photocopy, $6.25.--3 multiple-choice projective tests were administered to each of 3 Army samples: privates and noncommissioned officers in Leaders' Course, and West Point cadets. The Army Picture Story Test was patterned after the Thematic Apperception Test; the Picture Fill-In Test was patterned after the Picture Frustration Test; the Picture Interpretation Test involved selective identification with individuals depicted in various roles and activities. Item analyses (against leadership ratings made by associates) failed to reveal better than chance distributions of item validities for the samples of cadets and noncommissioned officers. For the privates, it was possible to develop stable scoring keys for the last 2 tests named. Cross-validation produced validity coefficients of .19 for the Picture Fill-In and .25 for the Picture Interpretation Test. Suggestions were inferred for improvement of the tests.

PRS REPORT NO. 958

VALIDATION OF THREE OBJECTIVELY SCORED PICTORIAL TESTS OF PERSONALITY FOR THE ASSESSMENT OF LEADERSHIP

by

Raymond A. Katzell
Principal Technician
Psychological Services Center
Syracuse University
Syracuse, New York

James J. Kirkpatrick
Project Supervisor

Frederick O. Carleton
Research Assistant

Nathan Rosenberg
Research Assistant

FINAL REPORT 31 May 1952

DEPARTMENT OF THE ARMY
Project No. 29534100 Subtask 76
RESEARCH CONTRACT NO. DA-49-083 OSA-64

PERSONNEL RESEARCH SECTION PR 3476

PRS reports are primarily technical. While conclusions affecting military policy or operations may appear in them, they are not intended as a basis for official action. Findings and conclusions contained in PRS reports are intended to guide the conduct of further research. When research findings suggest recommendations for administrative action, such recommendations are made separately to the appropriate military agency.

BRIEF

THE EFFECTIVENESS OF PICTORIAL TESTS OF PERSONALITY IN THE ASSESSMENT OF LEADERSHIP
(Based on PRS Report 958)

STATEMENT OF THE PROBLEM

One of the most important problems of personnel management in the Army is identifying (1) men with qualities of leadership, and (2) men who can readily be trained as officers and noncommissioned leaders. Various tests and procedures have been used; some have been more successful than others. The purpose of the present study was to evaluate three new pictorial tests of personality as predictors of leadership ability at the U. S. Military Academy and in Leader's Schools.

RESULTS

1. For a sample of privates enrolled at Leader's Schools, two out of three of the tests gave a better than chance differentiation between men rated high by their associates and men rated low.

2. The power of these tests to distinguish between high rated privates and low rated privates is about the same as a test already in use: "the Leaders Self-Description Blank." The new tests are not closely related to the old one.

3. However, for noncommissioned officers at Leader's Schools, no one of the tests differentiated between high rated men and low rated men.

4. None of the tests gave scores related to Aptitude-for-Service Ratings for cadets at the Military Academy.

CONCLUSIONS

1. The validity of pictorial tests used in this experiment was insufficient to add significantly to the validity attainable with Self-Description Blanks previously developed by the Personnel Research Section, Personnel Research and Procedures Branch, The Adjutant General's Office.

2. In order for pictorial tests to become effective leadership predictors, it appears necessary to effect improvement in item content and format. Whether such improvement would be sufficient to warrant the cost is debatable.

WORK SUMMARY

Three new tests, Picture Interpretation Test, Army Picture Story Test, and the Picture Fill-in Test, were administered to 216 cadets at the Military Academy and 958 enlisted men in Leader's Schools. In addition, the West Point Personal Inventory was administered to the cadets, and the Leaders Self-Description Blank to the enlisted men.

Responses to test items and total test scores were compared with an independent measure of leadership, the Aptitude-for-Service Rating for cadets or the associate rating for enlisted men and verified on additional groups of 238 cadets and 268 enlisted men.

TABLE OF CONTENTS

	Page
LIST OF TABLES	iii
PREFACE	iv
SUMMARY	v
I. INTRODUCTION	1
II. OBJECTIVES	1
III. METHOD	3
A. Tests	4
1. Picture Interpretation Test	4
2. Army Picture Story Test	5
3. Picture Fill-In Test	5
4. West Point Personal Inventory	6
5. Leaders' Self-Description Blank	6
B. Situational Validity at the Commissioned Officer Level	7
1. Sample	7
2. Tests	7
3. Criterion	7
4. Procedure	8
5. Analysis of Data	9
a. Item Analysis	9
b. Pattern Item Analysis	10
c. Cross-Validation	10
C. Enlisted Personnel Validity Studies	11
1. Samples	11
a. Item Analysis Group	11
b. Cross-Validation Group	11
2. Tests	11
a. Item Analysis Group	11
b. Cross-Validation Group	12
3. Criteria	12
a. Item Analysis Group	12
b. Cross-Validation Group	14
4. Procedure	14
a. Item Analysis Group	14
b. Cross-Validation Group	14
5. Analysis	15
a. Item Analysis Group	15
b. Cross-Validation Group	18

Page

IV. RESULTS 20

A. Results with Cadet Officers 20
1. Picture Interpretation Test 20
a. Item Analysis 20
b. Cross-Validation 20
2. Army Picture Story Test 22
a. Item Analysis 22
b. Pattern Item Analysis 22
c. Cross-Validation 22
3. Picture Fill-In Test 24
a. Item Analysis 24
b. Cross-Validation 24
4. West Point Personal Inventory 24
a. Item Analysis 24
b. Cross-Validation 24

B. Results with Enlisted Personnel 26
1. Item Analysis Group: All Tests 26
a. Item Analyses 26
b. Comparison of Item-Analysis Group and Dropouts 29
2. Cross-Validation 29
a. Reliabilities and Related Statistics 29
b. Validities 29
(1) Item Analysis Keys 29
(2) Item Classification Key for the Picture Interpretation Test 32
(3) Multiple Correlation 32

V. CONCLUSIONS 35

A. Validity of the three objective projective tests for measuring leadership performance of West Point cadets 35

B. Validity of three objective projective tests for measuring leadership performance of Army non-commissioned personnel 35

C. Inferences concerning the general promise of the objective projective type of test 37

VI. IMPLICATIONS AND RECOMMENDATIONS 38

LIST OF TABLES

Table		Page
1.	Intercorrelations among Various Criteria for Two Enlisted Samples	13
2.	Number of Cases in Item-Analysis and Cross-Validation Samples	15
3.	Critical Ratios and Variance Ratios for Testing Differences in Mean Associate Rating Scores	16
4.	Distribution of Item Validities in the Picture Interpretation Test for Two West Point Samples	21
5.	Distribution of Item Validities in the Army Picture Story Test for Two West Point Samples	23
6.	Distribution of Item Validities in the Picture Fill-In Test for Two West Point Samples	25
7.	Number of Significant Items for Three Experimental Personality Tests	26
8.	Distribution of Item Validities in the Picture Interpretation Test for Two Enlisted Samples	27
9.	Distribution of Item Validities in the Army Picture Story Test for Two Enlisted Samples	28
10.	Distribution of Item Validities in the Picture Fill-In Test for Two Enlisted Samples	28
11.	Significance of Mean Differences between Dropouts and Item Analysis Group for Picture Fill-In and Picture Interpretation Tests . .	30
12.	Reliabilities and Related Statistics Estimated from Cross-Validation Samples at Leaders' Schools for the Picture Interpretation and Picture Fill-In Tests	30
13.	Validity Coefficients for the Picture Fill-In and Picture Interpretation Tests in the Privates' Sample.	31
14.	Validity Coefficients of the Item Analysis and Item Classification Keys for the Picture Interpretation Test and the Correlation between Keys.	33
15.	Intercorrelations Among Predictor Variables, Correlations with Associate Rating Criterion, and Multiple Correlations for Cross-Validation Sample.	34

PREFACE

This is a report of a research study initiated in May, 1950, and performed under Contract Number DA-49-083 OSA-64, negotiated under authority of Section 2 (c) (5), Act of February 19, 1948 (Public Law 413 - 80th Congress).

The study was conceived and executed in association with the staff of the Personnel Research Section, The Adjutant General's Office, Department of the Army. Particularly noteworthy were the contributions made by Drs. D. E. Baier, H. E. Brogden, R. Perloff, E. K. Taylor, and the late Dr. C. I. Mosier.

On the staff of the contractor, indispensable collaboration was furnished by a number of individuals in addition to those whose names appear, somewhat arbitrarily, in authorship. Among them are Dr. Ernst G. Beier, Mr. D. K. Hable, Miss Mildred E. Leonard, Mr. Rolland Tougas, and Mrs. Elizabeth R. Coleman and her scoring staff.

The cooperation of the authorities at Forts Belvoir, Dix, Jackson, and Knox, and at the U. S. Military Academy, notably Lt. Col. Raymond Rumpf, Maj. Herman F. Smith, and Dr. Douglas Spencer, was invaluable in the acquisition of data.

To these individuals, and to many others of whom space or memory preclude the mention, the authors express their gratitude.

SUMMARY

A. Problem

1. To identify those items, in each of three new objectively scored projective tests, which discriminate between superior and inferior leaders among West Point cadets and enlisted trainees in Leaders Schools.

2. To determine the validity and reliability of the resulting scoring keys for the assessment of leadership in new samples of personnel.

3. To compare the validity of these keys with that of a biographical inventory currently used by the Army for leadership assessment.

4. To factor analyze the several tests found valid with West Point cadets along with other leadership measures, in order to investigate basic personality factors intrinsic to such measures.

B. Method

1. The Tests --

a. Picture Interpretation Test - involves elective identification with individuals depicted in various roles and activities.

b. Army Picture Story Test - involves the ranking of statements with regard to their appropriateness in describing each of a series of pictures.

c. Picture Fill-In Test - entails the rating of appropriateness of rejoinders in conversational situations depicted in cartoons.

d. West Point Personal Inventory - a series of biographical and self-descriptive questions, used with West Point cadets.

e. Leaders Self-Description Blank - a series of biographical and self-descriptive questions, used with Leaders School trainees.

2. The Criteria --

a. The West Point Aptitude Rating was used as the measure of leadership performance of West Point cadets. This is a composite rating on leadership made by the cadet's peers and tactical officer.

b. The Associate Rating, mainly a nomination rating by peers, was employed as the standard of leadership performance of Leaders School trainees.

3. The three projective tests were administered to 454 West Point cadets and 958 Leaders School trainees. The West Point Personal Inventory was also administered to these cadets. Criterion data were obtained for as many of these individuals as feasible.

4. All items on the three projective tests were biserially correlated with the criterion. This was done for four groups of subjects, as follows: two randomly selected groups of cadets, numbering 213 and 223, respectively; 385 privates enrolled in Leaders Schools; 228 non-commissioned officers enrolled at Leaders Schools.

5. Scoring keys were developed from this analysis, those items being keyed which had criterion correlations minimally significant at the 10% level of confidence.

6. The Picture Fill-In and Picture Interpretation Tests, and the Leaders Self-Description Blank, were administered to a new sample of 296 privates enrolled at Leaders Schools. Criterion data were secured for these individuals. Validity and reliability statistics were computed for this group.

C. Results --

1. In the two samples of West Point cadets, there was no better than a chance relationship between responses to the items on all three projective tests and the Aptitude Ratings received by the cadets. The West Point Personal Inventory had a correlation of .35 with the criterion in the two samples combined.

2. Similar negative results were obtained in the item analysis of the tests against Associate Ratings of non-commissioned officers enrolled in Leaders Schools.

3. In a sample of privates enrolled at Leaders Schools, it was possible to identify in two of the three projective tests an appreciably larger-than-chance number of items that distinguished between the higher- and lower-rated men. These two tests were the Picture Interpretation and Picture Fill-In Tests.

4. These tests, when scored for the new sample of privates by the scoring key developed on the first group, yielded validity coefficients of .25 for the Picture Interpretation Test, and .19 for the Picture Fill-In Test. The Leaders Self-Description Blank had a validity coefficient of .30 in this sample. Each of these coefficients differs significantly from zero at the 1% level of confidence.

5. The split-half reliability coefficients, augmented by the Spearman-Brown Prophecy formula, were .85 for the Picture Interpretation Test and .91 for the Picture Fill-In Test in the cross-validation sample.

6. The correlations of the two tests with each other and with the Leaders Self-Description Blank were all low and positive.

D. Conclusions

1. The three projective tests, as now constituted, are of no value for leadership assessment of West Point cadets.

2. These tests are also of no value for leadership assessment of non-commissioned officers in Leaders Schools.

3. Scoring keys were developed for both the Picture Fill-In and Picture Interpretation Tests on a sample of privates in Leaders Schools. Scoring the two tests for a new sample of privates by means of these keys yielded scores which were significantly correlated with the criterion of leadership in Leaders Schools.

4. The biographical inventories (West Point Personal Inventory and Leaders Self-Description Blank) showed significant criterion correlations in their respective samples.

5. Among the privates, the two valid projective tests did not add appreciably to the predictive power of the biographical inventory when combined with it in a multiple regression equation. Nonetheless, their correlations with the inventory are low (about .35), as is their correlation with one another (.18).

6. It is inferred that the Picture Fill-In Test and the Picture Interpretation Test show considerable promise as techniques for leadership assessment, although improvements are needed to translate this promise into a state of practical utility. Suggestions are made manifest in this study as to how improvements may be effected in regard to: (1) power to discriminate more accurately between superior and inferior leaders, and (2) extending the range of personnel with whom such tests would be useful.

7. In view of the lack of validity of the projective tests among West Point cadets, it was not meaningful to proceed with the factor analysis designed to reveal the basic personality factors common to these and other measures of leadership, so that this objective of the study could not be achieved.

I. INTRODUCTION

The identification of men with high potentialities as leaders is understandably a matter of prime importance to the Army. Accordingly, a considerable amount of research has been done or sponsored by the Army on techniques for accomplishing such identification.

Although it is commonly believed that non-intellective factors are of major importance in determining a man's leadership performance, methods for measuring such factors still leave much to be desired in the way of validity and accuracy. In recent years, the evidence has grown more suggestive that projective tests[1] may have promise along these lines. However, these tests are typically time consuming to administer and score, and typically require trained psychologists for their interpretation. These characteristics are manifestly unsuited for large-scale military classification purposes.

To circumvent these deficiencies, The Personnel Research Section of the Adjutant General's Office undertook the preparation of several tests which are fundamentally projective in nature but which are amenable to group administration and objective (even machine) scoring. When any new test is constructed, the questions of its validity and what it measures immediately arise. These questions become even more urgent when the test represents a radically new departure. Thus, in the case of the new objective projective tests, not only are their particular validities unknown, but also subject to question are the issues of the general fruitfulness of the approach and of the underlying psychological dimensions measured by such techniques.

The research described in this report was undertaken in an effort to shed light on these questions.

II. OBJECTIVES

More specifically, the objectives of this research may be described as follows:

A. To ascertain the validity of each of three objective projective tests for measuring leadership performance of Army commissioned personnel.

1. To determine the correlation of each item with a criterion of leadership performance, on the basis of which to develop a scoring key for each test.

[1] A projective test requires the examinee to interpret or structure a stimulus situation which lends itself to a variety of meanings, and thereby to reveal aspects of his personality.

2. To ascertain for each of these scoring keys its reliability and validity against a criterion of leadership performance at the level of commissioned personnel.

3. To compare the relative validities of these tests with one another, and with a self-description questionnaire.

B. To ascertain the validity of each of three objective projective tests for measuring leadership performance of Army non-commissioned personnel.

1. To determine the correlation of each item with a criterion of leadership performance, on the basis of which to develop a scoring key for each test.

2. To ascertain for each of these scoring keys its reliability and validity against a criterion of leadership performance at the level of non-commissioned personnel.

3. To compare the relative validities of these tests with one another, and with a self-description questionnaire.

C. From these data, to infer the general promise of this type of test, and to deduce indications of which lines of future development seem most fruitful.

It was also hoped originally to factor analyze the relationships among these tests, together with other personality measures including ratings and behavior measures, with the objective of determining basic personality factors gauged by such variables. Since the non-test variables were more appropriate and available in the commissioned personnel situation (West Point), the plan was to perform this analysis in connection with the data obtained from that sample. However, it was discovered in the course of the research that the projective tests were virtually uncorrelated with the leadership criterion in this situation, thus making the planned analysis pointless.

III. METHOD

The general plan of the study consisted of administering the tests to samples of personnel who were representative of the two levels of leadership activities for which such tests might be valuable assessment techniques. An additional requirement for selecting the samples was that the personnel be assigned to situations in which criteria of leadership performance would be available.

In accordance with these standards, cadets in the upper classes of the United States Military Academy at West Point were chosen as the sample whose characteristics and activities were approximately representative of personnel to be assessed for potential leadership at the commissioned level.

Students at Leadership Schools were selected as suitable for representing potential leaders at the non-commissioned level. This group is actually composed of two subgroups, as regards age, background, and previous experience: privates and non-commissioned officers. It was deemed advisable to investigate separately the validity of the tests for each of the two subgroups.

Thus, there were three categories of personnel who were the subjects of the investigation: West Point cadets, privates assigned to Leadership Schools, and non-commissioned officers assigned to Leadership Schools.

The research design involved the following steps for each of the categories of personnel:

1. Administering the three objective projective tests to samples of the personnel.

2. Collection of criterion data for these individuals.

3. Correlation of the test items against the criterion.

4. Development of a scoring key for each test.

5. Application of the key to the test results of new samples of personnel.

6. Correlation of the scores on each test with the criterion of leadership performance.

In the remainder of this chapter, the tests will first be described, followed by a description, for the cadet officers, of the samples, criterion, procedure for collecting data, and methods of analyzing the data. Finally, the same rubrics of information will be presented for the enlisted personnel.

A. Tests (Copies of the tests are included in the Appendix of this report.

The following tests were included in the validation study:

1. <u>Picture Interpretation Test, 1949</u>. (<u>DA AGO PRT - 1775</u>)

This 432 item test consists of a series of 268 pictures, some of which present individuals participating in military activities and others involving individuals in civilian situations. The general directions indicate that the test is a measure of interests, although it may be considered a projective instrument to the extent that the examinee tends to identify with the situations and individuals illustrated in the pictures.

Instructions for the first six parts of the test follow the same general pattern. For the individuals or situations presented in the pictures in each part of the test, the examinee is required to choose between two alternative reactions, as follows:

(1) Part I

(a) "<u>Yes</u>, I would like to do what he is doing," or
(b) "<u>No</u>, I would not like to do what he is doing."

(2) Part II

(a) "<u>Yes</u>, I would like to be that person," or
(b) "<u>No</u>, I would not like to be that person."

(3) Part III

(a) "<u>Yes</u>, this person is like me," or
(b) "<u>No</u>, this person is not like me."

(4) Part IV

(a) "<u>Yes</u>, I would admire this person," or
(b) "<u>No</u>, I would not admire this person."

(5) Part V

(a) "<u>Yes</u>, I am good at doing what this person is doing," or
(b) "<u>No</u>, I am not good at doing what this person is doing."

(6) Part VI

(a) "<u>Yes</u>, I like what is shown in this picture," or
(b) "<u>No</u>, I do not like what is shown in this picture."

Part VII differs from the rest of the test in that pictures of military situations and civilian situations are presented along with five descriptive statements for each picture. The examinee is required to make the following choice in regard to each statement:

"Yes, the picture made me think of this idea," or
"No, the picture did not make me think of this idea."

2. Army Picture Story Test, Series B, 1950, Syracuse University Press.

The Army Picture Story Test is an objective test, based on the general idea of the Thematic Apperception Test, consisting of a series of ten pictures. The pictures included in the Army Picture Story Test involve both military and non-military situations and are not the pictures used in the Thematic Apperception Test. For each picture, there are thirty items presented in groups of three. The items are relatively short statements which are descriptive of the picture. The examinee is instructed to read the statements within each triad and to select two statements: the most descriptive and the least descriptive.

The statements used in this test were obtained by administering the set of ten pictures to a large group of soldiers in a free response situation. The descriptions written by this group were edited and arranged in triads on the basis of their frequency of occurrence and with respect to a number of clinical categories. That is, triads were composed of items which were approximately equal in frequency of occurrence but which dealt with different personality needs.

3. Picture Fill-In Test, Second Form, 1949 (DA AGO PRT-1726)

The Picture Fill-In Test is an adaptation of the Rosenzweig Picture-Frustration Test. It differs from the Rosenzweig test in that the responses are obtained in objective form. A series of 43 cartoon-like pictures is presented, comprising a total of 392 items. In each picture, one individual is represented as saying something to another individual. Some of the pictures deal with military situations, while 24 pictures were taken directly from the Rosenszweig test. In an experimental administration of the Preliminary Form of the Picture Fill-In Test, the examinees wrote responses in the cartoon balloons. Responses made most frequently by this experimental group were selected for each of the pictures. Certain responses which seemed to be particularly revealing or measuring important factors also were included, regardless of their frequency of occurrence. From seven to ten responses were selected, and are presented below each picture in the Second Form of the test. This form, which was used in the present investigation, was developed so that it would be suitable for objective scoring in the following manner: The instructions require that the examinee rate each of the responses presented with the pictures with respect to how likely it is that the person shown would give that response. This rating of each response is accomplished on the following three-point scale:

A. "_Might_ say something like this."
B. "Is _likely_ to say something like this."
C. "Is _very_ likely to say something like this."

4. _West Point Personal Inventory, 1949 (DA AGO PRT-1756)_ (Also referred to as _ROTC Self-Description Blank, Form II, 1949, DA AGO PRT-1744_, and in previous progress reports as _Biographical Information Blank, ROTC edition_.)

The West Point Personal Inventory used in the present investigation with cadets consists of four sections and a total of 420 items. This test does not make use of pictorial material. The items are in the form of statements concerning various characteristics, as follows: Section I includes pairs of statements dealing with personal characteristics; the individual is instructed to select the statement in each pair that is the best description of him. In Section II, the individual makes a choice between each of two activities as to which he believes he can do better. Statements dealing with likes and dislikes are presented in Section III, and the individual again selects the statement in each pair that he likes the better. Section IV contains statements describing personal characteristics, likes and dislikes, abilities, and beliefs. For each statement, the individual indicates whether the statement applies to him or does not apply.

5. _Leaders' Self-Description Blank, Form E, 1951, Syracuse University Press._

The Leaders' Self Description Blank is a 342-item version of the Biographical Information Blank and was used at the non-commissioned level at the Leaders' Schools in the present investigation. It is similar in composition to the West Point Personal Inventory, but the exact content of the items is different. Like the West Point Personal Inventory, it does not present pictorial material and consists of four sections.

Section I contains pairs of statements dealing with personal characteristics. The examinee chooses the statement from each pair that describes him better. The pairs of statements in Section II describe various activities, and the individual selects the activity which he can do better. Pairs of statements are presented in Section III dealing with likes and dislikes, and the instructions require selecting the statement that you like better. Personal characteristics, likes and dislikes, abilities, and beliefs make up the content of Section IV, and the examinee is instructed to indicate whether each statement applies to him or does not apply.

B. Situational Validity at the Commissioned Officer Level

1. Sample

The first and second classes of cadets at the U. S. Military Academy, West Point, in July, 1950, were the subjects of the investigation. A total of 454 cadets was tested.

Available cases from this sample were later divided into two random sub-groups for purposes of performing a double cross-validation analysis. These subgroups comprised, respectively, 213 and 223 cases.

2. Tests[1]

The tests employed with this sample were:

a. Picture Interpretation Test, 1949, (DA AGO PRT-1775)
b. Army Picture Story Test, Series B, 1950, Syracuse University Press
c. Picture Fill-In Test, Second Form, 1949, (DA AGO PRT-1726)
d. West Point Personal Inventory, 1949 (DA AGO PRT-1756)

3. Criterion

The Aptitude for the Service System[2] was ascertained for each cadet for use as a criterion measure of leadership. The Aptitude for the Service System is used at West Point for the purpose of providing an accurate evaluation of the leadership effectiveness of cadets. The Aptitude Rating is a composite measure including the pooled opinion of the cadet's Tactical Officer and a small group of classmates within his Company. The evaluation by his classmates is accomplished through an associate (buddy) rating procedure.

Each cadet is ranked in order of merit by his Tactical Officer and by the cadets in his Company in regard to the following definition of leadership:

"The criterion of my appraisal is each cadet's ability (if or when placed in command of a group) to elicit the group's maximum cooperation; maintain the highest possible standards of administration and

[1] A description of the tests used in the study is presented in Section III, A.

[2] A detailed description of the Aptitude for the Service System may be found in "The Operation and Administration of the Aptitude for the Service System, U.S.M.A.", West Point, New York: United States Military Academy, 1951.

discipline; and at the same time, develop and preserve high morale and group spirit."[1]

From the raw ratings, the median ranking for each cadet is determined and transposed to a standard score (called Army Standard Rating). The Tactical Officer's rating is assigned a weight of one-third in combining it with the associate ratings. It is this final or composite Army Standard Rating (Aptitude Rating) that constituted the criterion of leadership in this investigation.

4. Procedure

a. Test Administration

On June 30 and July 1, 1950, the four tests were administered in group situations to 216 cadets of the first and second classes at the U. S. Military Academy, West Point. The test battery was divided into two sessions, two of the tests being administered in the first session and the other two tests being given in the second session. Each session required about three hours of testing time. A similar procedure was utilized when the second group of West Point cadets was tested on July 27, 28 and 29, 1950. This second group of cadets numbered 238 and were from the first and second classes.

b. Collection of criterion data and constitution of criterion groups.

Criterion data, entered on Hollerith cards, were received from the West Point statistical office. These cards contained the cadet serial number, the mean Aptitude Rating based on the first term and the second term of the second class, Aptitude Ratings for both terms, and year of expected graduation. The criterion of leadership effectiveness utilized in this investigation with the West Point sample was the mean Aptitude Rating which summarizes the cadet's leadership performance during his second class.

The total sample was divided rendomly on the basis of serial numbers, group A being composed of those cadets with even serial numbers, and group B having odd serial numbers. As a check on the randomness of this procedure, *t* and F statistics were computed between the mean Aptitude Index criterion scores of the two groups; this analysis indicated that the two groups may be considered as random samples from the same population in regard to the leadership criterion. The purpose of fractionizing the sample in this manner was to make it possible to perform a double cross-validation on the scoring keys derived in the item analyses.

[1] Ibid, p. 8

5. Analysis of Data

a. Item Analysis

The validity of the items in the experimental tests was estimated by computing the biserial correlation coefficient between dichotomized item responses and the Aptitude Rating criterion. The computation of the item validities was facilitated by making use of the Kolbe and Edgerton table for estimating biserial correlation coefficients.[1]

The Aptitude Rating criterion was normalized by dividing the distribution into equal frequency eighths, and assigning the standard score equivalent of the mid-point of each eighth in a normal distribution to each criterion score within that eighth. Thus, all cases falling in a given eighth of the obtained distribution of criterion scores received the same standard score equivalent.

While the same general item analysis procedure was followed for the West Point study, somewhat different techniques of dichotomizing the item responses were necessary for the different tests, as follows:

(1) Picture Interpretation Test, 1949 (DA AGO PRT-1775) The item responses in this test fit a natural dichotomy since the examinee is instructed to indicate either "Yes" or "No" for each item. Thus, there is no problem in dichotomizing the responses for the purposes of the biserial correlation type of item analysis.

(2) Army Picture Story Test, Series B, 1950, Syracuse University Press. As described in Section III, A, Tests, the Army Picture Story Test requires that the individual choose the most descriptive and the least descriptive statements from groups of three items. Within the triad, the item that is considered to be most descriptive is marked A, while the item that seems to be least descriptive is marked B, and the intermediate item is not marked. For purposes of obtaining item frequencies, I.B.M. graphic item counts were made for each item, for the A or B alternatives. The trichotomous alternatives for each item were dichotomized in order to apply the biserial correlation item analysis technique; in doing this, the extreme alternative ("Best" or "Worst") having the larger frequency of response was used as one category of the dichotomy, while the combination of the other extreme with the intermediate alternative constituted the other category. This arrangement was used in order to yield the closest approximation to a 50% 50% dichotomy, thus maximizing the stability of the resulting item validity coefficients.

[1] Kolbe, L. E., and Edgerton, H. A., "A Table for Computing Biserial r", J. Exp. Educ., 1936, 4, 245 251.

(3) *Picture Fill-In Test, Second Form, 1949* (*DA AGO PRT-1726*). This test requires that the individual rate each item on a three-point scale in regard to the degree of likelihood that the item is an appropriate statement, as explained in Section III, B, Tests. The dichotomy required by biserial item analysis was achieved by combining the B and C responses. Thus, the frequencies of responses were obtained for the A category vs. the combined B and C category. The basis for grouping the B and C responses rather than utilizing some other combination in order to dichotomize the responses was both logical and empirical. On *a priori* grounds, it seemed more reasonable to believe that B (Is *likely* to say something like this) is closer on a continuum to C (Is *very* likely to say something like this). Moreover, an inspection of the item responses indicated that by using the dichotomy of A vs. B and C, the ideal 50%-50% dichotomy was more closely approximated.

(4) *West Point Personal Inventory, 1949* (*DA AGO PRT-1756*). An item analysis of this test was not necessary since the scoring key had already been developed in a previous study and was made available by the Personnel Research Section, AGO, for the validation phase of the present investigation.

b. Pattern Item Analysis

Since the items of the Army Picture Story Test are grouped in triads, it was hypothesized that the pattern of responses might be significant. In order to investigate this hypothesis, the following pattern analysis was performed: For each triad, six patterns of responses are possible. For Group A of the West Point sample, frequency counts were made of the responses to each of the six patterns for each of the 100 triads. A level of significance test based on X^2 was made among the frequencies of the patterns for each triad, contrasting upper and lower criterion groups. This procedure made it possible to estimate the validities of the pattern responses.

c. Cross-Validation

In general, the validities of the tests were estimated by computing Pearson product-moment correlation coefficients between the scoring keys derived and the Aptitude Rating leadership criterion.

In following this procedure, the two samples, group A (even serial numbers) and group B (odd serial numbers) were treated separately, in order that the scoring key derived on group A could be crossed over and validated on Group B, and the scoring key obtained on group B could be validated on group A. This double cross-validation technique makes use of the principle of replication in determining which items are consistently valid in both samples and permits two minimum estimates of the validity of the test.[1]

[1] For a fuller discussion of the double cross-validation technique, see Katzell, R. A., "Cross-Validation of Item Analyses", *Educ. Psychol. Measmt.*, 1951, *11*, 16-22.

C. Enlisted Personnel Validity Studies

1. Samples

a. Item Analysis Group

The enlisted personnel samples in this study were drawn from Army Leader's Schools whose mission is to train personnel for leadership, primarily at the non-commissioned officer level. It is anticipated that the bulk of trainees will consist of privates who have just completed basic training and who have been recommended by their company officers as evidencing leadership potential. This source of students is called "pipe-line". Other sources have included reenlistments and National Guard personnel called up for active duty as a result of the Korean conflict. For these groups, training at the Leader's Schools is considered a refresher course. One other important category includes officer candidates who, at present are required to complete a leadership course before attending Officer Candidate Schools. Leaders Schools at Ft. Dix, N. J., and Ft. Knox, Ky., which train soldiers from ground force units, were visited to gather the data for item analysis. 958 men were tested at these two installations.

The sample was divided into two subgroups: privates (including privates first class) and non-commissioned officers. These groups differ in average age and military background, factors which might affect performance on the tests and criteria. Hence, it was considered desirable to perform separate item analyses and validations for the two subsamples.

b. Cross-Validation Group

Leaders Schools at Ft. Jackson, S. C., and Ft. Belvoir, Va., were visited to secure the data for cross-validation. These schools train personnel from infantry and engineering units, respectively. 368 cases were utilized for the cross-validation results.

2. Tests

a. Item Analysis Group

At Ft. Dix and Ft. Knox, the Picture Interpretation Test, 1949 (DA AGO PRT-1775), Picture Fill-In Test, Second Form, 1949, (DA AGO PRT-1726), and Army Picture Story Test, Series B, 1950, Syracuse University Press were administered.

b. Cross-Validation Group

On the basis of the item analysis performed, it was decided to administer only the Picture Interpretation Test and the Picture Fill-In Test to the cross-validation sample. In addition, at the request of PRS, the Leader's Self-Description Blank, Series E, 1951, Syracuse University Press, was administered to this same group.

Descriptions of the nature of all the above tests may be found in Section III, Part A, of this report.

3. Criteria

a. Item Analysis Group

The training cycle at Leader's Schools is divided into two four-week phases, Phase I and Phase II. Training during Phase I is primarily academic in nature, whereas Phase II consists primarily of practical leadership experience in field situations. During the training program, soldiers are periodically evaluated for their performance on different criteria by various kinds of raters, i.e., both commissioned and non-commissioned cadre as well as by their peers. All soldiers tested in this study were in Phase I of the training cycle at the time of testing.

The following criterion measures of leadership were obtained for the group tested at Ft. Dix and Ft. Knox: Faculty Board Rating, Associate Rating, Leaders' Reaction Test, Rating of Phase II Performance, and Total Rating (a weighted combination of the foregoing).

Intercorrelations among the above criteria were computed for a sample of the soldiers tested at Ft. Dix and Ft. Knox. These statistics are useful for estimating the extent to which the criteria measure different aspects of leadership. The following table shows these results. (See Table 1).

In both samples, it seems evident that the various criteria are somewhat unrelated to each other. Although the Associate Rating also appears to be somewhat different from other ratings of leadership potential, on the basis of its recommendation by Personnel Research Section for use in this study, it would seem to be the most appropriate measure of leadership available. Results from other Personnel Research Section studies[1] had shown associate ratings to be superior measures of leadership. In the present study, furthermore, Associate Ratings were available for more subjects tested than any of the other ratings.

[1] Wherry, Robert H. and Fryer, Douglas H., "Buddy Ratings: Popularity Contest or Leadership Criteria?", Personnel Psychology, 1949, 2, 147-159.

Table 1. Intercorrelations among Various Criteria for Two Enlisted Samples.

A. Fort Dix

N = 173

	F.B.R.	L.R.T.	A.R.	Perf. II	Total
Faculty Board Rating	---	.14	.45	.15	.74
Leaders Reaction Test		---	.04	.02	.37
Associate Rating			---	.06	.44
Performance during Phase II				---	.66
Total					---

Correlation between Phase II and sum of other three variables = .20

B. Fort Knox

N = 112-162

	F.B.R.	L.R.T.	A.R.	Perf. II	Total
Faculty Board Rating	---	.32	.15	.33	.78
Leaders Reaction Test		---	.00	.38	.50
Associate Rating			---	-.12	.20
Performance during Phase II				---	.78
Total					---

Correlation between Phase II and sum of other three variables = .24

All of the above considerations led to the decision to use Associate Ratings as the criterion for the enlisted sample of this study.

Since this criterion was adopted, it will be valuable to describe in somewhat greater detail the operations by which scores on this measure were obtained for the samples. Each student is evaluated by his fellow students at the end of Phase I training. The students are each given a "Student Leadership Evaluation Report-Rating Sheet". On this sheet is a roster of the men in the student's group, customarily numbering from nine to fifteen men. The student, from this roster, chooses those whom he thinks the three best leaders and the three poorest leaders. On the next day, each student is given "Student Leadership Evaluation Report-Description Sheet". On this sheet are printed the names of the men in the group. There are also ten pairs of descriptive statements. For each man on the roster, the student is to choose the description in each pair of statements which most appropriately describes the man being rated. These sheets are then scored by using the keys furnished by Personnel Research Section. One score is based on the nominating technique, weights being given to the number of nominations received. The more nominations a soldier receives which are indicative of better leadership, the higher his Associate Rating score. The other score is derived from weights based on empirical evidence as to which descriptions are more characteristic of better leaders. The scores from the rating sheets and description sheets are averaged. These scores constituted the Associate Rating criterion used for item analysis purposes.

b. Cross-Validation Group

Associate Ratings were also obtained for the soldiers tested at Ft. Jackson and Ft. Belvoir. The men were rated in the period from about January through March, 1952. Test scores were obtained by use of the keys developed from the item analysis group. These scores were correlated with the Associate Ratings.

4. Procedure

a. Item Analysis Group

Trainees in Phase I at Ft. Dix were tested in August and October of 1950 and January 1951. The Picture Interpretation, Picture Fill-In, and Army Picture Story Tests were administered to groups ranging in size from approximately 50 to 100 trainees. Every effort was made to elicit the cooperation of the soldiers tested, including some explanation of the purpose of the study. A total of 480 subjects in Phase I was tested at Ft. Dix. In field trips made during October 1950 and January 1951, 478 subjects in Phase I were tested at Ft. Knox. Thus, the total number of trainees tested for the item analysis group was 958.

b. Cross-Validation Group

Trainees in Phase I at Ft. Jackson, S. C., and Ft. Belvoir, Va., were tested in January, 1952, under conditions similar to those obtaining for the item analysis group. On the basis of the results of the item analysis, it was decided to administer only the Picture Interpretation and Picture Fill-In tests to these groups. An additional test was administered at the request of Personnel Research Section, the Leader's Self-Description Blank. At Ft. Jackson, the number of privates whose test papers were adequately filled out was 166; non-commissioned officers numbered 50. At Ft. Belvoir, test papers from 143 privates were acceptable; the non-commissioned officer sample numbered 9. The total number for all three tests included 309 privates and 59 non-commissioned officers.

Table 2 summarizes the number of cases used.

Table 2. Number of Cases in Item-Analysis and Cross-Validation Samples.

A. Number of Cases in the Item Analysis Group

	Ft. Dix	Ft. Knox	Total
Privates	246	139	385
Non-coms	90	138	228

B. Number of Cases in the Cross-Validation Group

	Ft. Jackson	Ft. Belvoir	Total
Privates	166	143	309
Non-coms	50	9	59

5. Analysis

a. Item Analysis Group

For the purpose of item analyzing the three tests used, it was desired to combine the installation samples, since the resulting keys would be used irrespective of installation. The following analyses were performed in order to determine the most appropriate statistical method for combining the Associate Rating scores from the two installations.

Critical ratios were computed comparing mean Associate Rating scores obtained by soldiers at Ft. Dix with those at Ft. Knox. The differences between installations were significant at the 1% level of confidence.

Variance ratios were computed for these data to test the significance of differences in variability for Associate Rating scores. Differences in variance were significant at the 2% level of confidence for non-coms at Knox versus non-coms at Dix. This significant difference in variability makes ambiguous the interpretation of tests of significance for mean differences reported, since significant critical ratios between means may arise because of differences in variability.

The following tables present data from which the above interpretations were made.

Table 3. Critical Ratios and Variance Ratios for Testing Differences in Mean Associate Rating Scores

A. Privates

	Fort Dix		Fort Knox
N	247		139
Mean	79.16		72.44
Standard Deviation	4.79		4.02
Critical Ratio		14.6^{***}	
Variance Ratio		1.42^{*}	

B. Non-commissioned Officers

	Fort Dix		Fort Knox
N	90		137
Mean	80.44		72.30
Standard Deviation	3.94		5.12
Critical Ratio		13.6^{***}	
Variance Ratio		1.68^{**}	

*** Significant at the 1% level of confidence
** Significant at the 2% level of confidence
* Significant at the 10% level of confidence

On the basis of the preceding analyses, the best procedure for combining the two installations seemed to be conversion of Associate Rating raw scores to standard scores within each installation before pooling the two. Although about 950 men had been tested on each of the three experimental tests administered, attrition in the number of cases had occurred as a result of improperly answered tests and also by the inability to secure criterion measures on some of the subjects. The graphic item counts are based on a sample of 385 privates and 228 non-commissioned officers. From the graphic item counts, biserial r's were computed for each item of each of the three tests administered. The method by which these were computed was analagous to that used for the cadet officer sample.

The Associate Rating criterion was normalized by dividing the distribution into equal frequency eighths, and assigning the standard score equivalent of the midpoint of each eighth in a normal distribution to the criterion scores within that eighth.

Graphic item counts for the three projective tests were obtained separately for the samples of privates and non-commissioned officers. Each of these two samples had been fractionized into eight subsamples of equal frequency, after first arranging the cases in descending order according to their Associate Rating standard scores.

While the same general item analysis procedure was followed for the three tests, somewhat different techniques for dichotomizing the item responses were necessary for the different tests. For the method of dichotomizing responses used for the different tests, see Section B, part 5, under cadet officers.

Scoring keys were developed for those tests with promising validity based on the item analysis results. Items were selected whose biserial r's were significant at the 10% level of confidence. The value of biserial r necessary for significance was determined from the standard error of biserial r computed from the following formula[1]:

$$SE_{r_{bis}} = \frac{\frac{\sqrt{pq}}{z} - r^2_{bis}}{\sqrt{N}}$$

Significant biserial r's are a function of the percentage of cases in each dichotomized group, as well as the confidence level adopted. For a 50% dichotomous split, an r of .105 was required for significance at the 10% level of confidence in the private's sample. For the non-commissioned sample, an r of .137 was required for significance at this level of confidence.

One key was developed for the Picture Interpretation Test in addition to those obtained as above. This test was selected for special study in an effort to discover the nature of those items which yielded significant biserial r's. Two judges classified the significant items from this test into 13 categories suggested by the kinds of pictures which produced significant responses. Those non-significant items whose content fitted the classification scheme adopted were also placed into these categories. It was reasoned that if the classifications used were indicative of real relationships between item content and criterion, non-significant items in the same classification might show correlations with the criterion having the same direction as that found for the statistically significant items classified in the same category.

To check on the extent to which non-significant items were predicted with correct signs for the various categories, the following analysis was performed. The proportions of positive and negative biserial r's among all non-significant items classified were determined. Likewise the proportion of positive or negative items allocated to each category was determined. The differences between proportions in each category and in the total were then tested for significance. If these differences were significant, it was concluded that the classification of items within these categories was meaningful. By this procedure, an item classification key of 89 items was developed. The item-analysis key for this test contained 99 items.

Certain trainees had been eliminated from the item analysis

[1] Kelley, T.L., Fundamentals of Statistics. Cambridge: Harvard, 1947, p. 375.

because they lacked Associate Rating scores as a result of having been dropped, for various reasons, from the Leader's Course. There were 25 such cases who had complete sets of tests. Mean scores were obtained for this group on the Picture Interpretation Test and the Picture Fill-In Test by scoring their papers with the item analysis keys developed as described previously.

It was desired to compare the mean score for dropouts to the mean score of the total item analysis group. It was postulated that, if a positive correlation existed between criterion and test scores, dropout mean scores on the tests would be significantly lower than the means of the item analysis group. This assumes that dropouts, had they been rated, would have received relatively low Associate Rating scores.

In order to estimate the mean test scores of the item analysis group, it was necessary to score their papers with the item analysis keys. Rather than scoring test papers for the entire sample of 385 privates, 50 cases were selected at random from this group. A stratified random sampling technique was used since the 385 privates had been fractionated into eighths on the basis of Associate Ratings for item analysis purposes (see page 9). Furthermore, for both this group and the dropouts, a different keying of responses was used to simplify scoring than that used later for the cross-validation sample. Since the keying of items is arbitrary, the two keying methods used result in scores which differ only by a constant.

For testing the significance of the differences between means of the dropout and item analysis group, *t* tests were computed. The standard error for the difference between means was adjusted in the *t* formula to take account of the use of a stratified random sample.[1]

b. Cross-Validation Group

Using the item-analysis keys, the Picture Fill-In and Picture Interpretation Tests were scored for the cross-validation sample. The Picture Interpretation Test was also scored for the item classification key described above. The Leader's Self-Description Blank was scored by using the key furnished by the Personnel Research Section.

Reliability coefficients were computed for the Picture Interpretation and Picture Fill-In tests. The method of computation used was the correlation between scores from odd-and-even numbered items, augmented by the Spearman-Brown prophecy formula to estimate the reliability of the whole test.

[1] McNemar, Q., *Psychological Statistics*, John Wiley and Sons, Inc.: New York, 1949, pp. 333-336.

Validity coefficients were computed for the cross-validation sample. Scores on each of the three tests, obtained as described above, were correlated with the Associate Rating criterion. The correlations were computed separately for the Ft. Jackson and Ft. Belvoir samples, and for the two combined.

Before combining the two installations into a total sample, it was advisable to test whether mean Associate Rating scores for the two installations differed significantly. A critical ratio was computed testing the significance of this difference. If this ratio were not statistically significant, Associate Rating scores would not be converted to standard scores for computation of the validity coefficients.

To test whether the validity coefficients differed significantly for the two installations, critical ratios were computed for the difference between two sample correlation coefficients. An r to z transformation was made prior to this statistical test.

Two multiple correlation coefficients were computed, by the Wherry-Doolittle method, with Associate Ratings as the criterion variable in both cases, and as the predictor variables (1) Picture Interpretation Test and Picture Fill-In Test, and (2) Picture Interpretation Test, Picture Fill-In Test, and Leaders' Self-Description Blank. Only those trainees who had completed all three tests and had received an Associate Rating were utilized for this analysis.

IV. RESULTS

A. Results with Cadet Officers

1. Picture Interpretation Test, 1949 (DA AGO PRT-1775)

a. Item Analysis[1]

The proportion of items with statistically significant validities failed to exceed chance expectancy, indicating a lack of validity for the test with this sample. A scatterplot was prepared showing the relationship of the obtained validity coefficients of Group A (N = 211) vs. the corresponding coefficients of Group B (N = 223). The correlation between the two sets of validity coefficients was approximately zero, indicating little consistency of item validity from sub-sample to sub-sample. This finding, together with the low proportion of significantly valid items, strongly suggests that the test does not possess sufficient validity for the prediction of leadership with West Point cadets.

A qualitative investigation of those items for which combined validities exceeded chance expectancy did not yield logical categories or trends which were considered to be psychologically meaningful.

Table 4 shows the distribution of the item validities in the Picture Interpretation Test.

b. Cross-Validation

To substantiate the evidence from the item analysis, the validity of the Picture Interpretation Test was estimated by computing the correlation coefficient between the scoring keys derived on the item analysis samples and the Aptitude Rating leadership criterion. For Group A (N = 210) the scoring key yielded a correlation coefficient of .12. For Group B (N = 222) the correlation coefficient was found to be .07, indicating the lack of appreciable validity of this test for these samples.

[1] Item validities have been reported in detail for each of the tests in tables included in regular monthly progress reports submitted to the Department of the Army during the course of the study. Slightly different N's from sample to sample and from test to test are the result of incomplete data on a few cases in the sample tested.

Table 4. Distribution of Item Validities in the Picture Interpretation Test for Two West Point Samples.

Group A		Group B	
N = 211		N = 223	
r	f	r	f
.00-.04	172	.00-.04	152
.05-.09	100	.05-.09	140
.10-.14	87	.10-.14	166
.15-.19	38	.15-.19	45
.20-.24	16	.20-.24	15
.25-.29	11	.25-.29	11
.30-.34	4	.30-.34	2
.35-.39	1	.35-.39	0
Total	429 items*	.40-.44	0
		.45-.49	1
		Total	432 items

* The total number of items for which it was possible to compute validity coefficients was 429 in Group A, since three items (no. 184, 344, and 362) yielded no responses in one category of the dichotomy.

2. Army Picture Story Test, Series B, 1950, Syracuse University Press.

a. Item Analysis

The item analysis of the Army Picture Story Test revealed only a chance proportion of statistically significant items. The scatterplot between the validity coefficients in Group A (N = 213) and Group B (N = 222) for the Army Picture Story Test indicated a near zero relationship, suggesting little consistency of item validity. However, as in the case of the Picture Interpretation Test, the subjects in the two samples, Group A and Group B, did respond similarly to individual items indicating a marked degree of inter-sample consistency.

A qualitative analysis of the significant items of the Army Picture Story Test also failed to disclose meaningful categories or trends in terms of postulated leadership characteristics.

Table 5 shows the distribution of item validities in the Army Picture Story Test.

b. Pattern Item Analysis[1]

Of the 600 possible patterns of response in the Army Picture Story Test, 30% of the patterns showed significantly high criterion relationships at the 20% level of confidence in Group A (N = 213). The percentage of significant patterns may not be beyond chance expectations because of inter-pattern correlation. However, the degree to which the relationships among patterns affect the number of patterns appearing to possess significant validities is impossible to determine. Thus, a scoring key was constructed on the basis of the significant patterns by assigning a weight of +1 to patterns with positive validity at the 20% level of confidence, and -1 to patterns with negative validity at this level.

c. Cross Validation

The validity of the item analysis keys of the Army Picture Story Test was estimated by calculating the Pearson product-moment correlation coefficient between test scores and the Aptitude Rating leadership criterion. Group A (N = 210) and Group B (N = 222) yielded validity coefficients of -.08 and -.05 respectively, indicating essentially zero validity for the test with these samples.

[1] Validities of each pattern have been reported in regular progress reports submitted to the Department of the Army during the course of the study.

Table 5. Distribution of Item Validities in the Army Picture Story Test for Two West Point Samples.

Group A		Group B	
N = 213		N = 222	
r	f	r	f
.00-.04	144	.00-.04	156
.05-.09	79	.05-.09	69
.10-.14	49	.10-.14	45
.15-.19	18	.15-.19	20
.20-.24	7	.20-.24	6
.25-.29	1	.25-.29	1
.30-.34	1	.30-.34	1
.35-.39	1	.35-.39	0
Total	300 items	.40-.44	0
		.45-.49	1
		.50-.54	0
		.55-.59	0
		.60-.64	1
		Total	300 items

Using the key derived by the pattern analysis on Group A, the scores of Group B (N = 222) were calculated. The Pearson product-moment correlation coefficient between the test scores and the Aptitude Rating leadership criterion was .01, indicating the lack of validity in the pattern key for this sample.

3. Picture Fill-In Test, Second Form, 1949 (DA AGO PRT-1726)

a. Item Analysis

The proportion of statistically significant items failed to exceed chance expectancy. The scatterplot between the correlation coefficients for Group A (N = 213) and Group B (N = 223) indicated approximately a zero relationship for the Picture Fill-In Test. Thus, again negative evidence was found in regard to the inter-sample consistency of item validity. As in the case of the two tests mentioned previously, a qualitative analysis of the significant items found in the two samples failed to reveal categories of responses which seemed to be psychologically meaningful.

However, there was evidence of considerable consistency between samples in the proportion of individuals who responded in the same way to the items in the test. This same indication of the inter-sample consistency of the responses was also found for the two tests discussed previously: The Picture Interpretation Test and the Army Pictrre Story Test.

The following table presents the distribution of item validities: (See Table 6).

b. Cross-Validation

Estimates of the validity of the Picture Fill-In Test were obtained by computing the Pearson product-moment correlation coefficients between the item analysis keys and the Aptitude Rating criterion of leadership. The validity of the test for the West Point samples was found to be approximately zero; in Group A (N = 210) the validity coefficient was .04 and in Group B (N = 222) the validity coefficient was .03.

4. West Point Personal Inventory, 1949 (DA AGO PRT-1758)

a. Item Analysis

As explained in Section III, B, 5, a, (4), the scoring key for this test was provided by the Personnel Research Section, AGO.

b. Cross-Validation

The West Point Personal Inventory was the only one of the four experimental tests which manifested appreciable validity for

Table 6. Distribution of Item Validities in the Picture Fill-In Test for Two West Point Samples.

Group A		Group B	
N = 213		N = 223	
r	f	r	f
.00-.04	164	.00-.04	161
.05-.09	106	.05-.09	119
.10-.14	69	.10-.14	57
.15-.19	32	.15-.19	29
.20-.24	12	.20-.24	20
.25-.29	4	.25-.29	2
.30-.34	3	.30-.34	2
.35-.39	1	.35-.39	0
.40-.44	0	.40-.44	1
.45-.49	0	.45-.49	1
.50-.54	0	Total	392 items
.55-.59	0		
.60-.64	0		
.65-.69	1		
Total	392 items		

the West Point sample. The Pearson product-moment correlation coefficient between the scores on this test and the Aptitude Rating leadership criterion was .351, based on a sample of 426 cadets. This value of .351 is a statistically significant validity coefficient, since a value of .128 is required for significance at the 1% level of confidence.

B. Results with Enlisted Personnel

1. Item-Analysis Group: All Tests

a. Item Analyses

Tables 8, 9, and 10 show the distribution of biserial r's obtained for items in each test. The sign of the coefficients was disregarded in tabulating these results, since keying the responses to each item is only arbitrary.

Table 7 summarizes the number of significant items found for the two samples at the 10% level of confidence in each of the three tests.

Table 7. Number of Significant Items for Three Experimental Personality Tests.

	Total No. of Items	Number of Significant Items
Picture Interpretation Test		
1. Privates (N = 385)	432	99
2. Non-Coms (N = 228)	432	53
Picture Fill-In Test		
1. Privates	392	130
2. Non-Coms	392	38
Picture Story Test		
1. Privates	300	48
2. Non-Coms	300	45

From this table, it is apparent that more significant items are found for the sample of privates than is true for non commissioned officers. By inspecting the individual items, it is also apparent that those items found significant in the private's sample generally are not found significant in the non-commissioned officer sample. The results indicate, furthermore, that the Picture Fill-In Test and Picture Interpretation Test are functioning in the sample of privates at a level appreciably better than chance expectancy. Since the 10% level of confidence was adopted, chance, on the average, would result in approximately 39 significant items for the Picture Fill-In Test, 42 for the Picture Interpretation Test, and 30 for the Army Picture Story Test, assuming no correlation among the i[illegible] each test.

Table 8. Distribution of Item Validities in the Picture Interpretation Test for Two Enlisted Samples

A. Privates

N = 385

r	f
.00-.04	152
.05-.09	131
.10-.14	90
.15-.19	44
.20-.24	13
.25-.29	1
.30-.34	1
Total	432 items

B. Non-Commissioned Officers

N = 228

r	f
.00-.04	168
.05-.09	107
.10-.14	88
.15-.19	45
.20-.24	15
.25-.29	2
.30-.34	2
.35-.39	1
.40-.44	2
.45-.49	0
.50-.54	1
.55-.59	0
.60-.64	0
.65-.69	0
.70-.74	1
Total	432 items

Table 9. Distribution of Item Validities in the Army Picture Story Test for Two Enlisted Samples.

A. Privates N = 385		B. Non-Commissioned Officers N = 228	
r	f	r	f
.00-.04	132	.00-.04	112
.05-.09	105	.05-.09	91
.10-.14	51	.10-.14	59
.15-.19	11	.15-.19	29
.20-.24	1	.20-.24	7
Total	300 items	.25-.29	2
		Total	300 items

Table 10. Distribution of Item Validities in the Picture Fill-In Test for Two Enlisted Samples.

A. Privates N = 385		B. Non-Commissioned Officers N = 228	
r	f	r	f
.00-.04	140	.00-.04	176
.05-.09	99	.05-.09	109
.10-.14	71	.10-.14	72
.15-.19	42	.15-.19	27
.20-.24	30	.20-.24	5
.25-.29	9	.25-.29	2
.30-.34	1	.30-.34	1
Total	392 items	Total	392 items

For the Army Picture Story Test, the number of significant items was about 1.6 times what would be expected on the basis of chance. This figure is not large enough to warrant concluding that non-chance relationships are involved, particularly since the assumption of non-correlation among items is untenable. It is possible that a pattern analysis might reveal more convincing evidence for the validity of this test, in view of the triad form of item responses. Experience with the pattern analysis performed for the cadet officer sample did not encourage a parallel analysis for the enlisted sample. In view of the relatively low number of significant items found for this test, it was not administered to the cross-validation sample.

For the Picture Interpretation Test, about 2.3 times as many items were found significant for the sample of privates as would be expected on the basis of chance. For the Picture Fill-In Test, this figure was about 3.3. These tests were selected to be administered to the cross-validation sample, since the evidence is suggestive of validity.

b. Comparison of Item Analysis Group and Dropouts

Table 11 shows *t* tests for the significance of the mean differences between dropouts and the item analysis group. It had been expected that dropouts would show lower mean test scores for the Picture Fill-In Test and for the Picture Interpretation Test. Such is the case, and furthermore, this difference is significant at the 5% level of confidence for the Picture Interpretation Test, and at the 1% level of confidence for the Picture Fill-In Test.

2. Cross-Validation

a. Reliabilities and Related Statistics

Table 12 summarizes reliabilities and related statistics calculated from the cross-validation sample for the Picture Fill-In Test and the Picture Interpretation Test.

These results indicate that scores for these instruments are sufficiently reliable to be useful for large-scale classification purposes.

b. Validities

(1) Item Analysis Keys

Table 13 summarizes validity coefficients for the keys developed from item analysis of the Picture Fill-In and Picture Interpretation Tests.

Table 11. Significance of Mean Differences between Dropouts and Item Analysis Group for Picture Fill-In and Picture Interpretation Tests

	Picture Fill-In	Picture Interpretation
Mean of Dropout Group	68.8	50.4
S. D. of Dropout Group	21.1	9.6
N of Dropout Group	25	25
Mean of Item Analysis Group	90.6	56.2
S. D. of Item Analysis Group	21.6	11.8
N of Item Analysis Group	50	50
t Ratio between groups	4.26**	2.44*

* Significant at 5% level
** Significant at 1% level

Table 12. Reliabilities and Related Statistics Estimated from Cross-Validation Samples at Leaders' Schools for the Picture Interpretation and Picture Fill-In Tests

Picture Interpretation Test

	N	Mean	S	S^2	S.E. of Meas.	Odd-Even Rel.	Spearman Brown
Ft. Jackson	158	61.34	10.31	106.25	4.50	.81	.90
Ft. Belvoir	138	60.51	9.58	91.80	5.58	.66	.80
Total Sample	296	60.95	9.97	99.47	5.08	.74	.85

Picture Fill-In Test

	N	Mean	S	S^2	S.E. of Meas.	Odd-Even Rel.	Spearman Brown
Ft. Jackson	145	97.39	12.53	157.08	5.46	.81	.90
Ft. Belvoir	133	91.91	18.61	346.25	5.58	.91	.95
Total Sample	278	94.77	15.85	251.28	6.34	.84	.91

Table 13. Validity Coefficients for the Picture Fill-In and Picture Interpretation Tests in the Privates' Sample.

	Ft. Jackson	Ft. Belvoir	Combined
Picture Fill-In Test			
r	.19*	.24**	.19**
N	124	132	256
M	97.0	91.3	94.1
S. D.	13.2	18.8	16.7
Picture Interpretation Test			
r	.36**	.16	.25**
N	133	136	269
M	61.3	60.4	60.8
S. D.	9.9	9.7	9.8

* Significant at the 5% level of confidence.

** Significant at the 1% level of confidence.

All correlations reported are significantly different from zero at the 5% level of confidence with the exception of r = .16 for the Ft. Belvoir sample on the Picture Interpretation Test. This correlation approaches significance at the 5% level of confidence so closely that it, too, is unlikely to be considered a sampling fluctuation from a correlation of zero.

Differences in mean criterion scores for the two installations were not statistically significant at the 5% level of confidence. On the basis of this result, validity coefficients were computed using Associate Ratings as raw scores rather than converting them to standard scores.

Tests of the significance of the difference between sample correlation coefficients were performed in order to compare validity coefficients at the two installations. These results fail to reveal a difference, significant at the 5% level of confidence, between the validity coefficients at the two installations. The tests seem to be about equally valid in both of the cross-validation samples.

(2) Item Classification Key for the Picture Interpretation Test.

Table 14 presents validity coefficients for the item classification key, for the item analysis key, for the combination of the two, and the correlation between the item analysis and item classification keys. The criterion used for the validity coefficients was that of Associate Ratings.

The fact that the item classification key correlates significantly with the item analysis key is interpreted to mean that these categories have an appreciable degree of internal consistency. The failure of the items classified to correlate significantly with the criterion is an indication of their consistent lack of validity for this criterion even in a cross-validation sample.

(3) Multiple Correlation

Table 15 shows the multiple correlations for the total cross-validation sample, and the intercorrelations from which the R's were computed.

The standard errors of these R's are such as to render no combination of the tests appreciably superior in prediction to another, nor, indeed to the Leaders Self-Description Blank alone.

Table 14. Validity Coefficients of the Item Analysis and Item Classification Keys for the Picture Interpretation Test and the Correlation between Keys.

Privates

	Ft. Jackson	Ft. Belvoir	Combined
Item Classification Key Alone	-.01	-.18	-.06
N	134	136	270
Item Classification Key and Item Analysis Key	.25	.06	.16
N	133	135	267
Item Analysis Key Along	.36	.16	.25
N	133	136	269
Item Analysis Key vs. Item Classification Key	.54	.26	.40
N	158	138	296

Table 15. Intercorrelations among Predictor Variables, Correlations with Associate Rating Criterion, and Multiple Correlations for Cross-Validation Sample

(N = 234 cases in cross-validation sample with all 4 measures)

	Assoc. Rating	P.F.I.	P.I.T.	L.S.D.B.
Associate Rating (0)	....			
Picture Fill-In (1)	.17	....		
Picture Interpretation (2)	.23	.18	....	
Leaders' Self Description Blank (3)	.30	.34	.37	....

Multiple Correlations

$R_{0.12} = .27$

$R_{0.123} = .32$

V. CONCLUSIONS

The conclusions to be derived from the results described in the preceding section will be presented below in the sequence corresponding to the objectives set forth in Section II.

A. Validity of the three objective projective tests for measuring leadership performance of West Point cadets.

1. In none of the three tests does the aggregation of items correlate with the leadership criterion (Aptitude Rating) appreciably beyond chance expectation. This follows from the fact that the number of items found statistically significant at a given level of confidence is no greater than the number which would manifest that degree of validity through sampling fluctuations about a true validity of zero.

2. When scoring keys are developed independently for each of two representative samples by keying items whose individual validity coefficients appear to be statistically significant, there is practically no better than chance correspondence in the items comprised within the two keys. Hence, it can be inferred that there is inadequate consistency of the scoring keys from sample to sample. Furthermore, each of the two keys for each test correlates approximately zero with the leadership criterion in its cross-validation sample.

3. Duplicating previous findings of the Personnel Research Section, the West Point Personal Inventory is found to correlate appreciably and significantly with the leadership criterion. In addition to demonstrating the validity of the test, this indicates that the criterion is predictable.

Discussion: This investigation failed to reveal a stable and valid method of keying the responses to three objective projective tests so as to predict leadership among West Point cadets with better-than-chance efficiency. This failure cannot be attributed totally to inadequacy of the leadership criterion, the Aptitude Rating, for it is predictable from scores on the West Point Personal Inventory.

B. Validity of three objective projective tests for measuring leadership performance of Army non-commissioned personnel.

1. a. Among privates enrolled at Leaders' Schools, the Army Picture Story Test failed to yield a proportion of items which correlates with the Associate Rating appreciably beyond chance expectation. However, in both the Picture Interpretation Test and the Picture Fill-In Test, considerably more than 10 per cent of the items were valid at least at the 10% level of confidence. Hence it was possible to develop a scoring key for each of these two tests with some expectation that the resulting scores would correlate appreciably with the leadership criterion in new samples.

b. Among non-commissioned officers enrolled in Leaders' Schools, none of the three tests yields an aggregation of items which correlates with the leadership criterion (Associate Rating) appreciably beyond chance expectation. This follows from the fact that the number of items found to be statistically significant at a given level of confidence is no greater than the number which would manifest that degree of validity through chance fluctuations about a true validity of zero.

2. a. When these keys for the Picture Interpretation and Picture Fill-In Tests are applied in new samples of privates enrolled in Leaders' Schools, the Spearman-Brown reliabilities of the scores are .85 and .91, respectively.

The correlation of these scores with the Associate Rating criterion yields a validity coefficient of .25 for the Picture Interpretation Test and .19 for the Picture Fill-In Test: these validity coefficients are significant at beyond the 1% level of confidence.

These keys also discriminate, on the average, between trainees who, for various reasons (including lack of leadership potential), are separated early in the program and those who are graduated.

b. (No cross-validation was undertaken with non-commissioned officers, in view of the negative results of the item analysis.)

3. a. The Leaders' Self-Description Blank is found to correlate appreciably and significantly with the Associate Rating criterion for privates. The validity coefficients of this and the other two tests do not differ from one another at the 5% level of confidence.

b. The multiple correlation between the criterion and the two projective tests is not appreciably higher than the validity coefficient of the Picture Interpretation Test alone. The multiple correlation between the criterion and the two projective tests plus the Leaders' Self-Description Blank is not appreciably higher than the validity coefficient of the last-named test alone.

c. The Picture Interpretation and Picture Fill-In Tests are virtually uncorrelated with one another. This, in the light of their relatively high reliabilities, suggests considerable independence of the factors measured. Both of the tests correlate somewhat more highly with the Leaders' Self-Description Blank, although still at a level far below their reliability coefficients.

Discussion: The Army Picture Story Test revealed evidence of validity in neither the sample of non-commissioned officers nor of privates. The Picture Interpretation Test and the Picture Fill-In Test evidenced no validity in the non commissioned officer sample but both showed appreciable and significant validity in the sample of privates.

It should be recognized, in this connection, that the Associate Rating criterion does not necessarily represent a total appraisal of leadership performance. This *a priori* consideration is substantiated by the fact that this particular criterion was only slightly correlated with other assessments of leadership obtained for those personnel in the Leaders' School. It is, therefore, conceivable that the validity of these tests for a more comprehensive criterion would be considerably higher.

C. Inferences concerning the general promise of the objective projective type of test.

The fact that these tests correlate significantly and appreciably with a limited criterion of leadership performance among privates, and their relative independence of one another and the Leaders' Self-Description Blank, combine to suggest that we have here a type of instrument that may constitute an important contribution to the techniques for identifying potential leaders. However, in the light of the results, this statement must be made with certain limitations. In the first place, the failure of the tests to function with samples of cadets and non-commissioned officers indicates that they may be most effective with personnel of relatively little sophistication and limited experience. Secondly, it should be recalled that the predictive value of the tests has been shown only for the Associate Rating criterion; this criterion probably does not reflect all aspects of leadership performance.

Conversely, the positive results that have been obtained with these measures by no means define the limits of their value. They may, for example, correlate as well or better with other aspects of Army leadership performance, or with similar criteria in other types of leadership situations, such as under field conditions. Their validity for other types of performance involving personality factors, *e.g.*, adaptability to stress situations, is likewise in the realm of possibility. At present, these are, of course, moot points. But the promise evidenced by these tests for the limited criterion and situations involved in this study imply the advisability of exploring their value for other criteria and situations.

Such further work should, however, be done with revisions of the present tests. Many items which failed to show validity in this study could well be eliminated, and replaced by other items constructed along lines which the present study suggests are more fruitful. Suggestions for such future modifications are incorporated in the following section.

VI. IMPLICATIONS AND RECOMMENDATIONS

The major implication of the present study is that there are at least two of these objective projective tests (the *Picture Interpretation* and *Picture Fill-In Tests*) which evidence validity for leadership criteria. Yet, the results indicate that there is considerably more work to be done before these instruments will have the utility which practical considerations require.

For one thing, the evidence is that the tests, as presently constituted, are valid only within a limited segment of personnel in connection with whom leadership assessments would be made, namely, the youngest and least experienced group. It behooves us to consider why this should be the case and what may be done to augment the range of utility of these instruments.

For another thing, the evidence does not suggest that these tests should be used to supplant or supplement the Leaders' Self Description Blank, which is currently available for assessing leadership potential of enlisted personnel. The reason for this statement is that neither of the two new tests correlates better with the criterion than the *Leaders' Self-Description Blank*, nor is the multiple correlation of the three tests appreciably higher than the validity of the *Leaders' Self-Description Blank*. However, it should be noted that the failure of the new tests to add appreciably to the validity of the *Leaders' Self-Description Blank* is not a function mainly of the degree of overlap among the measures; rather, the failure is primarily attributable to the relatively low (even though significant) validity manifested by the new tests. It would thus seem that revision of these tests so as to augment their predictive value could well result in useful additions to current selection instruments.

There are therefore two dimensions along which revision should proceed: (1) broadening of the range of personnel to which the tests are applicable, and (2) intensifying the discrimination power of the items within this range. Conceivably, of course, the two objectives may not be attainable with a single form of each test, so that a different form may be required for each of several types of personnel, *e.g.*, commissioned and enlisted.

Examination of the characteristics of the various items serves as a source of hypotheses for effecting these improvements. For example, extending the effective range of these tests to reach more highly educated and experienced personnel seems to require modification in two respects: (1) the situations depicted should be more appropriate to the interests and vital experiences of such personnel; at present, most of the situations appear to be rather simple, socially and motivationally, making it difficult to elicit real identification and projection on the part of more sophisticated subjects; (2) in spite of the projective approach, there may still be too much transparency in regard to what constitutes "right" and "wrong" answers; the use of more subtle situations and responses, and possibly also the use of forced-choice responses, may produce marked improvements in the suitability of such tests for higher-level personnel.

The latter point, namely forced-choice coupling of response alternatives, may also be a fruitful method of improving the discrimination power of the items for any level of personnel. But more than this should be done to attain this objective: new items should be prepared whose content is likely to generate validity. Clues from the scrutiny of present valid and invalid items should be used as a basis for this work.

Among these clues are the following:

A. For Picture Interpretation Test, items depicting high or low prestige activities seem more frequently to be valid. The same is true of items portraying leadership or dominance.

B. For Picture Fill-In Test, the discriminating responses seem to be those which reflect of extra-punitiveness and intra-punitiveness. Social appropriateness or inter-personal skill seems to be another dimension along which a number of the items discriminate.

The elimination of the many non-discriminating items from the present forms of both of these tests, and their replacement by items constructed along the lines suggested above may well be productive of greater validity.

Most of the facts needed to effect modifications along lines suggested above already are at hand. For example, the types of items which seem most productive of validity could readily be classified beyond the point already accomplished. Also, as regards the forced-choice possibility, preference values (defined in terms of frequency of response selection) and validity coefficients are available for all items.

The basic decision that must first be made is whether or not to proceed further with instruments of this type. It is felt that the evidence disclosed in this investigation is that the Army possesses at least two new type tests which show appreciable validity for leadership criteria while being, at the same time, relatively independent of existing instruments used for leadership assessment. These considerations cogently denote the promise inherent in further exploration and development along these lines. This is particularly true in view of the critical need for techniques of leadership assessment and the paucity of existing means for meeting this need.