Department of the Army
Pamphlet 611–2

Personnel Selection and Classification

ARMY PERSONNEL TESTS AND MEASUREMENT

Headquarters
Department of the Army
Washington, DC
22 June 1962

UNCLASSIFIED

SUMMARY of CHANGE

DA PAM 611–2
ARMY PERSONNEL TESTS AND MEASUREMENT

o

o

RESERVED

FOREWORD

The purpose of this pamphlet is to provide an understanding of how the Army applies personnel psychology and statistical methodology to its personnel problems.

This pamphlet is addressed to two audiences. The first is composed of those officer and enlisted personnel whose assignments entail responsibility for using personnel measurement techniques and procedures or for instructing others in their use either in schools or on the job. The second audience consists of officer and enlisted personnel interested in gaining a better understanding of technical aspects of personnel management.

Technical content of the pamphlet is treated in the light of practical military personnel problems. The solutions to these problems are difficult; therefore, considerable study and thought will be required of the reader. So far as possible, technical terms have been avoided, but where such terms are used, their meanings are explained.

The pamphlet emphasizes basic principles involved in Army personnel problems. Theoretical considerations have been included only when necessary to an understanding of Army personnel problems and the techniques employed in dealing with them. This is a pamphlet on Army personnel psychology, not on general personnel psychology. Specific instructions for the various personnel measurement instruments used by the Army are contained in appropriate directives.

Throughout the text, where the word “man” is used, it may be understood to mean “man or woman,” since the same principles govern testing of both men and women.

Headquarters
Department of the Army
Washington, DC
22 June 1962

*Department of the Army Pamphlet 611–2

Personnel Selection and Classification

ARMY PERSONNEL TESTS AND MEASUREMENT

By Order of the Secretary of the Army:

G. H. DECKER,
General, United States Army
Chief of Staff

Official:

J. C. LAMBERT
Major General, United States Army
The Adjutant General

History. This publication has been reorganized to make it compatible with the Army electronic publishing database. No content has been changed.

Summary. See chapter 1, section V, paragraph 14.

Applicability. Not Applicable.

Proponent and exception authority. The proponent is the Office of the Secretary of the Army.

Suggested Improvements. Not Applicable.

Distribution. *Active Army:* To be distributed in accordance with DA Form 12–9 requirements for DA Regulations, Military Personnel Procedures, Officer and Enlisted—B.
NG and USAR: To be distributed in accordance with DA Form 12–9 requirements for DA Regulations, Military Personnel Procedures, Officer and Enlisted—A.

Contents (Listed by paragraph and page number)

Chapter 1
PERSONNEL MEASUREMENT AS AN AID IN PERSONNEL MANAGEMENT, *page 1*

Section I
THE EFFECTIVE UTILIZATION OF MANPOWER, page 1
General • 1, *page 1*
Analyzing Job Requirements • 2, *page 1*
Differences among Army Personnel • 3, *page 1*
Classification a Continuing Process • 4, *page 2*
The Army's Needs and Civilian Specialties • 5, *page 2*

Section II
NEED FOR SCIENTIFIC PERSONNEL METHODS, page 4
Inadequacy of Personal Judgment • 6, *page 4*
The Problem of Measurement • 7, *page 4*

Section III
PERSONNEL MEASUREMENT IN PRACTICE, page 5
General • 8, *page 5*
Progress In Personnel Measurement • 9, *page 5*
Calculated Risks in Personnel Measurement • 10, *page 5*
Personnel Research as a Practical Approach • 11, *page 5*

*This pamphlet supersedes TM 12–260, 9 April 1953.

UNCLASSIFIED

Contents—Continued

Section IV
THE SCOPE OF PERSONNEL RESEARCH, page 5
Classification • 12, *page 5*
Personnel Research and Its Relation to Other Areas of Research and Policy in Human Resources • 13, *page 6*

Section V
SUMMARY, page 6
The Purpose of Personnel Research • 14, *page 6*
Title not used. • 14B, *page 7*

Chapter 2
HOW THE ARMY DEVELOPS ITS PERSONNEL MEASURING INSTRUMENTS, *page 7*

Section I
MAJOR CONSIDERATIONS, page 7
General • 15, *page 7*
Types of Instruments • 16, *page 7*
Importance of Understanding How Personnel Measuring Instruments Are Constructed • 17, *page 7*
Tests Are Designed To Meet Specific Army Needs • 18, *page 8*
Suiting the Instrument to Its Purpose • 19, *page 8*
Form of Test • 20, *page 8*
Multiple–Choice Items • 21, *page 11*
Length of Test • 22, *page 11*
Time Limits • 23, *page 12*
Other Considerations in Planning Tests • 24, *page 12*

Section II
PREPARING THE TEST, page 12
General • 25, *page 12*
Construction of Test Items • 26, *page 12*
Preparing Directions • 27, *page 13*
Preparation of Scoring Directions • 28, *page 13*

Section III
TRYING OUT THE TEST, page 13
General • 29, *page 13*
Field Trials and Analysis • 30, *page 13*
Selection of Test Items • 31, *page 15*
Final Form of Test • 32, *page 17*

Section IV
ESTABLISHING THE SCALE OF MEASUREMENT, page 17
Purpose of Standardization • 33, *page 17*
Standard Reference Population • 34, *page 18*
Minimum Qualifying Scores • 35, *page 18*

Section V
INSTRUMENT BATTERIES, page 18
The Need for More Than One Instrument • 36, *page 18*
Which Instruments Should Be Used • 37, *page 18*
Other Ways of Using a Number of Instruments • 38, *page 19*
Continuing Studies • 39, *page 20*

Section VI
SUMMARY, page 20
Steps in Instrument Construction • 40, *page 20*

Contents—Continued

Title not used. • 40B, *page 20*

Chapter 3
CRITERIA, *page 21*

Section I
CRITERIA AND VALIDITY, page 21
Definition of Criteria • 41, *page 21*
Part the Criterion Plays in Validating a Test • 42, *page 21*
How Criteria Contribute to Classification • 43, *page 21*
How the Criterion Measure Is Used • 44, *page 22*
Importance of the Criterion • 45, *page 23*

Section II
CHARACTERISTICS OF ADEQUATE CRITERIA, page 23
What Makes a Satisfactory Criterion • 46, *page 23*
Comprehensiveness and Weighting • 47, *page 23*
Freedom From Bias • 48, *page 24*
Consistency of the Criterion Measure • 49, *page 24*
How Criteria Are Selected • 50, *page 24*

Section III
KINDS OF CRITERION MEASURES, page 25
Production Records • 51, *page 25*
Cost Records • 52, *page 25*
Rank or Grade • 53, *page 25*
Course Grades • 54, *page 25*
Job Performance Measures • 55, *page 25*
Ratings • 56, *page 26*

Section IV
RATINGS AS CRITERION MEASURES, page 26
Ratings Are Recorded Judgments • 57, *page 26*
Description of Rating Scales • 58, *page 27*
Limitations of Rating Scales • 59, *page 29*
Other Methods—Ranking, Nomination • 60, *page 29*
Improving Criterion Ratings • 61, *page 30*
Criterion Studies • 62, *page 30*
Relation to Policy • 63, *page 32*

Section V
SUMMARY, page 32
The Criterion Plays a Critical Role in the Development of Personnel Measuring Instruments • 64, *page 32*
Title not used. • 64B, *page 32*

Chapter 4
THE MEANING OF SCORES, *page 32*

Section I
THE NATURE OF PERSONNEL MEASUREMENT, page 32
Measurement Is Approximate • 65, *page 32*
Scores Are Relative • 66, *page 33*

Section II
TYPES OF SCORES AND STANDARDS, page 33
Adjectival Measures • 67, *page 33*

Contents—Continued

Raw Numerical Scores • 68, *page 33*
Percentile Scores • 69, *page 34*

Section III
STANDARD SCORES, page 35
General • 70, *page 35*
Standard Scores Measure Relative Performance • 71, *page 35*
Computation of Standard Scores • 72, *page 36*
The Army Standard Score • 73, *page 36*
Interpretation of Army Standard Scores • 74, *page 36*
Normalized Standard Scores • 75, *page 38*
Conversion Tables • 76, *page 38*
Advantages of Army Standard Scores • 77, *page 39*
Army Standard Scores Are NOT "IQ's" • 78, *page 39*

Section IV
RELIABILITY, page 39
Definition • 79, *page 39*
Importance of Reliability in Psychological Measurement • 80, *page 39*
Conditions Affecting Reliability • 81, *page 39*
Estimating the Reliability of a Test • 82, *page 39*
Uses of the Reliability Coefficient • 83, *page 40*

Section V
SUMMARY, page 40
Standard Scores Aid Interpretation • 84, *page 40*
Title not used. • 84B, *page 40*

Chapter 5
THE PRACTICAL VALUE OF SCORES, *page 41*

Section I
THE PRACTICAL SIGNIFICANCE OF SCORES, page 41
General • 85, *page 41*
Ambiguity of Instrument Titles • 86, *page 41*
Factors Affecting the Use of Scores • 87, *page 41*

Section II
VALIDITY AND SCORES, page 41
The Correlation Coefficient • 88, *page 41*
The Validity Coefficient • 89, *page 43*
Non–Empirical Validity • 90, *page 44*

Section III
SELECTION RATIO, page 45
Definition of Selection Ratio • 91, *page 45*
Significance of the Ratio • 92, *page 45*
Application in the Army • 93, *page 45*

Section IV
MINIMUM QUALIFYING SCORES, page 45
Factors in Determining Minimum Qualifying Scores • 94, *page 45*
Conflicting Demands for Men • 95, *page 45*
Minimum Standards of Performance • 96, *page 45*
Need for Uniform Standards • 97, *page 46*

Contents—Continued

Section V
SOME GENERAL CONSIDERATIONS INVOLVING STATISTICAL SIGNIFICANCE, page 46
Necessity for Sampling • 98, *page 46*
Errors Due to Sampling • 99, *page 46*
Factors Affecting the Adequacy of a Sample • 100, *page 46*
Statistical Results as Estimates • 101, *page 46*
Test Scores as Estimates • 102, *page 47*
Significance of Differences • 103, *page 47*
Practical Significance as Compared With Statistical Significance • 104, *page 47*

Section VI
SUMMARY, page 48
The Interpretation of Test Scores • 105, *page 48*
Title not used. • 105B, *page 48*

Chapter 6
USE OF APTITUDE MEASURES IN INITIAL CLASSIFICATION, *page 48*

Section I
INITIAL CLASSIFICATION, page 48
Purpose • 106, *page 48*
Basis for Initial Classification • 107, *page 48*

Section II
APTITUDE AREAS, page 49
What Differential Classification Is • 108, *page 49*
Definition of Aptitude Areas • 109, *page 49*
Development of Aptitude Areas • 110, *page 49*
How Aptitude Areas Are Used in Classification • 111, *page 50*
Gain Resulting From Classification by Aptitude Areas • 112, *page 50*
Problems in Classification by Aptitude Areas • 113, *page 52*

Section III
SUMMARY, page 53
Aptitude Measures as Aids in Classification • 114, *page 53*
Title not used. • 114B, *page 53*

Chapter 7
ACHIEVEMENT TESTS, *page 53*

Section I
CONSTRUCTION AND EVALUATION OF ACHIEVEMENT TESTS, page 53
What Achievement Tests Are • 115, *page 53*
Types and Characteristics of Achievement Tests • 116, *page 53*
Requirements for Job Proficiency Measures • 117, *page 53*
Planning Achievement Test Content • 118, *page 54*
Are Paper–and–Pencil Tests Practical? • 119, *page 54*
Evaluating Achievement Testing • 120, *page 54*
Objectivity of Achievement Tests • 121, *page 54*
Reliability of Achievement Tests • 122, *page 55*
Validity of Achievement Tests • 123, *page 55*

Section II
USES OF ACHIEVEMENT TESTS, page 57
Classification of Army Personnel • 124, *page 57*

Contents—Continued

Use of Achievement Tests in Training Programs • 125, *page 57*

Section III
SUMMARY, page 58
Achievement Tests as Measures of Proficiency • 126, *page 58*
Title not used. • 126B, *page 58*

Chapter 8
INTERVIEWING AS MEASUREMENT, *page 59*

Section I
PURPOSE OF INTERVIEWS, page 59
General • 127, *page 59*
Fact–Finding Interviews • 128, *page 59*
Interviews to Survey Attitudes • 129, *page 59*
Assessment Interviews • 130, *page 59*

Section II
THE INTERVIEW AS A MEASURING INSTRUMENT, page 59
The Value of Interviews as Measuring Instruments • 131, *page 59*
Limitations of Interviews as, Measuring Instruments • 132, *page 60*

Section III
SUMMARY, page 60
Interviews as Personnel Measuring Instruments • 133, *page 60*
Title not used. • 133B, *page 60*

Chapter 9
SELF–REPORT FORMS, *page 60*

Section I
THE NATURE OF SELF–REPORT FORMS, page 60
General • 134, *page 60*
Kinds of Self–Report Data • 135, *page 60*
Advantages of Self–Report Forms • 136, *page 61*
Uses • 137, *page 61*

Section II
CONSTRUCTING THE SELF–REPORT FORM, page 61
General • 138, *page 61*
Constructing Items for Self–Report Forms • 139, *page 61*
Validation of Experimental Self–Report Items • 140, *page 61*

Section III
A SUPPRESSOR METHOD OF IMPROVING VALIDITY OF SELF–REPORT FORMS, page 64
Errors of Conscious Distortion • 141, *page 64*
Errors of Unconscious Distortion • 142, *page 64*
Control of Distortion by a Suppressor Key • 143, *page 64*

Section IV
FORCED–CHOICE METHOD OF IMPROVING VALIDITY, page 64
General • 144, *page 64*
What Is a Forced–Choice Instrument? • 145, *page 64*
Grouping Alternatives • 146, *page 64*
Checking Validity of Scoring Key • 147, *page 65*

Contents—Continued

Section V
SUMMARY, page 67
The Value of Self–Report Forms • 148, *page 67*
Title not used. • 148B, *page 67*

Chapter 10
RATINGS AS MEASURES OF USEFULNESS, *page 67*

Section I
GENERAL CHARACTERISTICS OF ADMINISTRATIVE RATINGS, page 67
How Ratings Differ from Tests • 149, *page 67*
Title not used. • 149B, *page 67*

Section II
PURPOSES OF RATINGS, page 68
Ratings Classified According to Purpose • 150, *page 68*
Specificity of Purpose and Effectiveness in Rating • 151, *page 68*

Section III
ADMINISTRATIVE RATING METHODS, page 68
Need for Effective Performance Ratings • 152, *page 68*
Methods Unsuitable for Administrative Ratings • 153, *page 69*
Rating Methods Suitable for Administrative Reports • 154, *page 69*
Administrative Rating as Procedure • 155, *page 71*

Section IV
MAJOR PROBLEMS IN ADMINISTRATIVE RATING, page 72
Acceptance of the Rating Procedure • 156, *page 72*
Rater Tendencies in Army Administrative Rating • 157, *page 72*
Validating Administrative Rating Procedures • 158, *page 73*
Standardizing Administrative Ratings • 159, *page 73*
Improving Administrative Ratings • 160, *page 75*

Section V
SUMMARY, page 75
Ratings for Administrative Purposes • 161, *page 75*
Title not used. • 161B, *page 76*

Chapter 11
THE ADMINISTRATION OF ARMY TESTS, *page 76*

Section I
INTRODUCTION, page 76
General • 162, *page 76*
Authorized Test Instructions • 163, *page 76*
Testing Situation Must Be Standard • 164, *page 76*

Section II
PRINCIPLES AND PROCEDURES FOR ADMINISTERING GROUP TESTS, page 76
General • 165, *page 76*
Physical Surroundings • 166, *page 77*
Testing Session • 167, *page 79*
Preparation for the Testing Session • 168, *page 79*
Administering the Test • 169, *page 79*
Collection and Disposition of Test Materials • 170, *page 81*
Care of Booklets • 171, *page 82*

Contents—Continued

Answering Examinees' Questions • 172, *page 82*

Section III
ADMINISTERING INDIVIDUAL TESTS, page 82
General • 173, *page 82*
Individual Testing Session • 174, *page 82*
Timing • 175, *page 83*

Section IV
SUMMARY, page 84
Importance of Proper Administration • 176, *page 84*
Title not used. • 176B, *page 84*

Chapter 12
SCORING ARMY TESTS, *page 84*

Section I
SOME GENERAL CONSIDERATIONS, page 84
Importance of Accurate Scoring • 177, *page 84*
Title not used. • 177B, *page 84*

Section II
SCORING PROCEDURES, page 84
What Is "Scoring" • 178, *page 84*
Meaning of the Scoring Formula • 179, *page 84*
Hand Scoring and Machine Scoring • 180, *page 85*
The Scoring Team • 181, *page 85*
Checking • 182, *page 85*

Section III
HAND SCORING, page 85
Hand Scoring Test Booklets • 183, *page 85*
Hand Scoring Separate Answer Sheets • 184, *page 85*

Section IV
MACHINE SCORING, page 86
Requisites for Machine Scoring • 185, *page 86*
How the Machine Obtains a Score • 186, *page 86*
IBM Manual of Instruction • 187, *page 86*
Scanning Answer Sheets Before Scoring • 188, *page 87*
Setting Up and Balancing the Scoring Machine • 189, *page 88*
Scoring and Checking the Answer Sheets • 190, *page 90*
Additional Checking of Machine Scoring • 191, *page 90*
Summary of Steps in Machine Scoring • 192, *page 90*

Section V
RECORDING SCORES, page 90
Importance of Accurate Recording • 193, *page 90*
Title not used. • 193B, *page 90*

Section VI
SUMMARY, page 90
Accuracy in Scoring • 194, *page 90*
Title not used. • 194B, *page 90*

Contents—Continued

Chapter 13
HOW THE ARMY USES PERSONNEL MEASURING INSTRUMENTS, *page 91*
Administration and Use of Army Personnel Measuring Instruments • 195, *page 91*
The Modern Army a Challenge to Personnel Management • 196, *page 91*

Appendix A. References, *page 94*

Table List

Table 4–1: Raw Numerical Scores tables, *page 34*
Table 4–2: Scores illustration tables, *page 35*

Figure List

Figure 1: How men in a unit may differ in quality of performance, *page 3*
Figure 2: Aptitude and achievement tests serve different purposes, *page 10*
Figure 3: A good sample comes from all parts of the lot, *page 12*
Figure 4: Internal consistency is not necessarily an index to the usefulness of the item, *page 16*
Figure 5: Illustrations of high validity and low validity for items at three levels of difficulty, *page 17*
Figure 6: Criterion tests a test, *page 22*
Figure 7: Examples of three types of rating scales, *page 28*
Figure 8: An example of ranking as a rating method, *page 29*
Figure 9: Comparison of raw score units, percentile score units, and Army standard score units for a distribution of test score with raw score mean of 26 and standard deviation of 8, *page 35*
Figure 10: Graph of a distribution of test scores, *page 37*
Figure 11: Standard scores, Army standard scores, and percentile scores in relation to the normal curve of distribution, *page 38*
Figure 12: Illustration of the relationship between test and criterion scores for 5 men, *page 42*
Figure 13: Correlation scatter diagrams illustrating different degrees of relationship, *page 43*
Figure 14: Aptitude Area profiles for two soldiers, *page 51*
Figure 15: Percentage of enlisted input with standard scores of 100 or higher on AFQT and on best 1958 aptitude area, *page 52*
Figure 16: Comparative objectivity of standard answer tests and essay tests, *page 56*
Figure 17: Types of items used in self–report forms, *page 63*
Figure 18: Important characteristics of items considered in developing self–report forms, *page 66*
Figure 19: Rating scale from Commander's Evaluation Report, DA Form 2166, *page 70*
Figure 20: Forced–choice phrases may be assembled in pairs or tetrads, *page 71*
Figure 21: Two important difficulties in obtaining valid ratings, *page 74*
Figure 22: Factors affecting test performance, *page 78*
Figure 23: Some conditions detracting from good test environment, *page 80*
Figure 24: Errors frequently found on answer sheets, *page 88*
Figure 25: Some typical uses of personnel measurement instruments, *page 92*
Figure 25: Some typical uses of personnel measurement instruments—Continued, *page 93*

Glossary

RESERVED

Chapter 1
PERSONNEL MEASUREMENT AS AN AID IN PERSONNEL MANAGEMENT

Section I
THE EFFECTIVE UTILIZATION OF MANPOWER

1. General

An effective fighting force needs the right kind of man as well as the right kind of equipment. The right kind of man is the man who is suited to his job; he meets the requirements and he is not wasted. If it is known what the requirements are and what the characteristics of the men are, the jobs and the men can be matched.

a. Effective Utilization of Manpower Means Matching Jobs and Men. It does not mean that only the best possible men will be accepted. The manpower barrel has a bottom and the "cream of the crop" is at best only a thin layer. How far down the barrel it is necessary to go is a matter of high–level policy and depends on how great the need is. It also depends on how successfully the men who are taken are used. The Army needs to know how the manpower barrel is made up. It is the overall objective of personnel measurement to provide the essential information on the abilities which make up the Army's manpower.

b. Army Personnel Research Is Concerned With Discovering Techniques That Will Help Match Army People With Army Jobs. On the one hand, each Army job must be analyzed into the tasks which make it up and the skills needed to do the job. On the other, the abilities and skills that men and women bring with them from civilian life, or that they acquire in the Army, must be identified and described.

2. Analyzing Job Requirements

New occupations are constantly being developed to meet the changing needs of the Army. There are over 300 military occupations in the Army, all of which have been established out of practical necessity. Once a job is recognized as essential in the work of a unit, it is added to the table of organization and a job description is worked out for it. The job description defines what a man is expected to do on the job and the degree of proficiency required. Selection standards for the job are established, and training courses organized, when required, to enable Army personnel to meet the job requirements.

3. Differences among Army Personnel

a. General. The men and women who must be selected and assigned to Army training courses and jobs vary in many ways. The fact that each individual possesses more of some abilities than of others is important to the Army. Effective utilization of manpower is not possible without knowledge of the strengths and weaknesses of the individuals that make up the manpower pool. It is a severe loss to the Army to fail to use to advantage a man capable of developing into a good leader, just as it is a severe loss to place a man in a position of leadership and find out too late the he is incapable of carrying out his responsibilities.

b. Differences in Physical Characteristics. Even after screening by the physical examination at the induction station and subsequent physical conditioning, soldiers differ widely in health, strength, size, and endurance. Some men can march 30 miles a day with full field equipment; others can cover only a few miles under the same conditions. Some can resist extremes of temperature, others cannot; some can maintain their efficiency at high altitudes, others lose it; some can see well at night, others are practically night blind. No matter how effective the screening or conditioning, soldiers will not show any great uniformity in physical characteristics.

c. Differences in Psychological Characteristics—Ability and Personality. The differences among soldiers in abilities and personality characteristics are just as large and as important as are the differences in physique, stamina, and the keenness of their senses. However, these abilities and personality characteristics cannot be observed as directly as can physical traits. It is not possible to tell by looking at a man whether he can add or spell, repair a carbine, lead a squad, or be aggressive under fire. Nor is it possible to tell by such direct observation whether the man can learn to do these things in the relatively short time available for training. To obtain useful estimates of psychological characteristics, it is necessary to use other methods—the methods which are described in this pamphlet.

d. Differences in Effects of Training. Among Army men in the same training situation or in the same job the range in ability, and in capacity to absorb training, may be tremendous. Figure 1 illustrates the spread of men in one training unit rated according to the quality of their performance in the unit. Although all of these men had been subjected to the same training methods and schedule, they were not all rated as of equal value. Some were rated as of little value, some were rated as of "average" value, and some were rated as outstanding. They differed in their capacity to acquire the skills in which they were being trained Men differ also in the level of skill they can reach with training. For instance, it is exceedingly doubtful if the poorest performers could be brought up to the level of the best even if limitless time were available. Differences among men are not always ironed out by training. Differences in level of performance may even increase under a program of instruction. Those who are more skillful to start with may improve faster than the less skillful, with the result that the range of abilities after training may be even greater than before.

4. Classification a Continuing Process

The jobs which the Army must fill rarely, if ever, require exactly the same abilities and aptitudes as are possessed by the men available for assignment. Since the Army requirements must be met, it is necessary that they take precedence over the soldier's abilities and interests whenever there is a conflict. The Army frequently will have to use men in duty assignments in which their abilities are not fully utilized. Later, as requirements change with the tactical situation, there may be a demand for more men with their level of ability, and consequently, they may be shifted to more appropriate jobs. Effective utilization of manpower means that such men are not lost sight of, that their superior abilities are known, recorded, and, whenever possible, utilized.

5. The Army's Needs and Civilian Specialties

a. Many persons coming into the Army would like to perform jobs corresponding to those they performed in civilian life. Very often, since a man's work experience is carefully considered in the classification process, that is exactly what happens. Sometimes, however, there is a conflict between the needs of the Army and a man's expectation that he will be assigned to a particular occupation in which he believes—and with reason—that he will do best. In that case, the individual's desires must give way to the overall needs of the Army. It must be recognized that the Army jobs to be filled and the men qualified to fill them never match exactly. Some compromise must always be made, and this compromise, of necessity, will be in the direction of adjusting the supply of manpower to the requirements of the Army. A typical assignment problem will show how such conflicts usually are resolved—To fill 30 truck driver jobs there may be available 100 men whose outstanding civilian skill was truck driving. Since only 30 of these men can be assigned to truck driving duties, the simplest procedure would be to pick the 30 best drivers. The remaining 70, obviously, would be utilized in assignments other than truck driving. However, the solution is not always this easy. Suppose the demand for tank crewmen is more critical than the demand for truck drivers. To assign the 30 best civilian truck drivers to truck driving may result in sending poorer quality to tank crewman school—to train for the more critical specialty. Hence, it may be more desirable to assign the 30 best civilian drivers to tank crewman school.

b. Consider another example—a civilian lawyer may feel that by virtue of his training and experience he could serve best as a legal officer. This may be true as far as he is concerned. However, as far as the Army is concerned, there may be an oversupply of legal talent and an under–supply of artillery officers, a job which has no civilian counterpart. The problem then resolves itself into determining what the characteristics of good artillery officers are, and whether some of the lawyers possess those characteristics, so that they may be assigned to artillery training. A good lawyer may thus become a good artillery officer.

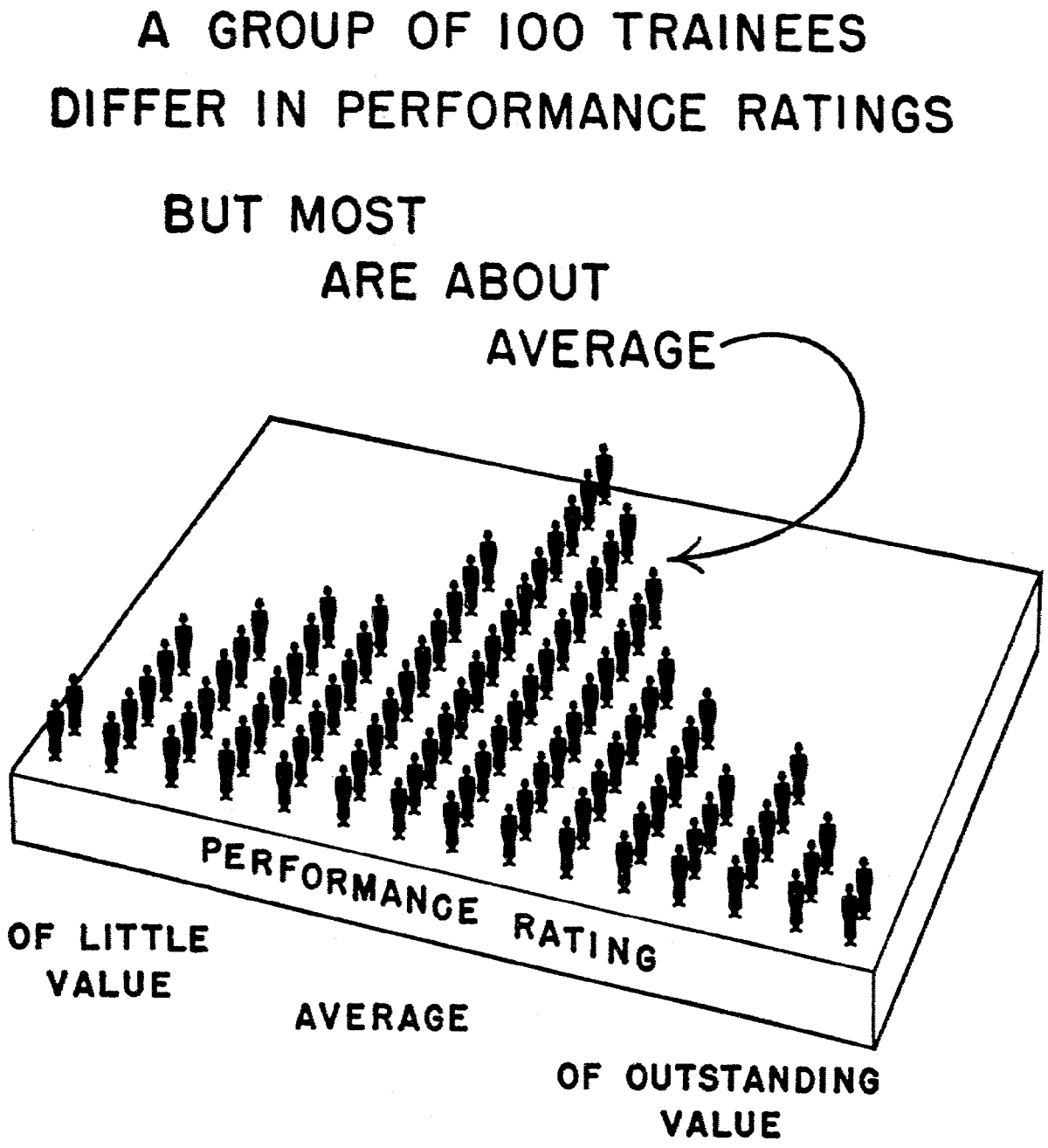


Figure 1. How men in a unit may differ in quality of performance

Section II
NEED FOR SCIENTIFIC PERSONNEL METHODS

6. Inadequacy of Personal Judgment

Some men are good judges of human nature. If there were enough good judges of human nature and if they had enough time to use their judgment critically, and if their judgments could be passed along to someone else without misunderstanding, there would be no need for other methods. Since none of these ideal conditions exist, it is necessary to rely upon the tools of personnel measurement, on aptitude tests, ratings, and the like.

a. Need for Objectivity. Some men may be able to "size up" a person and his qualifications with considerable accuracy. More often, however, such judgments are affected by prejudice. There may be insurmountable difficulties in passing along the judgment without misunderstanding. What one judge considers "good" performance may seem inadequate to another. Furthermore, personal judgments too frequently are not reasoned judgments. Too often they are no more than guesses—and uninformed guesses at that. There are some people who still believe that a man's ability and personality can be judged by his appearance. There is, however, abundant evidence that such judgments are generally too inaccurate for any important decisions to be based on them.

b. Need for More Efficient Methods. Personal judgment may serve as a personnel tool in Army units where a commander knows his subordinates, where he can decide personnel actions on an individual basis and observe how men work out in their assignments. However, this kind of classification is not common in the Army. Assignments to job or training more often are determined in classification or training centers where the information available about the man being assigned is limited to what has been obtained in the course of his initial classification and in subsequent actions such as performance ratings. It is essential, then, that this information provide the essential clues to a man's usefulness to the Army. It is not enough, however, to provide all the information essential to proper classification of a recruit. Such information must be in usable form. A classification officer, for example, who had at his disposal all the details of a soldier's personal and work history might be able to select an occupation which would be best for the man and, at the same time, work to the best interests of the Army. But it would be a difficult task and would mean consideration of a great number of nonessential facts. It is obvious that such a detailed study of every man and woman coming into the Army is neither feasible nor desirable. It would take too much time; there would be no objective guide to the soundness of the decision. Nor is it necessary; there are other methods which are as effective and even more effective—the methods of scientific personnel measurement. The effectiveness of selection, classification, assignment, and utilization would be seriously reduced if scientific personnel measurement tools were not available.

c. Applying Scientific Tools. The application of scientifically derived objective tools does not result in "an inhuman, mechanistic system." On the contrary, it helps organize and put to effective use more pertinent information about a man than an individual appraiser or interviewer normally could be expected to take into account. Scientifically derived formulas or measures such as are used in the aptitude area system represent the distillation of the best brains and the most skillful techniques available to the Army. Therefore, every man being processed receives the benefit of their efforts.

7. The Problem of Measurement

a. The problem of measurement in personnel psychology is fundamentally the same as in other scientific fields. It is the problem of defining the units in an unmistakable and useful way. What does the number 25 mean unless there is an added element such as "inches," "pounds" or "degrees"? Similarly, in personnel measurement, a number has no useful meaning unless the unit of measure is attached. Thus, a man may be assigned the number 25 because he gave correct answers to 25 test questions out of 26; he may also be assigned the number 25 because he gave correct answers to 25 test questions out of 100. Obviously, the number 25 does not mean the same thing in both instances; or stated another way, the unit needed is not just the number of problems correctly answered—the unit must also describe the number of problems he was required to answer.

b. Even this addition is not enough to give useful meaning to the number. How many problems should he be required to answer? Suppose two tests of 100 problems each are given to a group of men. Suppose, further, that on one of the tests perfect scores are made by a number of the men, but that on the other test, the best score ever made was, say, 65 right. Again obviously, a score of 25 on the two tests does not mean the same thing, even though both tests had the same number of problems or items. It does not mean the same thing because it does not represent the fact that one test was relatively easy (perfect scores were made by a number of the men), whereas the other test was relatively difficult (no one ever approached a perfect score).

c. Without attempting to be exhaustive at this point, only one further consideration will be raised. Even after all these definitions are attached to the number, there is still lacking one essential element of meaning. Is the number a "good" score or a poor score? Is the man who gets that numerical score a "good" man or a "poor" man? What, if anything, does the number reveal about the man's performance on other tests, the adequacy of his duty performance, his capacity for benefiting by further training, his promotability?

d. To sum it up, the problem of measurement in personnel psychology is not usefully solved until the numerical values used make clear how a man stands in comparison with other men and how well performance on the test predicts other performance. That is to say, standardization and validation are necessary to give useful meaning to the numerical values.

Section III
PERSONNEL MEASUREMENT IN PRACTICE

8. General

The primary purpose of all Army personnel evaluation instruments is to aid in measuring actual or potential job proficiency. Job proficiency does not mean just the performance of specific operations or movements. In some jobs, proficiency consists, to a large extent, of the ability to work with others as a member of a crew or team. On other jobs, leadership ability may be a prime requisite. Most combat jobs demand "courage under fire," and certain other jobs require a large amount of ability to understand people.

9. Progress In Personnel Measurement

a. Some human characteristics are easier to measure scientifically than others. For example, it is relatively easy to develop measures of ability to do the work at an Army technical school, but it is much more difficult to develop measures of "courage under fire" not so much because courage cannot be measured, but because it usually cannot be studied scientifically where it counts most, in combat, for example. But even here a start has been made.

b. The greatest progress has been made in measuring physical traits, characteristics such as dexterity and intelligence, and skills and knowledges of many kinds. Considerable progress has been made in measuring personality characteristics such as those involved in leadership, although much work remains to be done to obtain more exact and dependable methods of measurement. An important source of difficulty is that such characteristics are very complex and may depend on specific situations. That is, a person who is a good leader in one kind of situation may be a poor leader in another. It is apparently not useful to obtain a measure of "leadership"—it appears necessary to add "leadership for what, in which circumstances, and of which kinds of men."

10. Calculated Risks in Personnel Measurement

No personnel measuring instrument is perfect; that is why continuing research is needed. A major objective of personnel research in the Army is to strive continuously to improve the tests, the rating scales, the personality inventories and the other instruments used so that the amount of "error" in the measurements will be a minimum. What is of equal importance is knowing the limits of accuracy with which an instrument may be safely used, and separating what is feasible to measure from what it is not feasible to measure. This knowledge is important not only for the proper interpretation of particular scores but also for establishing policies governing personnel action. Should additional personnel information be sought? What followup is needed? Who should make the final decisions? To answer such questions it is necessary to know the limits of accuracy of the measuring devices. Thus, "calculated risks" are involved in personnel actions as well as in tactical operations. Research in personnel measurement provides a basis for calculating the risks in personnel actions.

11. Personnel Research as a Practical Approach

In order to furnish a basis for calculating risks in personnel actions involving personnel tests and other instruments, it is not enough to develop a theory or to speculate as to an instrument's value. It is necessary whenever possible to "test the test." How consistent is the instrument every time it is applied—that is, what is its reliability? And even if it is highly consistent, what is it really measuring? Not what someone thinks it measures, but what it really measures—that is, what is its validity for the use intended? The answers to these vital questions require a practical approach—field testing whenever possible. It is not an easy approach, for field testing requires a yardstick or criterion with which the results of administering the instruments can be compared, and such criteria are not easily arrived at. Problems concerning the use of yardsticks or criteria for determining the effectiveness of personnel measuring instruments are so important that a whole chapter of this pamphlet is devoted to them (ch. 3).

Section IV
THE SCOPE OF PERSONNEL RESEARCH

12. Classification

a. Personnel research in the Army is directed primarily at improving methods for selecting and evaluating personnel as individuals. The psychological and physical characteristics of a man must be measured before the Army knows whether he is soldier material. It is necessary to measure these characteristics at various stages of a soldier's career to determine what assignments he can be expected to perform satisfactorily and what special training it is profitable to give him. It is necessary to measure the outcomes of the training given him. Throughout his military career his abilities

must be reviewed and evaluated. As he acquires new military skills through training and experience or as the changing needs of the Army demand, his assignment is subject to revision.

b. Personnel research is also directed at improving methods for selecting and evaluating personnel as teams. It is necessary to fit together men with different characteristics to permit effective accomplishment of the mission of the team. The effectiveness of the team must be evaluated to determine what missions it can be expected to accomplish, what special training it is ready for, how much it has profited by the training given it, and how it has performed during tactical operations.

c. It is the business of personnel research to develop tools for selecting and evaluating personnel, as individuals and as teams, so that assignments and changes in assignment may have a sound basis.

13. Personnel Research and Its Relation to Other Areas of Research and Policy in Human Resources

a. Personnel research in the Army is concerned primarily with the scientific development of personnel and industrial psychological devices to select, classify, assign, utilize, and evaluate personnel. As a research activity, personnel research may have important relation to other areas of human factors research such as those concerned with physiological psychology, engineering psychology, learning psychology, training methods and devices, social psychology, work methods, systems research and analysis, and operations research. As a tool to improve the Army's utilization of manpower, personnel research may contribute to the making of personnel policy and is often the key to its execution.

b. Some examples may be given of the relations between personnel research and other problems involving military personnel. Suppose that a service school wishes to improve its tests and its grading system. Study may indicate that the tests and the grading system are adequate. Question may then arise as to the adequacy of the methods of instruction. Accordingly, research may be undertaken to compare several methods of instruction, and personnel measurement may be used to evaluate the effectiveness of the various methods. Or a question may be asked: What is the relationship between qualification in arms and overall value as an infantryman? In searching for the answer to this question, another question may arise–Are the current methods of weapons training in need of improvement? This may lead to still another question–Are the weapons designed to fit the men's abilities and thus permit the effective use of the weapons?

c. Suppose that research is under way to develop tests of night vision as part of an effort to improve combat effectiveness in night patrols. Before such tests can be developed, it may be necessary to study the difference between, say, night vision and day vision. Or suppose that research is under way to identify those Army jobs and courses of training for which selection is made on the basis of abilities in short supply but which could be selected on the basis of abilities in plentiful supply with little or no reduction in job effectiveness. In this way, the limited supply of abilities for the more critical highly skilled jobs would be conserved for use where it is most needed. Before such research can be successfully concluded, it may be necessary to study factors contributing to attrition in courses of training. Since attrition is not always a function of selection standards, it may be necessary to identify to what extent current selection standards, and to what extent standards other than selection standards, contribute to attrition. Or, for a final example, when the minimum qualifying score on the test for induction into the Army is to be set, a basis for the decision is provided by personnel research findings. The score as set reflects the best balance that can be struck between the number of men of marginal usefulness available to the Army and the number of such men who can be absorbed into the Army's job structure.

Section V
SUMMARY

14. The Purpose of Personnel Research

a. Army personnel research is concerned with discovering techniques that will facilitate the effective utilization of its manpower. This requires—

(1) Analysis of the psychological requirements for each job.

(2) Analysis of men in terms of—

(a) Individual differences in abilities.

(b) Their civilian specialties in relation to the manpower needs of the Army.

b. The size and scope of the Army requires the application of scientific personnel methods so as to increase efficiency in its personnel actions. Objective indicators of performance should replace personal judgment where possible, and effort should be directed at improving and systematizing personal judgment where needed.

c. The Army emphasizes a practical approach in its personnel research. Its purpose in using personnel evaluation instruments is to aid in measuring actual or potential job proficiency. Achievement of this purpose requires knowledge of the limits of accuracy of the instruments and is conditioned by the following facts:

(1) Some human characteristics are harder to measure than others.

(2) Validating measuring instruments requires considerable time and effort.

d. Personnel research is one area of human resources research. The products of personnel research are used by the Army to aid in the classification and utilization of personnel, as individuals and as teams. These products may also be used in other areas of human resources research, such as the evaluation of various training methods. Personnel research

may be used in the solution of other problems of personnel management and in the establishment and execution of personnel policy.

14B. Title not used.

Paragraph not used.

Chapter 2
HOW THE ARMY DEVELOPS ITS PERSONNEL MEASURING INSTRUMENTS

Section I
MAJOR CONSIDERATIONS

15. General

a. Most Army personnel measuring instruments are designed to be used for a particular personnel program. To apply an instrument to a personnel program other than the one for which it was designed is risky. The nature of the population taking the test may make considerable difference in the usefulness of the test. For example, an instrument developed for inductees is not necessarily applicable to trainees or NCO's even though the psychological characteristics to be measured appear to be the same. A test to measure how much men in basic training have learned about the functioning of a particular weapon might well be too easy for NCO's. If it is used with NCO's, it will not be possible to determine which of them know most about the weapon, since they will answer most of the questions correctly. Another test with more difficult questions is necessary.

b. The principles and practices followed in the development of personnel measuring instruments are described in this chapter. For simplicity's sake, the discussion is oriented toward the development of tests. It will be understood that the discussion applies, in general, to other types of personnel instruments as well.

16. Types of Instruments

a. The Personnel Measuring Instruments Used by the Army Are of Several Different Forms. Distinctions are often made between tests on the one hand, and rating scales, questionnaires, and interviews on the other. The term "test" is usually used for an instrument that requires answers which are either right or wrong. By way of contrast, the other types of instruments do not have right and wrong answers but rather yield indication of the degree to which a characteristic is possessed. In either case, it is important to determine the significance of the scores, whether they represent the number of right and wrong answers or the possession or absence of particular characteristics. As stated in chapter 1, it is necessary that instruments be valid if they are to be used effectively.

b. Personnel Measuring Instruments May Be Classified According to What They Are Intended to Measure. Some are intended to measure knowledge of a job or subject matter; others are intended to measure ability to acquire such knowledge.

(1) Some instruments (ratings) are intended to evaluate a man's performance, usually in terms of his value to his organization; others (self–description, forms) provide a means for the man to describe his past history, his likes and dislikes.

(2) A standard interview is an instrument used to evaluate how a man acts under prescribed circumstances. Sometimes an interview is used to find out what a man knows, although usually this can be done better with a suitable test or records (ch. 8). For special purposes, a "stress" interview may be used to evaluate the actions of a man under severe pressure. This type of interview is not widely used in the Army.

(3) Some measuring instruments are directed at obtaining knowledge of a man's attitudes and beliefs. Such instruments are not discussed in detail in this pamphlet. AR 600–45 describes the Sample Survey of Military Personnel, a periodic survey required by Headquarters, Department of the Army, employing questionnaires in which information about a man's attitudes, beliefs, opinions, and personal circumstances is obtained. The purpose of the survey is to provide information to assist the Department of the Army staff agencies in planning budget requirements, establishing personnel policies, supporting proposed legislation, answering inquiries from Congress and Governmental agencies, and determining attitudes, opinions, and characteristics of Army personnel.

c. Personnel Measuring Instruments May Be Classified in Other Ways. Some are administered to one man at a time; others—and in the Army, most—are administered to large groups. Some (most) require the use of language, usually English; others require only a minimum of language. Some (most) are of the familiar "paper–and–pencil" type; others require performance in a typical assignment situation. Some instruments are called aptitude tests; others, achievement tests.

17. Importance of Understanding How Personnel Measuring Instruments Are Constructed

All classification personnel in the classification system and all officers exercising a command function in regard to classification or assignment need a thorough understanding of the nature of Army personnel instruments and the proper

interpretation and use of scores. Such understanding can be aided by a brief description of the principles and practices of constructing instruments.

18. Tests Are Designed To Meet Specific Army Needs

a. Defining the Army Need. The first step in test making is to study the classification problem to be solved. The particular need of the Army must be clearly defined both in terms of the psychological requirements of the job or course of training and in terms of the number of men needed and the apparent supply. As a rule, the difficult and time–consuming process of instrument construction is warranted only when the problem is urgent and important. If a simple interview, a survey of past experience, or an examination of personnel records will do just as well, or if the number of men and jobs involved is small, the construction of a special instrument is not warranted.

b. Finding the Characteristics to Measure. Study of the problem also brings to light the characteristics which are related to successful performance. It is necessary to find characteristics, which are possessed in high degree by most (if not all) of the men who have demonstrated their ability in a particular course or assignment and which are not possessed by those who have failed. By measuring the amount of these traits possessed by untried men, it is possible to predict their performance on the job. In some cases, it is a simple matter to find characteristics highly correlated with success—successful carpenters know such things as the proper use of various tools and the various types of joints. Obviously, these are the skills and knowledges that should be measured in order to select the Army's carpenters. In other cases, the particular traits must be discovered through systematic trial made after the general purpose of the test is decided upon. For example, it was necessary to measure "cryptography aptitude" in order to predict which men would be most likely to pass a course in cryptography. It took considerable research to find which particular traits made up cryptography aptitude and to choose those most highly correlated with successful performance in the course. It is essential to select for measurement those traits which reflect actual job requirements.

19. Suiting the Instrument to Its Purpose

The use to which an instrument will be put is one of the principal factors which determine the kind of test to be developed and the form it will be given.

a. Achievement Tests. Achievement tests are instruments intended to find the men who possess the skill and knowledge necessary for successful performance in a particular assignment. The Army Language Proficiency Test (Spanish) is designed to measure the degree of fluency in Spanish an officer or enlisted man possesses. Scores on such tests may be used as aids in estimating future level of performance as an interrogator or translator or interpreter. Achievement tests may be used as a basis for assignment to some tactical or service organization. They contribute to the prediction of job proficiency by measuring how much knowledge or skill the man has acquired.

b. Aptitude Tests. Aptitude tests are instruments designed to predict which men are most likely to perform successfully in a training course or duty assignment. They predict success by measuring the degree to which examinees possess certain critical traits found in men who have been successful in the particular training program or duty assignment for which selection is being made. They are useful in selecting from among a mixed lot of available soldiers the most promising trainees for a particular course. They are particularly useful with the recruit whose job experience has been limited and whose potentialities for various kinds of available training are unknown. The fact that certain men get high scores on such tests would have little significance were it not for the fact that research studies have shown that such men are likely to benefit more by subsequent training for certain jobs than are men who get low scores.

c. Aptitude vs. Achievement Tests. Whether a personnel instrument is considered as measuring aptitude or achievement depends mainly upon the purpose of the instrument.

20. Form of Test

Two principal factors determine the form in which a test is cast—the purpose to be served and practicability. If a test is to be used on illiterates, obviously it must be in a form which minimizes the amount of language required. If a test is intended to show how a man does a job requiring physical handling of equipment and materials, a paper–and–pencil test of his knowledge of the job is not appropriate. Furthermore, a test is a field instrument and must be designed to serve efficiently under conditions likely to be found in the field. It must be adaptable to Army necessities which limit the facilities available and the time that may be allotted for its administration.

a. Verbal Test. A verbal test is one in which the examinee is required to talk, write, or mark correct responses stated in language. It is not limited to tests of vocabulary and reading. Most Army tests are verbal tests. The type commonly used is a paper–and–pencil test administered and scored in a short time and with great efficiency by men who need not be highly trained specialists. Conditions of administration can be made uniform, and highly objective scoring methods may be employed. A wide range of ability and knowledge may be sampled in a relatively short time. When many men are to be tested, or highly trained personnel are not available to administer tests, a verbal test is likely to be the most practical, unless it is clearly unsuited to the purpose at hand.

b. Nonlanguage Test. Where the examinees are illiterate or do not understand English, it is necessary to use pictures

and diagrams instead of verbally–stated problems, and to substitute demonstration for verbal instruction in administering the test. The Nonlanguage Test (NLT–2a and NLT–2b), administered to non–English speaking aliens enlisted by the U.S. Army in foreign countries, is an example of this type of test.

c. Performance Test. A performance test is one in which the examinee is required to manipulate objects, or make practical application of knowledge. It is the most efficient form to use when it is necessary to observe how the examinee does the job as well as his ability to do it, or when it is essential to measure ability to do a job rather than knowledge about the job, although the two are usually correlated with each other. The practicality of performance tests is limited in that they are time–consuming, expensive, usually require highly trained personnel for their administration, and are frequently difficult to score. The Driving Performance Test, including Manipulation of Controls, Practice Run, Test of Depth Perception, Check for Emergency Equipment, Before Operation Check, and Location of Instruments, is an example of this type of test.

d. Group Tests. Verbal (paper–and–pencil) tests are widely employed by the Army because they are best adapted to group testing. The large number of men to be classified and the pressure of time make it necessary to test in groups whenever possible. Performance tests, by their very nature, are better adapted to the testing of men one at a time. It is sometimes possible, however, to devise a performance test to be given to groups.

e. Individual Tests. Individual tests are administered to one man at a time. They may be of either the verbal or performance form, depending upon the particular purpose to be achieved; or a test may involve both performance and verbal responses. Individual tests are constructed to accomplish the following purposes for which group tests are not suited:

(1) Screening and classifying men whose language deficiencies or other personal characteristics do not enable them to demonstrate their abilities adequately on a group test.

(2) Testing aptitudes or proficiencies which can best be revealed through actual work samples or manipulation of materials.

(3) Determining from an examinee's behavior while taking a group test whether his group test score is consistent with other information available about him.

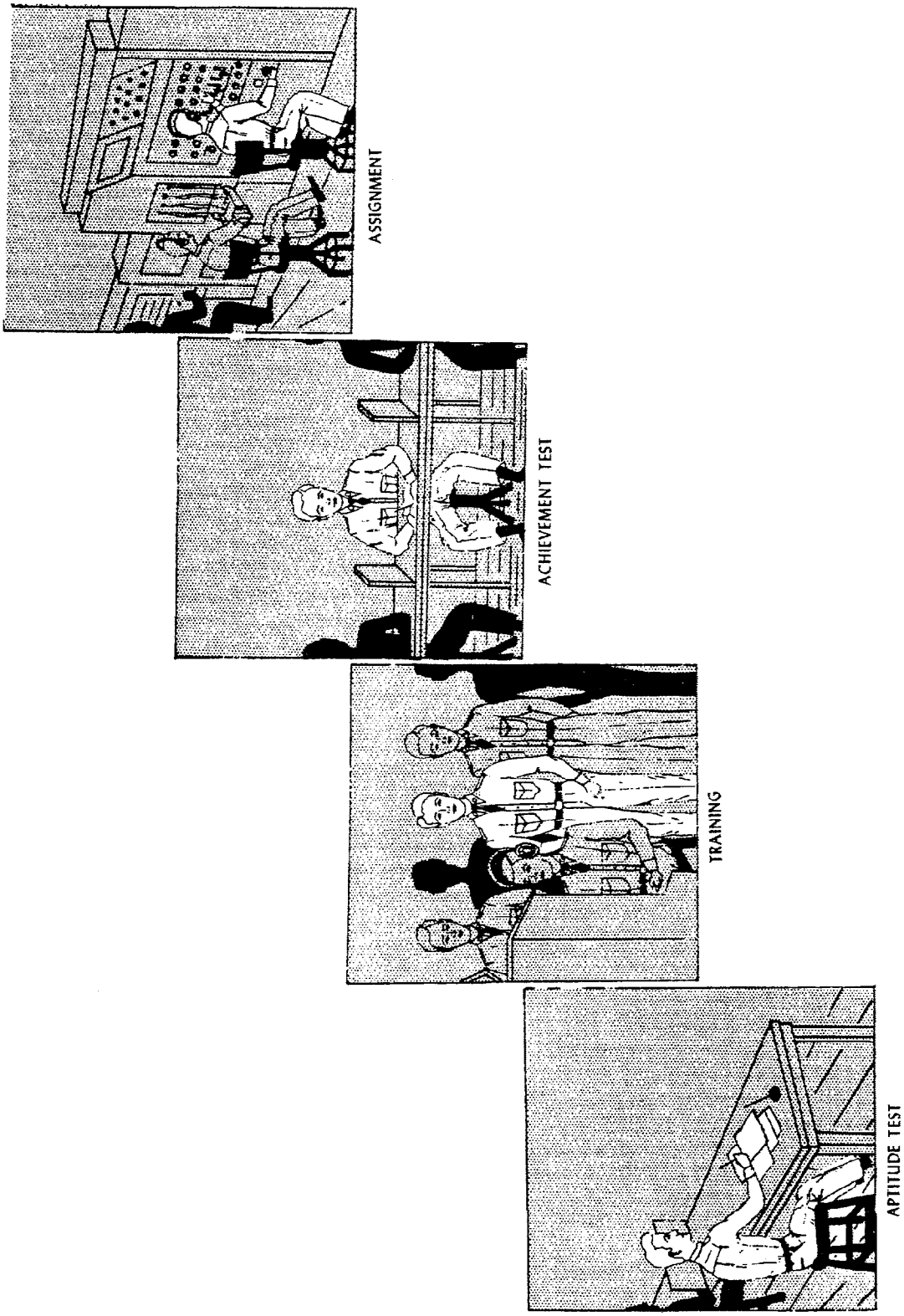


Figure 2. Aptitude and achievement tests serve different purposes

21. Multiple–Choice Items

Two steps in constructing the test thus far have been considered. The traits and characteristics to be measured have been decided upon, and the form which the test is to take has been determined. The next step is to determine in which form the questions or problems should be stated to reveal the desired information. They maybe brief statements requiring essay answers. Or they may be incomplete sentences or paragraphs requiring insertion of the correct terms. Other forms are possible. For a variety of reasons, the multiple–choice item has been adopted by the Army for large–scale classification testing.

a. Description of Multiple–Choice Items. In performance tests, the items usually are problems involving cards, blocks, tools, equipment, etc. In paper–and–pencil tests, the items are generally multiple–choice questions or statements. Multiple–choice items present the examinee with several (usually four or five) answers to a question. The problem then is to indicate the correct answer. Examples—

(1) Boston is the capital of—

(a) Maine.

(b) Montana.

(c) Massachusetts.

(d) Minnesota.

(2) In the Diesel engine, the gas mixture in the cylinder is ignited by the—

(a) Spark.

(b) Heat generated by compression.

(c) Ignition system.

(d) Firing order of the cylinders.

b. Advantages of Multiple–Choice Items. Multiple–choice items are preferred for most personnel testing for several reasons. Scoring is more objective because the right answer is already set down and is not subject to the varying judgments of testing personnel. The examine has only to recognize the answer, and is not burdened by having to search for it in his mind and then phrase it in his own way. Because the examinee does not have to write but is merely required to check the correct answer, he can cover many multiple–choice items in a short time. Multiple–choice items can be scored by machine (as described in ch. 12), which increases accuracy and saves time.

22. Length of Test

a. General. To include all the items pertinent to a given trait would make the test absurdly and inefficiently long. To include too few items pertinent to a given trait would result in a test which would not measure the trait reliably. The principle which governs the number of items used in the test, and therefore the length of the test, is that there must be enough to show the degree to which each examinee possesses the trait, and to show this in a measurable fashion so that men can be compared with one another in terms of the trait.

b. Classification Testing Employs the Same Sampling Principles Followed in Other Fields of Measurement. In grading a carload of wheat, for example, it is not practicable to examine the whole lot in order to compute the percentage of high–quality grain, of chaff, and of foreign materials. The examiner instead gathers samples, analyzes them, and assumes that the characteristics of the whole carload are the same as those of the sample which he tested. But he would be extremely naive if he took all his samples from the top of the car. An unscrupulous vendor could easily have filled the car with an inferior grade of wheat and placed a thin layer of first class stock on top. The sample taken from this top layer would not be representative of the whole, and the measurement based on this sample would be an exceedingly inaccurate indication of the quality of the entire lot. Aware of all the pitfalls of careless sampling, and wishing his sample to be representative of the whole carload, the examiner collects a number of smaller samples—from the top and bottom of the car, from different depths, from each end, and from the middle. The more samples he collects, the more accurate his grading. The car may contain a concentration of inferior wheat, constituting a very small fraction of the total amount. If the examiner takes only five samples, and happens to take one of them from this small concentration, the inferior grade will comprise one–fifth of his total sample, and conclusions based on it will not apply to the whole carload. So it is with testing. Enough items should be used to sample adequately all important parts of the area which the test is to measure. Further discussion of sampling is presented in chapter 5.

c. Practical Considerations Limit the Length of the Test. In practice, the test–maker includes as many items as are necessary to make the test sufficiently accurate for sound personnel measurement, but keeps it short enough to be practical under field conditions and within the limits of fatigue and boredom of the average man.

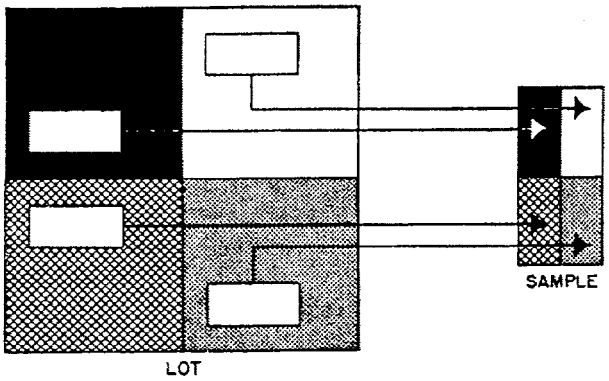


Figure 3. A good sample comes from all parts of the lot

23. Time Limits

Time limits are extremely important because they help to determine the qualities measured by a set of items. A test made up of items which are equal in difficulty measures speed if the time limit is so short that no one can finish all the items. A test in which the items get successively harder and harder measures power if examinees are given all the time they need to complete as many items as they possibly can. Most Army tests measure both power and speed. The Army must consider both how well and how fast a man can be expected to perform in a given assignment. The Officer Candidate Test is a good example. It is made up of questions and problems which become harder as the examinee forges through them. Only the man who is both able and quick can complete most of the 70 items in the 45 minutes allowed for the test.

24. Other Considerations in Planning Tests

The nature of the particular instrument may introduce special problems. The decision as to whether a test should measure speed or power may be important for achievement testing; it is not so important for such instruments as interviews and self–report forms, (chs. 8 and 9).

Section II
PREPARING THE TEST

25. General

After basic decisions have been made concerning the traits to be measured, the general form of the test and the type of test items, the next task is to prepare an experimental form of the test. Preparation of the experimental form includes the actual writing of the items and preparation of a scoring key and manual. This step is only the beginning. Only after a test has been subjected to field trial and a useful scale of measurement established can the test be considered as properly developed.

26. Construction of Test Items

Prior to the construction of items, it is necessary to study the performance of men in the training courses or duty assignments for which the test is to be used. This study may take the form of job analysis involving direct observation of the men. It may take the form of examination of training curricula, consultation with experienced men, or examination of official job descriptions. Such information is valuable in furnishing detailed ideas for the content of items. Then a large number of questions or problems, closely related to the trait to be measured, are assembled. When appropriate, the writer of the items consults a subject–matter specialist to aid in keeping the subject matter of the items on a sound basis. Each item is carefully checked to make sure that it is logically and clearly stated and that it seems likely to produce an answer that will indicate something about the trait which is being measured. Finally, all the items are reviewed by a group of experts who may recommend final changes in form, phrasing, or content. The items which remain are then collected into an experimental booklet. Directions for administration and scoring are prepared for use during an experimental try–out.

27. Preparing Directions

a. The Directions Which. Accompany Each Set of Test Items Constitute a Statement of the Conditions Under Which the Test Was "Calibrated" or Standardized. Great pains are taken to make these directions complete and clear. Only by following them can the standard conditions—the conditions under which the test was standardized—be repeated (ch. 11). Unless these same conditions prevail, a test gives results as unreliable and misleading as a thermometer reading taken when the patient has a mouth full of ice.

b. Two Sets of Direction Are Prepared for Each Test: One for the Examiner and One for the Examinee.

(1) Instructions and suggestions to the examiner are included in the manual that accompanies the test whenever it is administered. They indicate the general conditions under which the test should be given, list the materials required to give it, and state time limits for the parts and for the whole test. They also include introductory remarks that should precede and set the stage for the administration of the test, as well as recommended answers to questions that commonly arise during the testing session.

(2) The second set of directions provides the specific instructions to the examinees. These are printed as part of the test booklet itself to insure that instructions will be the same for every administration of the test. Their purpose is to make certain that each examinee understands just what he is expected to do. They touch upon such details as the advisability of guessing when not absolutely sure of the answer, the amount of time that will be allowed for the test, the relative importance of working for speed as against working for accuracy. And they give precise and detailed explanations, along with demonstration and practice items, of the correct manner of indicating answers. Since all of these directions are such a vital part of the test, they are prepared by experienced test–makers and subjected to independent check for completeness and clarity.

28. Preparation of Scoring Directions

a. Placing the Item and Answers in Proper Order. The final step in making an experimental model of a new test is to work out the proper technique for scoring. The items are first put in the order which it is believed will yield the best results. The position of the right answers is adjusted so that they fall in random order. That is, the right answers are so located from one item to the next that the examinee will not be able to "outguess" the test by discovering a particular pattern of right answers. A scoring key indicates the position of the right answer on the answer sheet. The correct answer is placed in the corresponding position in the test booklet. The incorrect choices or alternatives are so arranged as to eliminate any clues that the examinee might get from the order of the alternatives. A final check is made to be sure that the right answers shown on the key and the correct alternatives in the test questions correspond.

b. The Scoring Formula.

(1) *Correction for guessing.* In some types of tests guessing may give a man a higher rating than his abilities warrant. If an examinee selects one of four alternative answers to a multiple–choice item by pure guess, his chances of selecting the correct alternative are one in four. In a large number of such guesses, he is likely to be wrong three times for every time he guesses right. On a 100–item test, for example, he will usually obtain about 25 right choices by answering in this fashion. Since he answered one item right for every three he answered wrong, an estimate can be obtained of what his scores would be, if he had not guessed by subtracting from the number right one third of the number wrong.

(2) *Is correction for guessing useful?* The use of the correction formula described in detail in Chapter 12, is based on the logic of chance. In practice, however, guesses are seldom completely blind. An examinee may get an item right by knowing which alternative is correct or by knowing that the other three are wrong. Likewise, if he knows that two are wrong, he will have to guess only between the remaining two and will, therefore, stand a better chance of picking the correct one. Any error that results from the application of the correction formula will always be in favor of the examinee who uses such udicious "guessing". However, there will be other, more cautious, examinees who may know just as much but who will never put down an uncertain choice if they are to be "penalized for wrong answers" Guidance to the examinees as to the advisability of guessing is normally included in test directions.

Section III
TRYING OUT THE TEST

29. General

After planning and constructing the experimental model of the test, the next step is its trial run. The aims of the trial and analysis of the now test are, in general, the same as those for any other trial run. The pilot model of a now gun is tested to make sure that it will shoot accurately and consistently and according to its specifications. Similarly, the pilot model of a now test is given a trial to make certain that it will measure specified characteristics with sufficient accuracy to permit construction of an operational test that will be effective in classification of soldiers.

30. Field Trials and Analysis

The experimental form of a test contains many more items than will be used in the finished product, because, on the basis of the findings obtained in field studies, certain of these items will be rejected. In addition to making certain that

enough items are available to allow for dropping unsatisfactory ones, certain other considerations are important in field trials.

a. The Sample Used.

(1) Ideally, the men to whom the experimental test is administered should be representative of the men who will take the finally developed test. A test to be used with inductees should be tried out on a group of men representative of inductees, not on a group of men who have completed advanced schooling. A test to be used for selecting enlisted men for officer candidate school should be tried out on a group of enlisted men otherwise eligible, not on men already enrolled in officer candidate school. In brief, the experimental sample should truly represent the operational population from which selection is to be made.

(2) Unfortunately, this ideal can only be approximated and too often the extent to which the experimental sample deviates from the operational population is not known. For one thing, the operational population may vary markedly from time to time so that its characteristics are not stable. For another, the essential criterion data (ch. 3) needed to determine the effectiveness of the test may not be available for all the men who will take the operational test. For example, a test which is intended for use in selecting enlisted men for a school will obviously lack the school data for the men who are not accepted. Thus, in estimating the effectiveness of the test in predicting how well men will perform at the school, the estimate will contain a certain amount of error arising from the fact that the men in the school are only the better men—that is, they represent a range of ability restricted to the higher level.

(3) Various practical attempts are made to allow for the possible lack of representtiveness of the experimental samples. Statistical methods are available for correcting for restriction in range when such correction is justifiable. These methods are beyond the scope of this pamphlet and will not be described here. Another procedure is that of conducting follow–up studies of the effectiveness of the test on succeeding groups of men. This method has the advantage of taking into account possible variations in the range of abilities represented by the men from time to time.

b. Difficulty Index of Item. Another consideration in the field trial of an experimental test is the determination of the difficulty of each of the items. Difficulty is not determined by the subjective estimate of the test–maker, nor by the subject matter expert, nor by any opinion that the item "should be easy for anyone claiming to be familiar with the content of the test." The difficulty of a test item is determined from actual tryout; it is defined as the proportion of men in the group who actually answer the item correctly.

(1) *How item difficulty is expressed.* The estimate of item difficulty from the trial run, therefore, involves the task of counting, for each item, the number of examinees who answered it correctly and computing percentage of the total number of cases. Difficulty is thus expressed in percentage form; a difficulty of 70, for instance, means that 70 percent of the group answered the question correctly. This would be a relatively easy item. An item having a difficulty of 28, that is, an item answered correctly by only 28 percent of the group, is relatively hard. (It will be noticed that a low percentage means difficult and a high percentage means easy.)

(2) *How item difficulty is used in item selection.* In tests involved in personnel actions, there is little point in including a large number of very difficult or very easy items. If some of the items are so difficult that no one answers them correctly, then all that is accomplished is subtracting a constant from all the scores without affecting the ranking of the men on the test. Similarly, including easy items that every one answers correctly means merely that a constant is added to every man's score. Furthermore, too many difficult items may tend to discourage examinees; too many easy items may result in a loss of interest and carelessness.

c. Internal Consistency Index of the Items. The internal consistency index of an item refers to the relation it bears to other items in the test. If, for example, the men who answer a particular item correctly tend to answer most of the other items correctly, while those who answer it incorrectly also answer most of the other items incorrectly, then that item may be considered to be consistent with the other items. It will have a high internal consistency index. However, such consistency is no guarantee that the test is valid, i.e., that it agrees with the external criterion.

d. Validity of Items. There is one other characteristic of an item that it is necessary to know. Is it answered correctly by good men (high on the trait being measured) and incorrectly by poor men (low on the trait measured)? If half of the men who answer an item correctly are known to be competent and the other half are known to be incompetent, that item is useless in distinguishing the good from the poor. In a well constructed achievement test that is intended to determine, for example, what men have learned in a particular course, the validity of an item rests upon whether the content of that item was adequately covered in the course.

(1) However, when such a test is used to indicate how well the men will perform on the job after completion of the course, validity of items in the test must be established in terms of job success. Suppose that successful job performance requires not only knowledge of course content but ability to work with others, ability to adapt materials to the purpose at hand, ability to solve problems not covered in the course, and ability to lead. In such cases, it is necessary to discover if men who are good on the job answer the item correctly and men who are poor on the job answer it incorrectly. If this is the case, the item is said to have "high validity." If both competent and incompetent men on the job answer an item similarly, the item has low validity, and there is no point in including it in the instrument.

(2) The determination of the validity of items in such instruments as self–report forms is especially important since such items do not have correct or incorrect answers but rather indicate the degree to which a particular characteristic is

present. It is still necessary to determine whether men who are good in their job performance tend to select one answer and those who are poor select another. Where this is not true, the item is useless for discriminating between good and poor men.

31. Selection of Test Items

Field studies of the experimental form furnish data on the three main characteristics of each item—its validity, its difficulty, and its internal consistency. These data are used in selecting the best of all the items tried out.

a. Validity. The first and most important consideration is the validity of the item. The items are examined and those with little or no validity are eliminated, regardless of their other characteristics. There is little point to loading a test with items which have little to do with what the test is supposed to measure. Such items may be valid for other purposes, but if not valid for the purpose of this test, they add nothing to this test.

b. Difficulty. Once the items have been selected on the basis of their validity, they are examined to make certain that their range of difficulty is suited to the purpose of the test.

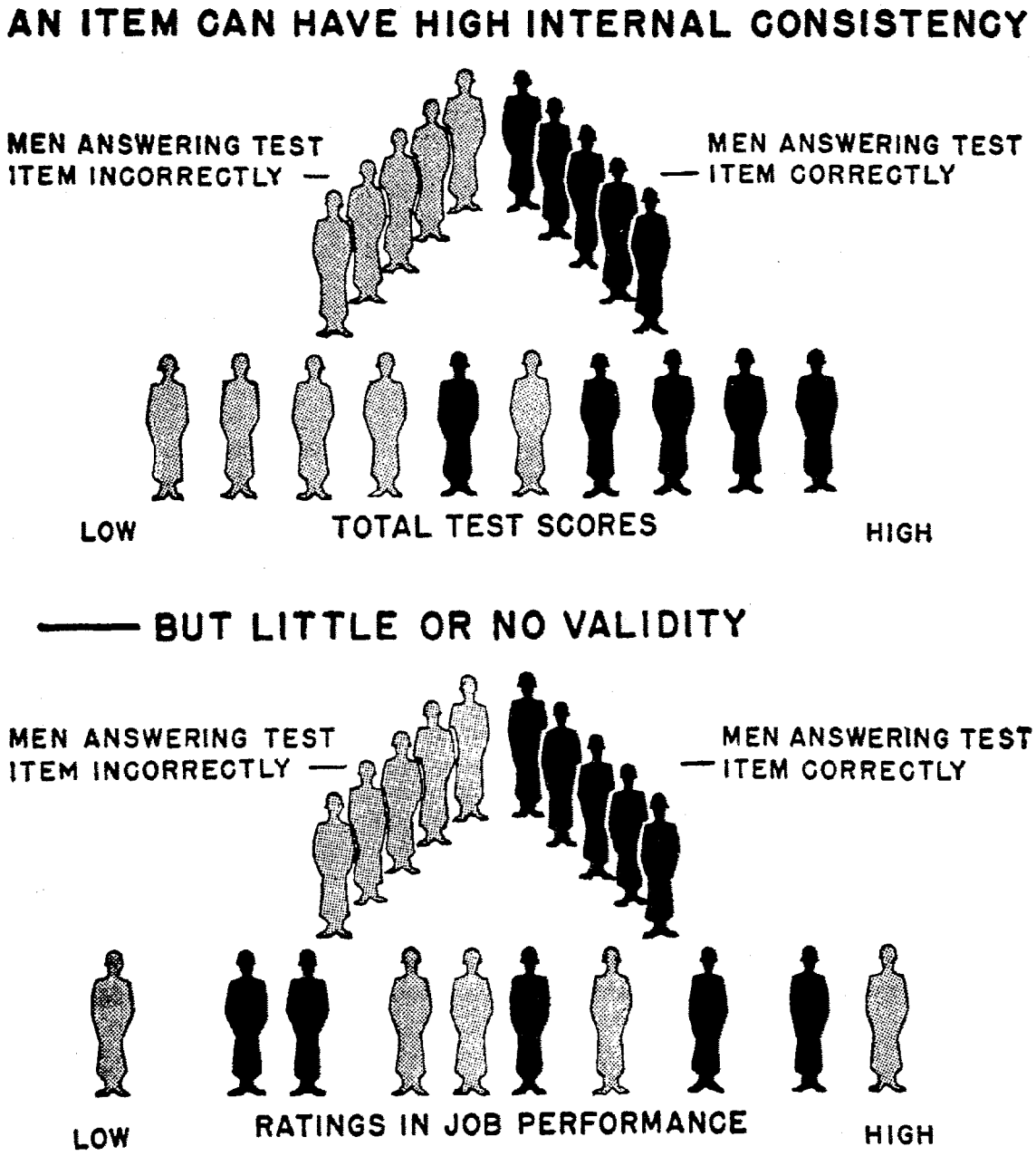


Figure 4. Internal consistency is not necessarily an index to the usefulness of the item

(1) If this purpose is to make the most efficient division of the whole population into a high and low group with reference to the trait being measured, the difficulties of the items selected should cluster around the division point, if no guessing is involved. More specifically, if it is desired to qualify the top 30 percent of a population for specialist training or assignment, then the difficulties of the items selected should cluster around 30 percent (items answered correctly by 30 percent of the population). If guessing is involved, the item difficulties should be somewhat higher than 30 percent.

(2) If, as is usually the case, it is desired to grade the whole population from highest to lowest with reference to a trait rather than merely to divide it into two groups, the difficulties of the selected items should be spread over most of this range.

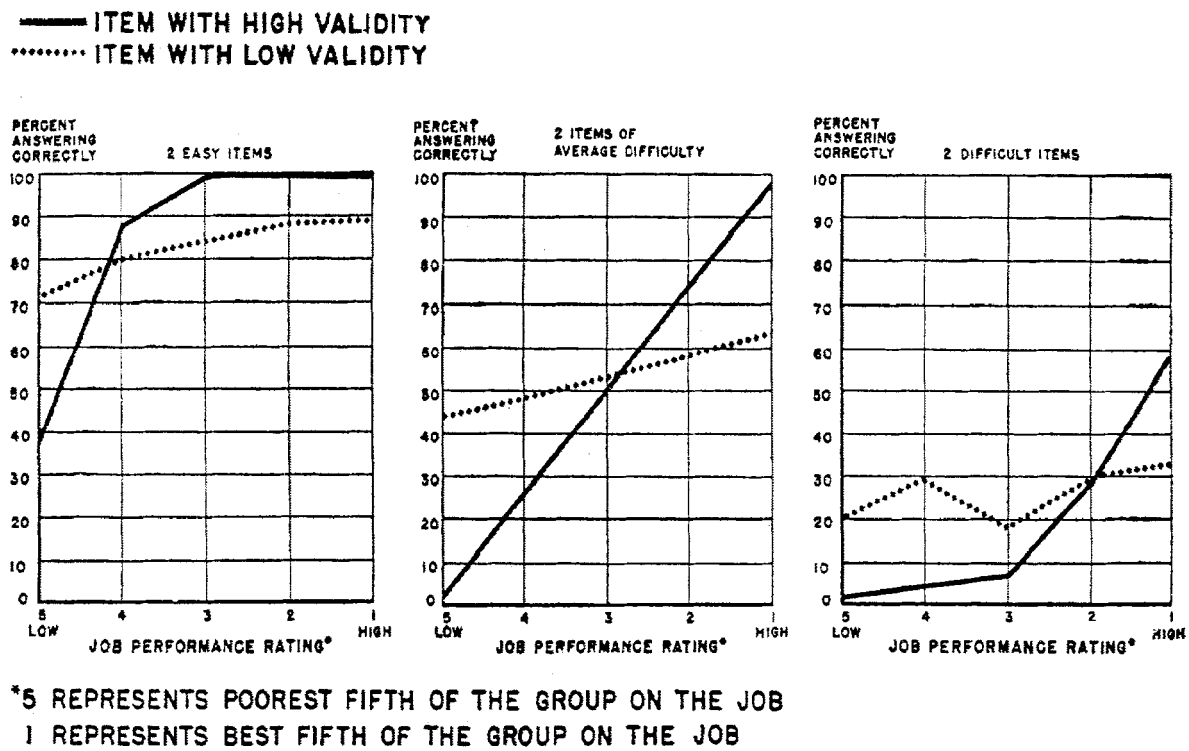


Figure 5. Illustrations of high validity and low validity for items at three levels of difficulty

(3) Selection of item difficulty distribution for optimum efficiency of a test in a specific situation is a complex problem involving a number of factors. A full discussion is beyond the scope of this pamphlet.

c. Internal Consistency Index. As stated in paragraph 30, this characteristic must be used with caution. If the test is supposed to measure a specific ability, such as ability to manipulate numbers, then the items which are selected should have high internal consistency indexes. However, if the test is supposed to cover a broad area of abilities, such as understanding verbal statements or ability to persevere in a disagreeable task, then high internal consistency indexes may be of little value.

32. Final Form of Test

When the best items for the purpose of the test have been selected on the basis of data from the trial run, the test is in its final form. It is then important to establish the validity and reliability of the finished product (see ch. 6 on validity, ch. 4 on reliability). The test is administered to another group of men, selected so that they will be representative as far as possible of the population for which the test was designed (see ch. 5 on sampling). The final directions, time limits, and scoring procedures are used at this time. Scores of men tested are compared with measures of job or training success, and the validity of the final test is estimated. Reliability is also computed from these data. This whole process of checking the usefulness of the final form of a test on another and similar group of men is called cross–validation. It is an important step in test construction, a step which aids in estimating the value of the test under operating conditions. Cross–validation of a test may be accomplished more than once, as the Army population for which the test was designed changes.

Section IV
ESTABLISHING THE SCALE OF MEASUREMENT

33. Purpose of Standardization

a. Upon completion of the item selection, the test is a finished product in the sense that it has been made as accurate and dependable as possible. But measurements made with it will still be in terms of “raw” scores, that is, the number

of questions answered correctly, or the number right minus a fraction of the number wrong, or, in the case of instruments which indicate the degree to which certain characteristics are possessed by the man, the total strength of all the characteristics covered in the instrument. By itself, a single raw score is seldom of much value to the classification officer who has to use it, regardless of how accurate and dependable the test may be. (See also pars. 65 through 69.)

b. A raw score does not tell the degree to which a man possesses a given skill or aptitude in comparison with other men in the Army and is, therefore, no clear indication that he will do better or worse than others on assignment. A raw score does not tell what proportion of Army men stand higher or lower in regard to the trait under consideration. For each test, therefore, it is necessary to know the scores of other men with which a mart is being compared and how they are distributed along the range from high to low. Personnel management officers have neither the time nor the facilities to collect this necessary data. Further, a time–saving and efficient technique for interpreting raw scores in terms of these data is required to make sound personnel measurement practicable in an Army of millions of men. The data concerning performances of other Army men are obtained by testing a standard reference population. The device for handy interpretation by the classification officer is the AMY standard score scale, developed to show what raw scores mean in terms of comparisons among men (pars. 70 through 78).

34. Standard Reference Population

a. The sample on which the experimental test is tried out is usually not the sample on which the final form of the new test is standardized. This standardization sample must be representative of the population that will take the test. It is neither feasible nor efficient to give each new test to the whole population of Army men concerned. Sufficiently dependable information can be obtained by using a carefully selected sample. Each new test is given in its final form to a large sample of men selected to represent as accurately as possible the whole Army population for which the test was designed. The size of the group varies considerably, depending upon the nature of the problem, the availability of groups, and the requirements of speed and economy. No practical advantage is gained by enlarging the sample at a high cost in time, energy, and personnel, since scientific control of its selection and the application of appropriate statistical techniques produce sufficiently dependable results. The representative group to which the final form of the test is given is the standard reference population. The administration of the test to this population, and the statistical computations which follow, is known as standardization.

b. Sometimes it is possible to standardize an instrument on the entire operational population, rather than on a sample, prior to adoption of the instrument for operating use. For example, no attempt is made to standardize the scoring of officer efficiency reports until all the reports for all officers for 1 year are available. This method eliminates the problem of adequate sampling since the total population is used. It also avoids the difficulties introduced by possible differences between experimental and operating conditions.

35. Minimum Qualifying Scores

The test score below which men may not be accepted for induction, duty assignment, or training course, is called a minimum qualifying score. Minimum qualifying scores for several different assignments may be set at appropriate points on the range of scores for a single test. The chances of success on a given job indicated by any obtained Army standard score can be computed. The minimum qualifying score is set at a point dictated by Army necessity. Thus, if it is desired that 80 percent of the men selected shall complete a course successfully, or perform satisfactorily in a given assignment, the minimum qualifying score could be set so that only men who stand a 4 to 1 chance of success, or better, will be selected. To select so high, however, may also mean that few will qualify. In establishing the minimum qualifying score, it is necessary to take into account the supply–demand ratio for the particular course or assignment in question. If the demand is small in relation to the supply, the minimum qualifying score can be set high and there, will be low probability of failure among those selected. Where the demand is large relative to the supply it will be necessary to lower the minimum qualifying score in order to qualify more men. When this is done, a higher percentage of failures must be expected among those selected.

Section V
INSTRUMENT BATTERIES

36. The Need for More Than One Instrument

The discussion thus far has been directed at the development of a single test. However, it is apparent that for many Army jobs the abilities required for successful performance are so complex that a single test, even a very good one, can hardly provide all the kinds of information needed to classify men. A battery of measurement instruments is needed to cover all the important abilities required.

37. Which Instruments Should Be Used

a. Before selecting or constructing measurement devices for a battery, a clear understanding is needed of the nature of the work for which selection is to be made. Consider an example, the enlisted job of Infantry Operations and Intelligence Specialist. The job duties in part include the following: “Prepares operations, situation, or rough topographic maps using field sketches, overlays, and other sources of information to indicate type, strength, and tactical

disposition of enemy and friendly units." To be successful in this job, the man must know theory and technique of infantry sound locating and principles of survey applicable to counterfire activity. He must also know duties of personnel comprising the infantry sound locating team. Much of this information can be tested by a written test devised for the purpose.

b. The Infantry Operations and Intelligence Specialist also must be able to supervise the sound locating team in installation and operation of counterfire equipment. He must actually demonstrate how to perform many of the tasks which are to be carried on. Tests of performance in operating a counterfire information center, for example, or of operating a sound recorder or computing and plotting counterfire data as a member of the sound locating team could be useful in determining how well he can do what he is supposed to do.

c. Not every man who is qualified to perform the technical aspects of this job can teach the job to someone else, nor have paper–and–pencil tests been developed as yet which give a useful measure of effectiveness as a teacher or supervisor. It has therefore been common practice to try to obtain some measure of effectiveness in teaching and supervising by having competent observers make ratings of past performance in such duties.

d. In addition, the Infantry Operations and Intelligence Specialist serves as a leader of an intelligence and reconnaissance platoon or squad in combat operations. To get estimates of this important aspect, other kinds of measuring devices are needed, such as an instrument which provides information on his personal history, schooling, interests, hobbies, etc. (self–report form). It would also be desirable to get some idea by direct observation of the man's typical effectiveness in dealing with people (standard interview).

e. At this point it can be seen that an adequate testing program for the job of Infantry Operations and Intelligence Specialist would appear to require—

(1) A paper–and–pencil test of some length, covering a variety of kinds of knowledge.

(2) A performance test or work samples.

(3) Ratings on past performance.

(4) A self–description form or standard interview, or both.

Note. Furthermore, each of these instruments must be valid for selecting men for this job; that is, the measures must be related to some independent measure or criterion of competence on this type of job. In addition, there should be a minimum of overlap in what each instrument measures.

f. The decisions reached thus far on what combinations of factors are important to success on a given job and how they should be measured have been based on judgment and logic. It is still necessary to determine whether these decisions are sound. Statistical techniques are available which will provide useful information on this point. A battery of a large number of instruments may give only slightly better prediction of job success than a reduced number of tests; the increased validity of the larger battery may not be of practical importance when weighed against the greater cost in time and convenience of using a longer battery.

g. This determination of the best combination of instruments for selection or classification of men is based on several considerations. Most important are the validity of each instrument and the degree to which each overlaps the others, that is, the correlations of each instrument with the other instruments (see ch. 5 for discussions of validity and of correlation). If two instruments are both valid but also extremely similar in what they measure, there is usually little point to using both in a battery. However, instruments which have some validity, but low correlation with each other, may make up a composite which is considerably more valid than any one of the instruments used separately. In such cases, a composite score may be obtained which represents performance on the battery as a whole rather than on each instrument separately.

h. Instruments in the battery may be given equal weight or differential weight. The weighting used is based on the results of statistical techniques. When the scores on the instruments are thus weighted, the composite score will yield the best prediction of success in job or training that can be obtained using the particular combination of tests making up the battery.

i. Selection of instruments for a battery and determination of the best weights are checked by administering the battery to a new group of men, representative of the group for which the battery was designed. This procedure, known as cross validation, has been discussed with regard to test construction in paragraph 32.

j. Other considerations involved in determining what tests should make up a battery before it is installed for operational use include the importance of the job, the number of men needed and available, time and expense.

38. Other Ways of Using a Number of Instruments

a. Instead of combining several instruments into a battery which yields a useful composite score, the instruments may be used separately in two ways—

(1) Minimum qualifying scores may be established for each instrument where a minimum level of competence is required for each of several aspects of a duty assignment or training course. This is not a very useful method since minimum qualifying scores are not fixed but should reflect variations in the supply–demand ratio. Furthermore, it is based on the assumption that a weak performance on one measure cannot be compensated for by very good performance on another.

(2) The instruments may also be used as successive screens. On the surface, this method is a very plausible one, but

in practice it has serious weaknesses. One is that the order in which the screens are used can seriously alter their effectiveness. Another is that separate minimum qualifying scores are needed, and, as already emphasized, these are not fixed. Other difficulties are present, not the least of which is the temptation to add more screens than is profitable.

b. Considering all the advantages and disadvantages, it is usually better to use a battery of instruments as a composite than to use the instruments independently or as successive screens.

c. In general, the basic procedures for constructing a battery of instruments are the same as those for a single instrument, which might be conceived as a battery of items. Just as the scoring of a single instrument is converted to a standard scale of measurement, so should the composite score be made meaningful.

39. Continuing Studies

When an instrument is released for field use, its usefulness remains a concern of the test–maker. Even if a test has been shown to have near perfect validity, it may still be necessary to check on it from time to time. The nature of a duty assignment or school course may be materially altered as a result of improvements in equipment and tactics. The nature of the Army population may change from a peacetime status to a mobilization status. Supply–demand ratios change. All these may affect the usefulness of a test. Follow–up studies need to be made from time to time to check the subsequent job performance of men who scored high and those who scored low on the test. Only through the study of accumulated evidence can the Army be sure that the test is doing what it is supposed to do, or ascertain when improved tests and procedures are necessary.

Section VI
SUMMARY

40. Steps in Instrument Construction

a. Study of the personnel measurement problem to determine the need for instruments, practical considerations influencing the structure of the instrument and the characteristics to be measured, and the specific way in which the scores will be used.

b. Design of the instrument in accordance with its purpose.

c. Construction of items.

d. Field trial of experimental form.

e. Analysis of results of field trial.

(1) Determination of validity, difficulty, and internal consistency of items.

(2) Preparation of scoring key.

f. Selection of items for final form.

(1) Use of indexes of validity, difficulty, internal consistency.

(2) Scoring based on selected items only.

g. Verification of scoring key.

(1) Application to a second group.

(2) Revision of items.

h. Computation of validity and reliability of revised instrument as a whole.

i. Preparation of operating form of test and procedures.

j. Establishment of scale of measurement.

(1) Standardization.

(2) Minimum qualifying scores as dependent upon supply–demand ratio.

k. Instrument batteries.

(1) Selection of instruments.

(2) Composite, independent, and successive screens.

l. Continuing verification studies when appropriate.

40B. Title not used.

Paragraph not used.

Chapter 3
CRITERIA

Section I
CRITERIA AND VALIDITY

41. Definition of Criteria

a. In chapter 2 it was shown how a newly prepared test is "tried out" to see how effective it is in differentiating between competent and incompetent men in a given occupation, or between men who are "good" on some characteristic and those who are "poor." But how is it decided which men shall be considered competent and which incompetent? What does it mean to be "good" or "poor" on some trait? It is clear that there has to be some standard to go by.

b. The measure of competence or of "goodness" used in making these distinctions is the criterion measure. To say that there is another measure by which the "test is tested" gives rise to a whole series of questions: How is this criterion decided upon? How can a measure of this criterion be obtained? How is it known whether or not it is a satisfactory measure? What is the essential difference between test and criterion? These and other problems will be discussed in detail in the following paragraphs.

42. Part the Criterion Plays in Validating a Test

a. The Criterion as Representing the Purpose of the Test. Determining the validity of Army personnel measurement instruments boils down to two basic problems, one somewhat theoretical, the other extremely practical. The theoretical problem is whether the test actually measures the trait or ability it was devised to measure. The chief limitation in the use of this approach is that even if it can be established that a certain trait is measured by the test in question, it still has to be established that the trait as measured is important to success on the job. The name of a trait or characteristic, in fact, means little. A man who is honest in money matters may not have the same scruples about taking advantage of his fellows when all are in line for promotion. The more practical question is: Is what the test measures, regardless of what it is called, important in performance of a given type of job? What must be ascertained is whether the test can provide a reasonably accurate means of estimating how good a soldier a man will be in the spot for which he is being tested. Only a measure of his performance on the job or in training can demonstrate whether or not the test is contributing as it was intended to do to the effectiveness of the selection process.

b. Criteria as a Check on Theory. The practical approach, too, has its limitations. Experience in research on the development of tests has shown that even after the most careful analysis of a job it may not be clear just what is required for success on that job. It is here that a criterion is most important. It is by checking theory about what is important to such success against a criterion of success that theory can be verified or errors in reasoning detected.

43. How Criteria Contribute to Classification

a. Each man could, it is true, be placed in a job or assigned to training in a specialty for which, by surface evidence, be seemed suited. If after 6 months or so he was doing satisfactory work or making satisfactory progress in training, his assignment could be made permanent. If not, he could be reclassified to some other occupation. That would amount to using the criterion itself as a basis for classification. What is wrong with that?

b. In the first place, it would be a time–wasting process. A man who was misclassified would have lost 6 months or more of training time. More important, nothing would have been learned about the man's ability to do other jobs, nor would anything have been found out about other men who might perform much better in that particular job or training assignment. Even in the case of those who met the minimum performance standards for the job, it is not safe to assume that they have been properly classified. They might better have been assigned to some other occupation.

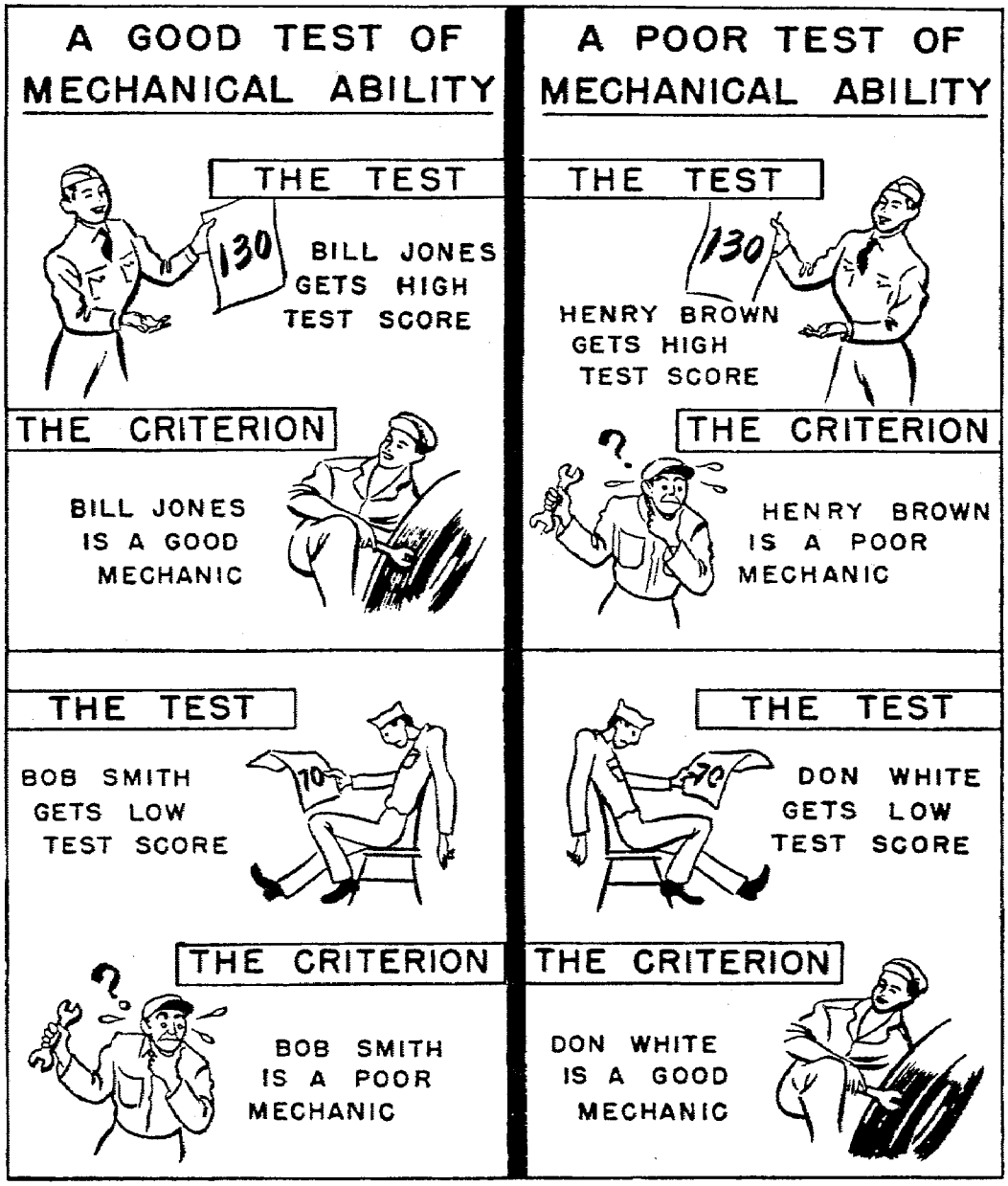


Figure 6. Criterion tests a test

c. Thus, the use of tests instead of criteria not only saves time and expense but provides a means by which the whole classification procedure in the Army can be integrated. It is only by having these advance measures of probable success for each man that the process of matching manpower and Army needs can go on efficiently. On the other hand, these advance measures would be impossible to obtain without the use of criteria.

44. How the Criterion Measure Is Used

a. To illustrate the part the criterion plays in the development of a personnel measuring instrument, consider the problem of finding out how effectively a set of tests will pick men who will do well as auto mechanics after an appropriate training course. The experimental tests might include a mechanical aptitude test, a test of automotive

information, a test of ability in visualizing how patterns fit together to form objects, and a test of how quickly a man can assemble some mechanical device such as a bicycle pump. These tests would be given to a group of men who had undergone training for the job of auto mechanic and were now assigned in that military occupational specialty.

b. At the same time, criterion measures would be obtained for each man in the group. Suitable criteria for this job might be performance ratings for each man by both his superiors and men working with him. How this is determined is discussed in paragraph 62. It could then be determined how closely scores on each of the tests correspond to the criterion scores, or, more accurately, whether the ranking of the men on the tests is the same as their ranking in ratings of performance on the job. This comparison of rankings is known as correlation, and the computation of the degree of correspondence between scores on a test and scores on a criterion gives a measure of the validity of the test (ch. 5). If the correspondence is close, that is, if men are in pretty much the same order on both measures, the test is said to be valid for this particular job. When the test, thus tried out against a criterion and found to be sufficiently valid, is given to groups of unselected men, there is considerable likelihood that the men who score high on the test will give good performance as auto mechanics after training, and the men scoring low on the test will perform poorly. The test is then said to be a good "predictor" of success as an auto mechanic. The degree of correspondence between the criterion and each test may be found in this way; or, as explained in chapter 2, the tests may be treated as a battery and a composite score on the tests, with each test weighted according to its validity and degree of relationship to the other tests, may be used.

45. Importance of the Criterion

It is necessary to determine at this point whether a certain test will be used for a particular purpose or which of several tests will be retained in a battery. Since this decision is made chiefly on the basis of the relative validity with which the tests predict the criterion, it is in effect the criterion which determines which tests will be used and which discarded. If the criterion measure is faulty, if it is not truly related to the essentials of job success, the wrong tests may be used. The men selected by these tests will have been selected on the wrong basis, and little will have been accomplished toward improving the method of selection for the job. The criterion, then, has a decisive effect upon the outcome of a selection procedure.

Section II
CHARACTERISTICS OF ADEQUATE CRITERIA

46. What Makes a Satisfactory Criterion

a. Since the criterion plays such a decisive role, it is essential that it be adequate and that it be adequately measured. If the criterion is to be adequate, it must be tied to the purpose of the job. It is necessary to determine what is required for success on the job and to determine how to measure the desired requirements for success in that job. For example, if it is desired that recruiters obtain as large a number of enlistments as possible regardless of the quality of the enlistees, the appropriate criterion measure is a production record. If it is desired that the recruiter enlist only high quality men, and not mere numbers, a production record is not a satisfactory criterion measure—the criterion must emphasize the recruiter's ability to size up men. Similarly, if a rifleman is required to be primarily an expert in marksmanship, a production record—target scores—might be the appropriate criterion measure. If he is required to train recruits in handling a weapon, ability to teach marksmanship would be the appropriate criterion measure.

b. Another aspect of the problem of determining whether a criterion is adequate is the relation of the job to the ultimate mission of the Army. Thus, the criterion of success of an infantryman in combat is performance in combat. If there is no combat, obviously no measures of the ultimate criterion can be obtained. It may be necessary to seek as an adequate criterion some training situation that approximates some of the stresses of actual combat—its urgency, fatigue, fear. As yet, no adequate substitute has been found.

c. Characteristics important in one job may not be important in other jobs. Military bearing may be important for an infantryman on garrison duty but may be of minor significance for an infantryman in combat.

d. It is apparent that considerable attention must be directed at studying possible criterion measures before it is decided which of them to accept (par. 62). Such studies are intended to discover which possible measures possess the necessary characteristics for use in criterion measures. These characteristics will be considered next in greater detail.

47. Comprehensiveness and Weighting

a. The criterion measures should cover in some way the important aspects of the performance, whether it be in a training course or on the job. For convenience, the discussion will refer only to job performance; it should be remembered that the discussion applies to training performance as well.

b. The important elements of the job must be analyzed by careful study. This can be done by interviewing men on the job and those who supervise them, by observing men at work on the job, or, on occasion, by actually performing the job long enough to discover its more subtle aspects. This job analysis must uncover not only the particular operations performed but also the personal relations involved. What is even more important than identifying the job elements is to determine their relative importance so that the more important ones may be weighted more in the criterion.

c. On many jobs, more is involved in success than a breakdown into individual job elements will show. This is most apparent in jobs where the mission must be accomplished by working through others. The only criterion measure that will suffice is one which in some way takes account of how skill in dealing with other people influences all other elements of success in such a job.

d. Looked at from the point of view of validating a particular instrument, it may turn out that the instrument is closely related to one criterion measure but not to another. This condition would not necessarily mean that the instrument is of no value. It might well mean that the criterion measure which is not correlated with the predictor instrument is an inappropriate criterion.

48. Freedom From Bias

a. The criterion data should also be free from bias, that is, free from the influences of any factors, other than chance, that are not pertinent to on–the–job success. A man's score on the criterion may be influenced by factors which have little or nothing to do with how well he does his job. For example, he may have produced less work than another man, not because he cannot, or will not produce more, but because he is working with poorer equipment, or because he has to make up for someone else's mistakes, or because he gets less to do than the others.

b. Criterion bias often occurs when a rater knows how the man he is rating scored on the test that is being tried out. Knowing that the man scored high on the test, he may consciously or unconsciously rate him high on the criterion, regardless of his actual performance. Bias of this nature will make a test seem more valid than it really is.

c. Again, an officer who is rating his men for criterion purposes may know that a man makes a good appearance and talks well, and therefore have the feeling that he does everything well. He may judge a man's skill in close order drill and infer from this that the man is doing well in squad tactics, instead of observing him in squad tactics and rating him on those duties. This type of bias is known as "halo."

49. Consistency of the Criterion Measure

One general requirement of a criterion measure is high reliability. Criterion measures should be consistent from one time period to another. A rating obtained in October should not be too different from one obtained for the same man in May, if neither the job situation nor the man has changed drastically. Nor should there be too great a divergence between the criterion ratings when a man is rated by different raters within the same time period. The standards of consistency are considerably more relaxed for ratings than for test scores, but even here if there is not a reasonable consistency the measure is worthless as a criterion. Obviously, a characteristic which varies markedly cannot be measured consistently.

50. How Criteria Are Selected

a. In any validation study, the criterion measure used is the one which is expected to prove most useful and practical. In making the selection, the type of instrument to be validated is considered and the purpose to which it is to be put after it has been tried out. Is it intended to screen out only the poorest men? Is it to select only a few best men for promotion? How specific is it to one particular type of job? Another consideration is the importance of this purpose and how much time and expense can be justified in obtaining—or developing—what is considered an adequate criterion measure. Timing, too, is important. How long can be spent in investigating the various possible criterion measures? In view of all these aspects of the problem, which of the available or obtainable measures related to job success will be the most satisfactory?

b. Some measure of job success may stand out as the most accessible and as bearing an obvious relationship to job success. To accept it as a criterion without investigation, however, may lead to attempts to validate an instrument against a misleading standard. In developing tests to select driver trainees and Army drivers it was agreed that the criterion was safe driving. It would have been easy to accept one of two criterion measures which seemed clearly related to motor vehicle safety and, at the same time, fairly easy to obtain–accident records and road test scores. Both of these, however, had turned out to be open to objections. Accident records, the most obvious and easily obtainable measure, were dependent upon too many factors other than the safe driving practices of the motor vehicle operator. Road tests, another seemingly logical device, failed to sample a driver's behavior in potentially dangerous situations out from under the surveillance of the examiner. Both measures were unstable from period to period. In view of the shortcomings of these measures, a special study was set up to develop an adequate criterion measure in the form of ratings by the drivers' supervisors and associates. In developing the rating procedure special attention was given to narrowing the content of the rating to observable–that is, ratable–behavior important in safe driving. The result was a more reliable evaluation of safe driving against which driving proficiency or aptitude tests may, in the future, be validated. This illustration points out the importance of conducting criterion studies before selection of the appropriate measures is made (par. 62).

Section III
KINDS OF CRITERION MEASURES

51. Production Records

a. Examples of production records that may serve as criterion measures are quantity of work produced, quality of product, number of errors. Such records are attractive as criteria because of their relative objectivity and their very apparent relationship to on–the–job proficiency. Only a careful examination reveals that they may not be as objective as they seem. And while they usually do measure some element of job success, it is not always the most important one. Amount produced depends in large measure on opportunity to produce. Quality in workmanship depends, in part, upon the condition of the materials and tools with which the work is done. These are matters over which the individual has no control but which may bias his record. Production criteria have their main usefulness in situations where opportunity to produce may be fairly uniform.

b. It is sometimes possible to keep variations in assignment and working conditions under control or to allow for them in some way in obtaining the criterion measure. Such a measure then gives a partial criterion of success on the job, and it may be a significant part. With a typist, the most important consideration may be how fast and how accurately he can type. With more complicated and responsible positions, such a measure would represent at best only the most routine parts of the work. In practice, production records are commonly used in combination with other criterion measures.

52. Cost Records

Cost records may provide the basis for a more useful form of production criterion. However, cost records must be translated into what may be termed a "dollar criterion," a measure of the individual's net worth to the organization. Essentially the dollar criterion is based upon what the organization must spend on a man in order for him to achieve a certain level of production. This kind of criterion is broad enough to include such factors as supervisor's time spent in training, absence from the job, and the cost of rectifying errors made by the individual.

53. Rank or Grade

a. Sometimes status or position is considered for use as a criterion. A man who holds the grade of sergeant may be thought of as representing a higher level of competence than a corporal. Stated more generally, grade, echelon, and other indicators of official status may be considered as representing levels of responsibility and these, in turn, as representing levels of competence. A criterion measure of this sort might be obtained on a group of men of various grades whose test scores would be compared with their grades to see if the men with the high test scores are the men who hold the high grades.

b. If such a criterion measure is used in a follow–up study—that is, a study to determine whether the personnel instrument used to select the men is really predicting their relative success on the job—certain precautions are necessary. The predictor measures must have been taken before the men start in on their assignments. Results on the selection test should not be made known to those responsible for assignment and promotion, or their promotions may partially reflect this knowledge. In that case, status would be a poor criterion.

c. The difficulty with this type of criterion measure is its assumption that status is determined primarily by the competence of the individual. Sometimes this is true; sometimes it is not true. The whole question of equality of opportunity, of equal responsibility associated with equal rank, enters in. Before status or position is accepted as a criterion measure, the basis on which it rests must be carefully examined.

d. Grade or rank may not offer sufficient spread to be a useful criterion measure. At times, promotion from private to private first class is practically automatic. In such instances, promotion would be of little use in differentiating among men of the lower grades.

54. Course Grades

Grades in Army service school courses may be appropriate criterion measures for instruments designed to pick men who are likely to succeed in acquiring certain knowledge's and skills. Such grades usually belong in the class of subjective measures, being based in part on the instructor's judgment of how well the man has done in the course. A number of schools have adopted grading systems based on an objective test or a series of such tests given at various training stages. Even when objectively arrived at, however, course grades in specific subjects are of doubtful value as criteria. They usually fail to take into account many elements of job success such as performance under unfavorable as well as favorable conditions, or the application of knowledge to an unfamiliar problem. On the other hand, successful completion of the training course is a prerequisite to success in job performance. School performance may then be thought of as another predictor of job success and, as such, has validity as a criterion. The purpose of the school course and the basis on which course grades are determined usually give a clue as to the value of a given set of grades as criterion measures.

55. Job Performance Measures

Attempts to use criterion measures as close as possible to what a man actually does on the job has led to the use of

performance tests measuring an individual's ability to perform a more or less complex task required in a specific job or type of job. The task is performed under standard conditions, and observations are recorded and evaluated. Evaluation may be objective, consisting simply of counting the steps performed correctly; or it may be more or less subjective—a judgment or rating of the quality of performance or of product. Recent research using performance tests as criteria has shifted toward measures in which rating has little or no part. Emphasis is rather on recording the facts of performance in such a way that they can be scored according to a pre–established formula. The suitability of such an approach to criterion measurement depends upon the relevance of the behavior tested to the job. Does the task constitute a substantial component of the job? Is it a crucial element? Are there other crucial job elements to be considered?

a. Automatically recorded performance. Only when a measure represents performance that is automatically recorded and scored by a set formula is the judgment element entirely eliminated. In studies of monitor performance, signals of various kinds and responses to signals can be recorded by electronic apparatus which leaves a cumulative record of correct and incorrect responses. Image interpreters may be given a set period of time to identify all military targets of a designated type to be found in an aerial photograph. When the results are compared with a pre–established key to the photograph, measures of speed, accuracy, and completeness may be obtained. In a test of this kind specific tasks can be simulated under laboratory conditions. It may even be desirable to try to simulate important environmental factors which normally are not reproducible in the typical laboratory. Each man is required to perform exactly the same task under exactly the same conditions. The important consideration is whether the performance measured represents adequately the total job performance or whether it can be combined with other measures to provide an evaluation of total job performance. In some jobs, the element subject to objective measurement may not be important enough to justify the time and expense required to develop the measuring device and the essential controls over the test situation.

b. Observer–recorded performance. In performance measures requiring the services of an observer, the element of judgment is minimized by precise definition of the task and the sequence of actions by which performance will be scored. The observer notes the examinee's behavior—did he or did he not set the safety catch? Did he pull the lever at the required time? Checklists specifying the items of behavior to be observed—and sometimes also possible actions, correct and incorrect, examinees may be expected to take—provide, the basis for systematically recording what the examinee is seen to do. The notations are then scored in the manner of an answer sheet to a test. In this type of performance test, the observer is not asked to judge or evaluate—merely to observe carefully according to instruction. Very difficult to control, however, is the accuracy of the observer's notations. Test procedures are usually subjected to repeated tryout and revision in an effort to minimize discrepancy between the examinee's action and the record of action.

c. Observer–evaluated and –recorded performance. When the observer is asked not only to note what he sees the examinee do but also to evaluate the quality of performance or of the results of performance, the subjective element in a performance measure may be considerable. However carefully the observer is instructed in the standards of performance to apply, his evaluation is essentially a rating, and shares the possibilities of bias inherent in a rating. Problems of obtaining reliable and accurate ratings is thus an important consideration in performance testing for criterion purposes.

56. Ratings

a. Ratings are familiar as devices for evaluating men as a basis for appropriate administrative actions. Ratings are used as criterion measures because they are frequently the most nearly adequate measures available for this purpose. The rater in appraising a man's performance will, in general, consider it in the light of the responsibilities and opportunities that are a part of the job. Thus the rater indirectly takes into account conditions which can make each job somewhat different from every other job.

b. A rating can be an overall measure. Frequently that is the kind of criterion that is necessary. For example, personnel actions involving infantrymen are seldom concerned with competence in some particular element of the job. An infantryman is not considered competent just because he is a good marksman, or a good map reader, or good at rendering first aid. There are men who are good marksmen but who are, nevertheless, not good infantrymen. To obtain criterion measures of competence as infantrymen—an overall measure—the only technique available at present is the rating technique.

c. A rating involves the personal reactions of the rater to the ratee. Some of this is undesirable and represents a source of bias. Yet when job performance involves the need to work with and through other people, it is important to have a measure involving personal reactions.

Section IV
RATINGS AS CRITERION MEASURES

57. Ratings Are Recorded Judgments

The term "rating" is used in a general sense to mean all methods of expressing judgments on the adequacy of a performance. All ratings are essentially recorded judgments. The familiar rating scale may be considered as an attempt to improve ratings by greater systematization of the method of recording the judgments.

58. Description of Rating Scales

a. In its simplest form, the rating scale is a series of steps representing levels of performance or degrees of a characteristic possessed by the ratee. The graphic rating scale is probably the most serviceable means of obtaining a rating. Indeed, the "cross–on–a–line" method has been so widely adopted that the term, "graphic rating scale" has been generalized to include all rating scales (except checklists) whether or not they are truly graphic. Such a scale consists of a line divided into intervals representing varying amounts of the characteristic on which the rating is made. Descriptive phrases may be placed along the line from one extreme to the other to define the points of the scale. A variation in this procedure is the man–to–man rating scale. Raters are asked to choose specific persons of their acquaintance as typical of each of the scale positions. The rater thus has a standard—set by himself, it is true—and can place the ratees by comparing them with the persons he has selected to represent the various points on the scale.

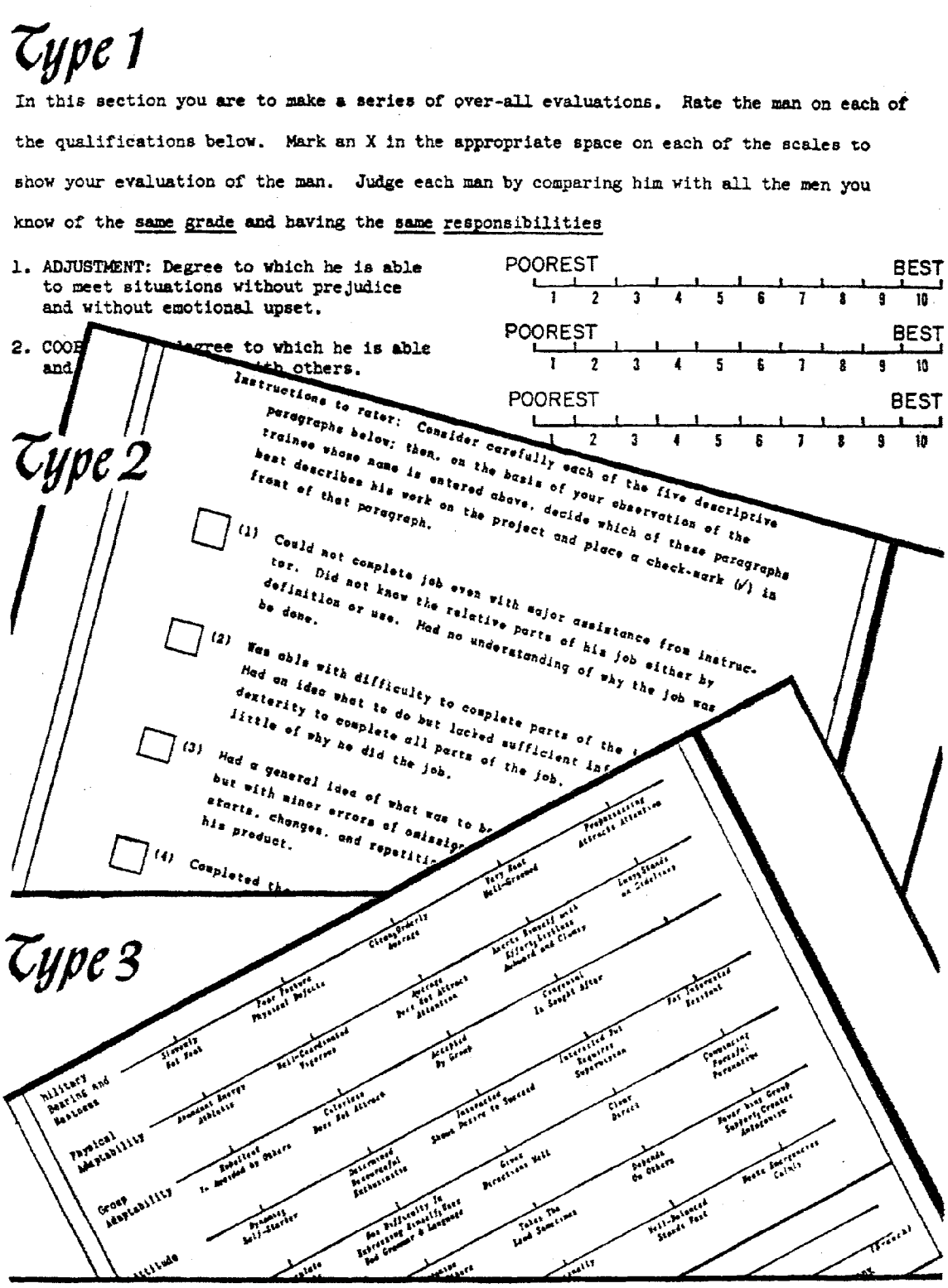


Figure 7. Examples of three types of rating scales

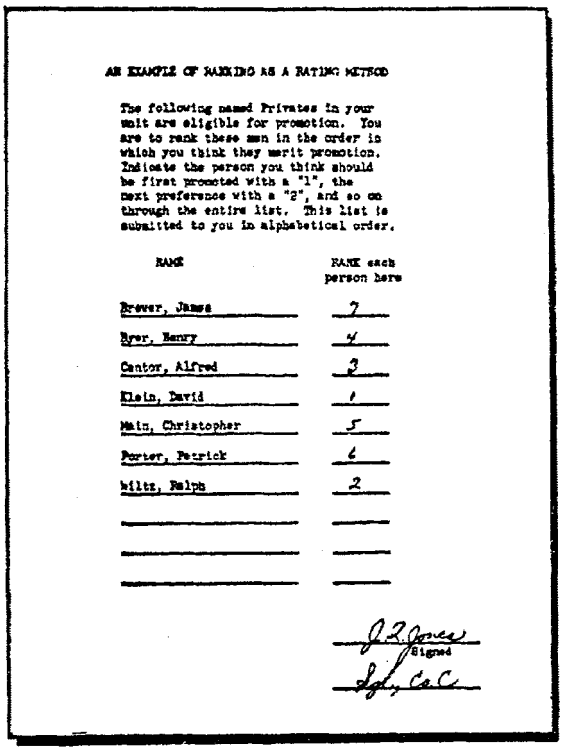
b. There may be one scale or many. With several scales it must be determined how much each shall count in the composite rating. The final numerical score may be converted to a standard score by the method explained in chapter 4.

59. Limitations of Rating Scales

Rating scales have become the traditional method of rating. Nevertheless, they possess certain serious defects which cannot be overcome simply by preparing a new form. The basis for these defects lies less in the form than in the rater. In the study of the rater, four general human frailties affecting ratings have been well established—tendency to rate on general impression (halo); tendency to keep all ratings close together on the scale (lack of dispersion); tendency toward leniency (rating high); and differences in raters' standards (differences in frame of reference). The combined effects of these tendencies for the traditional type of scale is to bring into operation quickly the law of diminishing returns in attempts to lengthen rating scales; to reduce the spread of rating scores, often to the point of making them useless for any practical purpose; and to obtain scores which are often more a reflection of difference in raters than in ratees. These limitations of ratings are discussed in more detail in chapter 10.

60. Other Methods—Ranking, Nomination

a. Ranking—Order of Merit. Ranking consists of placing men in order of merit. Each man is assigned a number, the one standing highest in the group is assigned number 1, the next, number 2, and so on through the group. The last number assigned is equal to the size of the group. When a number of raters evaluate each man, as is usual in criterion rating, an average is computed for all the ranks assigned the ratee by all his raters. In obtaining criterion measures, efforts are usually made to ask all raters to rank approximately the same number of ratees, thus avoiding having one rater rank only 2 or 3 men and another 15 or 20 men. The rankings may be converted to rank in 100 or rank in 1,000, or to a standard score. The standard score is the more useful conversion for criterion purposes.



AN EXAMPLE OF RANKING AS A RATING METHOD

The following named Privates in your unit are eligible for promotion. You are to rank these men in the order in which you think they merit promotion. Indicate the person you think should be first promoted with a "1", the next preference with a "2", and so on through the entire list. This list is submitted to you in alphabetical order.

NAME	RANK each person here
Brewer, James	7
Ryer, Henry	4
Cantor, Alfred	3
Klein, David	1
Main, Christopher	5
Porter, Patrick	6
Wiltz, Ralph	2

J. Z. Jones
Signed
Lt, Co C

Figure 8. An example of ranking as a rating method

b. Nomination—A Variant of Ranking. Instead of ranking all men in order of merit, criterion raters are sometimes asked to name a certain number as best and an equal number as poorest. The score is the number of times the ratee was nominated as best and as poorest. A similar method, the "sociometric method," requires each rater to name the man he would want most for a particular duty and the men he would want least.

61. Improving Criterion Ratings

The limitations described in paragraph 59 have one net effect; they introduce systematic error into the measurement. Such errors are said to bias the ratings. To reduce bias in criterion ratings several methods are available.

a. Averaging. When a number of measurements of an object or a person are available, it is usual to average them. The average provides one convenient number to represent a variety of numbers. In addition to being a convenience, the average has another valuable property as applied to ratings—the biases resulting from the peculiarities of the individual rater can be reduced and a more stable measure obtained. In addition to partial canceling out of the effect of individual rater bias, the use of an average of a number of ratings improves the basis of the rating by combining what a number of raters know about the ratee. A better estimate may thus be obtained of his effectiveness. Of course, increasing the number of raters, will be to no purpose if they are not competent raters of the ratee.

b. Guided Rating. Attempts to reduce bias in ratings may go so far as to provide actual guidance in making the ratings. An expert in rating methods works directly with the rater, or with groups of raters. Together, they reach an understanding of what qualities are to be rated and on what basis judgments are to be formed. It is the role of the expert to help the rater think back over his observations of the man he is rating and weigh them as indicators of typical action or performance. His guidance is intended to help the raters avoid as much as possible the weaknesses which usually detract from the value of ratings. This method offers a practical means of obtaining improved ratings for criterion purposes.

c. Special Instruction. In obtaining criterion ratings, special instructions are generally supplied to the raters, emphasizing the basic principles of accurate rating. Usually these instructions are given orally by a technician who also attempts to clarify the instructions, if necessary, and to remind raters of the thoughtfulness and care required. This practice is similar to guided ratings except that it is applied to groups instead of to individuals. In general, such instructions have proved helpful but by no means do they eliminate the various sources of bias in criterion ratings.

d. Selection of Raters. A rater, to be competent, must know what is required of the ratee and how well the ratee has met these requirements. Thus it may be specified that the criterion rater must have a certain minimum length of service and certain minimum length of work contact with the ratee. Command relationships are usually ignored and any rater who knows the job and the ratee's performance on the job is eligible. Superiors, equals, and subordinates may be used, if they meet the minimum requirements. In one study it was found that hard raters and easy raters rendered equally valid ratings—the raters placed their men at different parts of the scale, but they placed the ratees in much the same order. Another study indicated that ratings by enlisted men with general mental ability scores below an Army standard score of 90 were not as valid as ratings by men with higher scorer, even though special pains were taken with the low scoring men. The possible use of such characteristics in selecting raters to give administrative ratings is limited.

62. Criterion Studies

a. General. The critical importance of criteria in determining the usefulness of personnel measuring instruments and their effect on the management of military personnel makes it essential that the measures used as criteria be properly selected. This point has been emphasized in this chapter several times. To select the proper measure is not always easy, and considerable research effort may be needed.

(1) Sometimes this effort is directed at discovering whether existing measures are suitable for use as criteria. Too often, an existing measure seems suitable only until a more critical study is made of it. For instance, ratings by associates on overall value may be available and would seem usable as criteria. However, investigation may reveal that these ratings are used for administrative purposes, that they are not confidential, that raters did not know enough about some of the ratees to rate them accurately, that raters made little effort to spread their ratings, and so on. What appeared to be a suitable criterion measure may prove not to be so. Consider another example—leadership ratings may be available at a school and would appear at first glance to be suitable for criterion purposes. However, suppose it turns out that the raters were classroom instructors and that leadership ratings were highly correlated with academic grades. As criteria for validating academic aptitude tests, the leadership ratings might be suitable, but then the question would arise, are the ratings leadership ratings? As criteria for validating instruments intended to measure leadership potential, the leadership ratings could not be used unless it were known that academic performance and leadership were closely tied together, which is frequently not the case. In brief, then, considerable research may be required to determine if existing measures are appropriate for use as criteria, and, if not, to develop suitable criterion measures.

(2) Criterion measures may themselves be used to predict performance later on. For example, research may indicate that success in a particular service school is correlated with later success on the job. The measure used to evaluate

success in the school may be the one used to determine the validity of the instruments used to select applicants for the school.

(3) Criterion research may yield an important by–product. The criterion measures may have been developed to determine the validity of certain personnel measuring instruments. They may also be used as yardsticks in other areas of personnel management. For example, they may be used to evaluate training outcomes, the effect of assignment policies on performance of the men, and the effect of morale conditions on performance.

b. Relation of Criterion Measures to Purpose. It has been pointed out that the criterion measure selected must be related to the purpose of the instrument to be validated. Sometimes this purpose can be established as a result of the knowledge of the job requirements and advice of job experts. For example, assume it has been determined that an infantry soldier should have the ability to orient himself promptly behind the lines. A "locations" paper–and–pencil test might measure this ability. The criterion measure could consist of evaluations of the man's demonstrated performance in leading a patrol through the wilderness to determine how well the locations test measures ability to orient oneself behind the lines. But suppose it is desired to use the locations test not only to indicate how well a soldier can lead a patrol but rather how well he can perform as an all–around combat soldier. Another kind of criterion measure might then be called for, one which is more comprehensive. Further study may be required to determine the scope of the criterion measure and the method to be employed. Sometimes the use of a criterion is determined by policy. For example, instruments used to select officers might well be validated against overall criterion measures if the heterogeneous nature of officer requirements is emphasized. If ability to perform a technical specialty is required, on the other hand, the criterion of overall officer performance may be inappropriate. In general, then, both the purpose of the test and the purpose of the job must be understood in order to select the appropriate criterion measures.

c. Relations Among Criterion Measures. Sometimes a large number of possible criterion measures are available and it is desired to study the relationships among them as an aid in determining how they should be used. Measures which are highly correlated with each other may be examined to see which may be discarded to avoid unnecessary duplications. Those which are not correlated with each other are examined for suitability. They may be uncorrelated with other measures because they are poor measures, in which case they are dropped from further consideration. However, they may be uncorrelated with the others because they cover an aspect of performance not covered by the other measures. If this aspect is considered important in the successful accomplishment of the job and the measure is sufficiently reliable, it would be included as a criterion measure. As a matter of fact, considerable effort is directed at studying the correlations among the available measures to discover the genuine and important aspects of job performance. It is conceivable and has actually happened that none of the measures available for criterion use show any correlation with each other, in spite of the fact that the job requirements indicate that certain job aspects are not independent of each other. Put another way, measures which should be correlated with each other turn out not to be correlated. Such a finding suggests that the grading and evaluation systems employed are defective and points also to the possibility that the job duties or training content are not effectively organized. Until the necessary corrective steps are taken, them is little point in developing expensive measuring instruments to predict job performance when there is nothing to predict.

(1) There is another type of relationship among criterion measures which is of considerable importance—the relationship among measures at successive stages in a man's Army career. Are the men who are good in basic training good in combat? Are men who are good in school good on the job? With the introduction of an Army enlisted student evaluation roster and an enlisted on–the job data sheet to provide information on performance at school and on the job at a later date, sizable groups of men in various military occupational specialties have been available to provide some research answers on these criterion relationship questions. There is only a moderate relationship between school grades and job performance ratings, a finding which has raised these important questions for future criterion research: What factors are not included in the prediction of job performance on the basis of aptitude measures and how can these factors be taken into account? It is believed that the most prominent of such factors or variables may be differing motivation on the job, differences between requirements of the job and academic requirements laid down by the school, and differences in the bases on which job ratings are made from locale to locale for the same job.

(2) One final point should be made regarding the relations among criterion measures. In view of the Army's mission to fight and win wars, it might be thought that the criterion measure that should be used is performance in combat. This point is recognized and studies to validate various instruments against performance in combat are made when opportunity offers. However, it does not necessarily follow that the ultimate criterion—combat performance—is the only one that should be used. For one thing, combat criteria are obviously not always available, and during peacetime, when the Army still has a job to do, other criteria must be used. For another, combat criteria are not especially appropriate for all military occupations. For example, the closest to combat that automotive engine rebuilders may get is rear echelons and there is little point in attempting to got measures of performance under fire. On the other hand, there is little doubt that for riflemen a criterion measure based on performance in combat is appropriate. This discussion is another illustration of the principle that the selection of criterion measures must be related to the purpose of the instrument to be validated and the purpose of the job.

d. Criteria of Unit Effectiveness. The need to evaluate the effectiveness of small military units (squads, platoons, companies) and teams operating man–machine systems is an urgent one in the Army. Standard field problems have been used to aid such evaluations in connection with training and determination of combat and operational readiness.

These problems differ from traditional field and situational problems in that their design, development, and scoring have involved measurement principles discussed in this pamphlet. Yet they are just as useful as the traditional field problems in training and as aids to administrative and operational decisions, in addition to which they have been designed to meet the requirements of adequate criterion measures. Personnel instruments, training, and management policies can be validated against these measures, provided they are well designed. With good criterion measures of unit and systems effectiveness, it becomes possible to study systematically the impact on unit and systems effectiveness of varying the characteristics, training, and work environment of the members of the group and to answer questions such as the following: Should all the men in an infantry squad have high aptitude area GT scores or will the squad be equally effective if a certain number have high scores but the others have fairly low scores? Is it necessary for maximum effectiveness of a man–machine system that all the men in the system be thoroughly familiar with all major aspects of the functioning of the system? Have attempts to improve the morale or motivation of 9, unit or team proved successful? In short, recognition of the value of adequate measures of unit and systems effectiveness makes this an area of increasing importance in military personnel management.

63. Relation to Policy

The extended discussion of criterion problems should not be lost sight of in establishing policy regarding the use of personnel measuring instruments. The criterion measure plays a critical part in the development of an instrument, and should be kept in mind when establishing policy as an aid in guarding against overestimating or under estimating the effectiveness of the instruments. The criterion is a measure of purpose and hence determines the effectiveness of an instrument for that purpose. For other purposes, other criteria—and consequently other instruments—may be needed. In other words, the criterion should not be lost sight of in establishing policy to determine the use of certain measuring instruments.

Section V
SUMMARY

64. The Criterion Plays a Critical Role in the Development of Personnel Measuring Instruments

a. Criteria play a decisive part in determining the effectiveness of personnel measuring instruments. Criterion measures for validating instruments may also be used in other areas of personnel management.

b. The adequacy of the criterion measure is determined by the following:

(1) Relevance to purpose of instrument and job.

(2) Comprehensiveness and weighting.

(3) Freedom from bias.

(4) Consistency.

c. Job performance measures based on automatically recorded actions provide the most objective criteria. Measures based on observer–recorded performance are most objective when little or no evaluation is required of the observer. Checklists specifying the precise action to be observed may aid in reducing the element of judgment in the measure obtained.

d. A major research interest is the development of methods to improve ratings as criterion measures. The use of an average of a number of ratings instead of single ratings is a simple and effective way of increasing the value of ratings.

e. Criterion research is directed at studying available measures and developing new ones. There is an increasing need for criterion measures based on unit performance as well as on individual performance.

f. The nature of the criterion should be included in the factors considered when establishing policy governing the use of personnel instruments.

64B. Title not used.

Paragraph not used.

Chapter 4
THE MEANING OF SCORES

Section I
THE NATURE OF PERSONNEL MEASUREMENT

65. Measurement Is Approximate

a. Measurement in personnel research is approximate rather than exact. However, it is not fundamentally different in this respect from measurement in other fields. The difference is one of degree, not of kind. An engineer thinks in terms of “tolerances” and specifies what tolerances he desires. He knows that when he says a metal bar is 33 inches long, he

does not mean that it is exactly 33 inches long, but perhaps .014 or .045 inches longer or shorter than the stated length. The amount of difference which can be tolerated depends upon such factors as the additional time and expense of measuring the bar more exactly and whether the more exact measurement will make any difference in the use of the bar.

b. A similar situation exists in personnel measurement. A test score of 119 does not mean exactly 119. A score of 119 may mean, say, any value from 111 to 127, and the personnel psychologist may say something to the effect that in a particular case the chances are 40 out of 100 that the score is actually between 118 and 120, but 97 out of 100 that it lies between 111 and 127. The range of scores from 111 to 127 may represent an area in which we are almost completely certain that the man's true score lies. Factors affecting the size of these "tolerances" in personnel measurement will be discussed in this chapter. See also chapter 5.

c. Measurement in personnel research is always with reference to a specified population. As was stated in paragraph 34, each new test is standardized on a sample as representative as possible *of the Army population for which the test was designed.* Sometimes that population is a full mobilization population, sometimes a peacetime population of some restricted kind (Army driver trainees, Special Forces volunteers, Advanced ROTC applicants) and sometimes merely the trainees in a particular Army school course. Use of the various types of scores and standards discussed in this chapter is always contingent upon an understanding of the population a given test or measure has been selected to represent.

66. Scores Are Relative

It will be recalled (par. 7) that a number by itself is not a meaningful score and that standardization and validation are needed to provide the necessary meaning. A score must express a relationship between one man and the other men, either on a particular test, or on a criterion of value on the job. That is, the score must be interpretable as a "high," "low," "slightly above average," "average," etc., score on a test, or it must be interpretable as a score made by a "good soldier," "poor soldier," and so on. Scores, then, are relative.

Section II
TYPES OF SCORES AND STANDARDS

67. Adjectival Measures

a. A superior observes his subordinates and rates them as *excellent, superior, satisfactory,* or *unsatisfactory.* Another rates his subordinates as *outstanding, excellent, satisfactory, unsatisfactory,* or *very unsatisfactory.* This is a simple way of evaluating men. But is it as simple as it appears? What do the adjectives mean? Obviously, they do not mean the same things to all men—in the above example, *excellent* means "best" to one man, "next to best" to the other. Even if it were agreed to use *excellent* to mean only one thing, say "best," the problem is by no means solved. One man has high standards and very few of his men are evaluated as excellent. Another man has lower standards and a large number of his men are evaluated as *excellent.*

b. Even if differences in standards could be eliminated, adjectival measures still have a serious shortcoming. Suppose 100 men are rated on their value for a certain type of job and 15 are rated excellent. Suppose, further, that it is desired to pick the best, 10 for a special assignment. Which of the 15 "excellent" men are the 10 best? It is this question which illustrates a final difficulty in the use of adjectival measures for instruments involved in personnel actions—they indicate only very roughly how one man stands in relation to other men.

c. To sum up, then, adjectival *measures* are crude measures and are of limited usefulness in Army personnel actions.

68. Raw Numerical Scores

a. An improvement over the adjectival measure is the raw numerical score which may be a simple count of the number of correct responses made by an individual. The test answers are scored with a fixed key so that what the scorer thinks about the test or the answers is removed from the picture. Any two people scoring a paper should, except for errors, come up with the same total score. However, intelligent interpretation of such scores is still quite difficult.

b. The difficulty of interpreting raw scores can be seen from the example below. Jim Brown is given three tests on which he makes the following scores:

Table 4–1
Raw Numerical Scores tables

Test	Number of items	Incorrect answers	Correct answers
Clerical speed test	225	50	175
Shop mechanics test	40	1	39
Vocabulary test	53	5	48

Notes:

Brown missed only one question on the shop mechanics tests, but five on the vocabulary test and 50 on the clerical speed test. Yet his score of 39 in shop mechanics was lower than his score in the other two tests. How can the classification specialist determine from these scores the test on which Brown did the best? He can do so only by taking into consideration the difficulty of the test. It might be more of an achievement to get all but five questions right on a difficult test than all but one right on an easy test of the same length. Before Brown's three scores can be compared, they must be changed to some common scale to take account of different numbers of items, differences in difficulty of the tests, and differences in ability of the people taking the tests.

c. Similarly, it is difficult to compare two men on the same test on the basis of raw scores. Suppose, for example, it is desired to compare Brown with Smith who answered, say, 29 correctly on the shop mechanics test. Compared with Brown's 39, Smith's 29 is ten points lower. Is a difference of ten points a large one or a small one? Suppose, further, that in the clerical speed test Smith answered 165 correctly as compared with Brown's 175. Again, there is a difference of ten points, and again the question is asked, is this difference a large one or a small one? The only way to answer this question is, as before, to convert all scores to a common scale.

d. Raw numerical scores, then may have the advantage of objective and uniform scoring, but they permit only a very rough comparison between men on one test or between one man's performances on two different tests. Types of standards that have been developed and are being used will be discussed in the rest of this section. It should be kept in mind, of course, that the computation of raw scores is always preliminary to the calculation of any more refined scores.

69. Percentile Scores

a. General. One type of standard is the percentile score by which one person's performance on a test is evaluated in terms of the scores of all the other people who took the test. Percentile scores may range from 0 to 100, when rounded to the nearest whole number. If a man gets a percentile score of 90, this means that his raw score was higher than the raw scores of 90 percent of the men who took the test, and equal to or lower than the raw scores of 10 percent of the men. If he receives a percentile score of 50, he excelled 50 percent of the group.

b. Advantages. Percentile scores offer a realistic way of interpreting tested performance because they compare the individual with the group. Since it is virtually impossible to determine the absolute amount of ability possessed by an individual, the determination of his relative ability is the most meaningful approach.

c. Disadvantages. Percentile scores are generally favored because they are easily understood. But a percentile system suffers from one serious disadvantage. While it reveals what proportion of the total group tested an individual excels, it does not indicate how much better or poorer he is than any one of the others. This is true because percentile units are not equal to one another in terms of raw scores. For example, 5 percentile points difference for the best scores may represent a difference of 15 raw score points, while 5 percentile points for average scores may represent a difference of only 4 raw score points. This difference in meaning of percentile units arises from the well–known fact that the number of men making extremely high or extremely low scores is much smaller than the number of men who make more moderate scores. Thus, since percentile units are not uniform, and for this reason cannot be averaged, percentile scores have limited value. This problem is discussed further in paragraph 74.

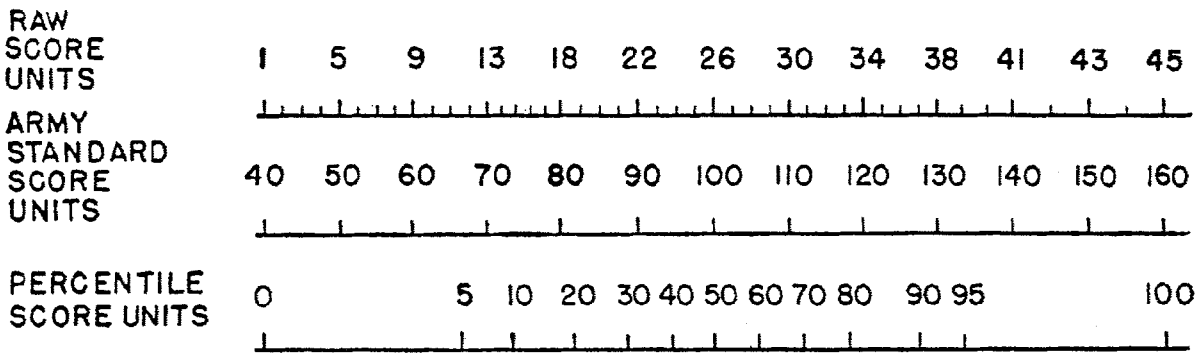


Figure 9. Comparison of raw score units, percentile score units, and Army standard score units for a distribution of test score with raw score mean of 26 and standard deviation of 8

Section III
STANDARD SCORES

70. General

a. A more flexible type of standard than any of the types mentioned is the standard score system. The standard score system retains most of the advantages of the percentile score with the added benefit of equal score units. Figure 9 shows how raw score units and percentile units may differ in value, whereas standard score units are of uniform value throughout the scale. A standard score shows more precisely how much better or poorer than the average score any particular score is.

b. Every classification specialist should be thoroughly trained in the statistical techniques involved in computing standard scores and converting them into Army standard scores. These techniques are described briefly in the following paragraphs of this section. More complete and detailed information can be found in any standard textbook in psychological statistics.

c. Readers of this pamphlet, other than classification specialists, may well omit paragraph 72 which describes the computation of the standard deviation and the standard score.

71. Standard Scores Measure Relative Performance

a. There is more to the standard score system than a comparison of the individual with the average performance. Standard scores also compare people with one another. Some tests are easier than others; some groups taking the test are more nearly of the same level of ability or knowledge than are others. In a group whose members are fairly equal in ability, a high raw score represents more of an achievement than in a group whose members are of a wide range of ability. In the first case, a high score stands out from the rest of the group. In the latter case it does not. The way this works can be seen in the following illustration:

Table 4–2
Scores illustration tables

Scores on Test A	Scores on Test B
80	80
60	77
58	74
56	68
53	46
49	40
44	30
40	25
8 /440	8 /400
55–Mean Score	55–Mean Score

Notes:
On Tests A and B the average or mean scores made were both 55. On Test A, however, the person achieving a score of 80 had 20 score points more than the next highest individual, while on Test B a score of 80 was only 3 points more than the next highest score. On Test B there were a number of scores very close to 80. The man achieving 80 on Test A was outstanding on that test. The man achieving 80 on Test B was only slightly superior to a number of other people who took the test.

b. A standard score system would take into consideration not only the mean performance of the group, but the relative performances of all others taking the test by indicating how far from the mean score each individual score is. How this is accomplished will be illustrated by describing the method of computing standard scores.

72. Computation of Standard Scores

a. There are two elements which must be known before the standard scores can be computed:

(1) *The mean scores,* which is merely the mean of all the raw scores of the standard reference population (par. 34).

(2) *The standard deviation,* a measure of the spread of scores or how much members of the tested group vary in ability among themselves. The standard deviation is represented as a certain distance on the scale of measurement above and below the mean score made by the people taking the test. It is computed in the following manner:

(a) Subtract the mean raw score from each raw score.

(b) Square each of the remainders, or deviations, so obtained.

(c) Add all of the squared deviations together.

(d) Divide this sum by the total number of scores.

(e) Extract the square root of this quotient.

b. When the mean raw score and the standard deviation of the raw scores are known, any individual's raw score on a test can be converted to a standard score by the following method: Subtract the mean raw score from the individual raw score and divide by the standard deviation. (Standard score=(Individual score minus mean score)/(Standard deviation))

Note. For any score on any test, therefore, a standard score can be computed and tables can be drawn up showing the standard score equivalents for all the raw scores. It can be seen that the standard score takes into account the man's deviation from the mean and the relative amount of variation in the group as a whole.

73. The Army Standard Score

There are two undesirable features of standard scores which have led the Army to make a slight modification in their use. For one, the standard score scale is short; a range from –3 to +3 takes in practically all cases. In order to make a fine enough differentiation between men, it is necessary to resort to the inconvenience of decimal scores. Another disadvantage is that all scores below average are negative in sign, another inconvenience. The Army, therefore, multiplies each standard score by 20 in order to get rid of the decimals, and adds 100 to each score to get rid of the negative sign. If a man gets a score which is exactly the same as the mean score on a test, his standard score is equal to 0 (computed as explained in par. 72 above), and his Army standard score is equal to (0 X 20) + 100=100. Thus, the Army standard score system has an average of 100 and standard deviation equal to 20 points. The possible scores that can be obtained on an Army standard score scale range from approximately 40 to 160. It should be realized that the selection of 100 to represent the mean and 20 the standard deviation is purely a matter of convenience. Other numbers could have been used without altering the relationships of the various scores to each other and to the mean.

74. Interpretation of Army Standard Scores

a. Normal Distribution of Human Abilities. In order to interpret Army standard scores, it is necessary to understand a little of the nature of the distribution of human abilities. If the various raw scores made by all persons taking a test are plotted on a graph, the resulting distribution of scores looks something like figure 10, unless the test is extremely easy or extremely difficult. In the typical distribution of test scores, however, notice that relatively few people receive either the extremely high scores or the extremely low scores, but that most people's scores tend to pile up in the center. The statement of the distribution of scores on a test may be generalized to include most human abilities. That is, a few people possess a great deal of an ability, a few others very little, while most people possess moderate or "average" amounts of an ability. The distribution of most human traits and abilities has a shape very similar to figure 11. This curve is usually referred to as the normal distribution curve. When an Army instrument is standardized, it is usually given to a large group of individuals representing the population with which it will be used. The group is the standard reference population. The distribution of scores for this group will, in most cases, follow closely the normal curve. The normal curve can then be assumed as a basis for setting up the standard score scale and interpreting scores.

b. Usefulness of the Normal Distribution curve. Because the distribution of most human abilities follows closely the pattern of the normal curve, this curve is very useful for interpreting scores.

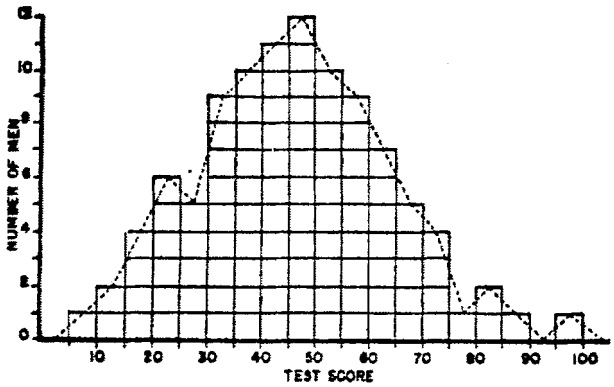


Figure 10. Graph of a distribution of test scores

(1) The central point on the baseline represents the mean or average score. All other points along the baseline represent scores of different magnitudes from low to high. It can be seen that them scores can also be conceived as differing by various amounts from the mean score. The standard deviation (par. 72), is the yardstick by which the relative values of the scores may be indicated. It is useful to note the number of people who score at levels of 1, 2, and 8 standard deviations above and below the average score. With a knowledge of the average score and the standard deviation of the distribution of scores, interpretations of any individual's performance on a test can be made and individuals can be compared directly.

(2) In addition to comparing people directly with one another, standard deviations give us much the same information that percentile scores yield—that is, they tell us what proportion of the population which took the test fall above or below any one score. When the distribution of scores on a test follows the normal curve, a score of one standard deviation above the mean score indicates that the individual achieving that score has done better than 84% of all those taking the test. Since all Army standard scores have a mean of 100 and a standard deviation of 20, the individual who receives a standard score of 120 is one standard deviation above the mean (100+20) and has exceeded the performance of 84% of the people. If his score had been expressed in percentile terms, he would have had a percentile score of 84. Each person's score can be compared in this way with those of the rest of the group. Standard scores have essentially the same meaning and the same interpretation from test to test.

(3) Figure 11 shows the relations among standard scores, Army standard scores, and percentile scores and relates them to the curve of normal distribution. It should be pointed out that, once the curve of distribution of scores is known, standard scores and percentile scores may be interpreted in exactly the same way—that is, there is an equivalent percentile score for every standard score. Standard scores, then, yield the same information that percentile scores do, and have the great advantage that since the units are equal, standard scores can be combined or compared with other standard scores. And from the technical point of view, the equal units are of great importance. Without them, statistical computations would be seriously handicapped.

c. Precautions in Using the Normal Distribution Curve. Percentile scores and Army standard scores in figure 11 were related to a theoretically perfect normal curve of distribution. In practical operations, raw scores may not only have the kind of sampling fluctuations that produced the irregularities shown in figure 10 but may deviate in some more significant fashion from the curve of normal distribution. For example, selection and classification test scores for the full mobilization population, on which many Army–wide tests are based, characteristically depart from normality along a certain range of their distributions. For this rea son, conversion of operational percentile scores, for example, Armed Forces Qualification Test scores, to their standard score equivalents will reveal a standard score distribution that is at odds with the conversions shown in figure 11 for an idealized normal distribution. Such standard scores are appropriate for use and interpretation in the manner described in this chapter, since the distributions they represent approach normality.

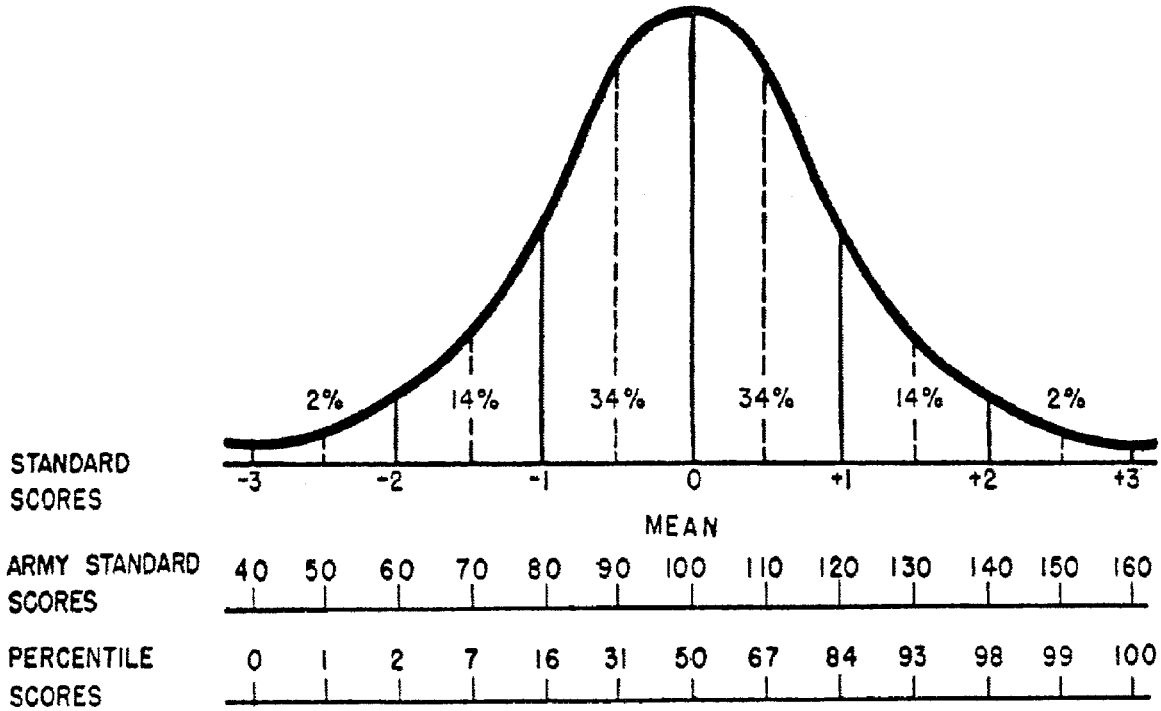


Figure 11. Standard scores, Army standard scores, and percentile scores in relation to the normal curve of distribution

75. Normalized Standard Scores

a. Except for special purposes, a test would not be used if its distribution of scores differed so markedly from the normal distribution as to show, for instance, a sharp peak at one end, or, to take another instance, several widely separated peaks. Abundant experience has shown that in the usual populations, well constructed instruments yield distributions of scores approximating the normal. This experience supports sound theoretical considerations that the traits being measured by the scores are normally distributed in normal populations. There is, then, reason for expecting most test scores to be distributed normally and it is therefore desirable to convert obtained distributions of scores to normal distributions when justifiable. Instances where the distributions obviously should not be normalized are when the number of men taking the test is small or when the men taking the test are highly selected for intelligence, length of military service, grade, etc.

b. Raw scores are normalized by conversion to percentile scores which, in turn, are converted to their equivalent standard scores. Normalized standard scores are thus merely standard scores based upon a distribution of raw scores that has been made normal. The distribution is normalized only when there is reason to believe that the underlying trait measured is normally distributed.

76. Conversion Tables

The Army provides tables with most of its tests so that the raw scores can readily be converted into Army standard scores. Each of these tables is based upon the distribution of scores achieved on the test by the standard reference population (par. 34). Thus, on most Army–wide tests, an individual's score is compared with scores of a large and representative group of Army men, and each individual's standing on the particular trait tested is given in terms of the Army as a whole. All individuals are compared with the same standards. Army standard scores, obtained by use of authorized conversion tables, are therefore much more useful in personnel measurement than would be such scores based upon data collected at any single Army installation. Moreover, the use of these conversion tables insures uniformity in translating raw scores into standard scores. The usual table involves two columns of numbers. The first is a list of all possible raw scores; the second contains the standard score corresponding to each raw score. Using the table involves looking up each obtained raw score, then reading off and recording the corresponding standard score. Computation of either percentile scores or Army standard scores is feasible for any distribution of raw scores obtained

on a personnel test—even scores obtained on a one–time basis in the classroom—but the user must continually ask himself what population is represented thereby and interpret the scores accordingly.

77. Advantages of Army Standard Scores

Army standard scores are useful for selection, classification, assignment and utilization purposes for the following reasons:

a. They state test performance in such a way that small individual differences in ability or achievement are clearly revealed.

b. They tell how a man ranks in comparison with other Army men.

c. They make it possible to compare an individual's expected performance with that of others, and to compare each man's performance on a number of tests.

d. They are mathematically convenient, and therefore make further statistical analysis of data more practicable.

78. Army Standard Scores Are NOT "IQ's"

Army standard scores bear no direct relationship to such concepts as the "IQ" (intelligence quotient), or "MA" (mental age), and Army test results must not be interpreted in terms of these concepts.

Section IV
RELIABILITY

79. Definition

a. A reliable measuring instrument is one which can be depended upon to measure a given object in exactly the same way every time it is used. But such reliability depends upon the refinement of the instrument and the care with which it is used. Successive measurements of a room with a yardstick may vary by several inches if the thumb is used to mark the point where the rule is to be re–laid each time.

b. With psychological measuring instruments the situation is somewhat aggravated. Tests themselves often fall farther short of perfection than is usual with physical instruments. And, while the dimensions of a room remain quite constant, human beings are changeable and are affected by many conditions which have no effect upon physical objects.

80. Importance of Reliability in Psychological Measurement

a. If the measurements resulting from a test are to be depended upon in making important decisions, they must be sufficiently reliable; there must be evidence that men, repeating the test under the same conditions and without changing significantly, will make nearly the same scores the second or third time as the first. A person's arithmetical ability cannot be represented by a score of 30 in the morning and 90 in the afternoon, any more than his height can be 5′ 2″ in the morning and 6′ 2″ in the afternoon. People don't grow that fast either mentally or physically.

b. A test's reliability is important also because of its relation to the test's validity. A test can be reliable without being valid–that is, it can be a consistent measure of something other than the particular trait you wish to measure; but a test cannot be valid without possessing some degree of reliability.

81. Conditions Affecting Reliability

There are four main types of conditions which affect the reliability of an instrument—

a. The Instrument. If the instrument is too short, that is, contains too few items, the scores are likely to be greatly affected by chance. If the items are poorly written so that the men have to guess at the meanings, their answers will reflect their guesses.

b. The Men. If the instrument is so long that the men get tired or bored, the responses to the later items will be affected and the test as a whole will not measure consistently whatever it is supposed to measure throughout its length. If some of the men have shown no interest, or if they were tired or in poor health, or if some of them have been specially coached, the consistency with which the instrument measures will be affected.

c. Administration. If the examiner varies his instructions from time to time, if he gives special instructions to some men but not to others, if his instructions are different from those of other examiners, the reliability of the instrument will obviously be affected. If the physical conditions of the room in which the test is being given are poor, the reliability may be reduced.

d. Scoring. Irregularities in scoring affect the reliability. However, in the Army, objective tests are used and the scoring is obviously objective. The evaluation of essay answers, as is well known, is much less consistent than the scoring of objective tests.

82. Estimating the Reliability of a Test

a. General. For two important reasons it is necessary to know just how reliable a test is prior to its use: to determine whether it is suitable for use, and to know just how much unreliability to allow for in interpreting the test results.

Before being released, Army tests are subject to rigorous and exhaustive statistical checks to determine the dependability of the measures which may be obtained with them.

b. Method. Reliability may be measured by comparing the scores achieved by men who take the same test two or more times under identical circumstances or who take equivalent forms of the test. If each man gets the same score—or almost the same score—each time, the test must certainly be reliable. But if many individuals get widely different scores each time, the test is unreliable, and no confidence can be placed in any single score. It is almost impossible, however, to give the test more than once under identical circumstances. Besides the obvious effects of familiarity with the test, there are bound to be other changes in the examinees from time to time, and it is often too difficult to construct an exactly equivalent form of the test. In practice, therefore, statistical techniques are used which estimate the result that would be obtained if two equivalent forms of the test had been administered at a single session. The techniques are accurate for most types of tests and have the important advantage of saving both time and effort. The mathematical statement of the result is called the reliability coefficient.

83. Uses of the Reliability Coefficient

The reliability coefficient gives important information about the test and the interpretation of test scores.

a. Improving Test Reliability. The first use of the reliability coefficient is to determine whether the test is reliable enough for classification purposes. If the reliability proves to be too low, it can usually be increased by the addition of more items of the same kind as are in the test (par. 22). But since long tests are time–consuming, reliability beyond practical usefulness is not sought. Army tests before being released for use are made sufficiently reliable for use in selection, classification, assignment and utilization, and they will usually remain so provided they are used as directed.

b. Reliance on Test Scores. As pointed out in chapter 1, some calculated risk is involved in personnel actions. Army tests are reliable, but they are not perfectly reliable. A few points difference in scores does not always mean a real difference between the individuals receiving the scores. Yet, when a minimum qualifying score is set, it is recognized and used throughout the Army as the dividing line between those who are acceptable and those who are not acceptable for a certain type of assignment. Admittedly, some of the men accepted would, if retested, fall below the minimum qualifying score and some of those rejected would reach or exceed it. This risk is taken when a decision is made to use a certain point on the measurement scale as a standard for acceptance or rejection. It is a decision essential to an effective personnel management system. The misclassification of a few men—and their number can be forecast—is weighed against the uncertainty of classification without such a standard. When the minimum qualifying score is disregarded, or when scores are juggled so as to nullify its effect, just that much more of uncertainty—of uncalculated risk–enters into the classification and assignment of men in the Army.

c. Retesting. Because Army tests are reliable it cannot be expected that retesting will, on the whole, result in marked score increases. Familiarity with the test and the test situation may contribute a few points of increase, but experience proves that this increase, on the average, will be small. Moreover, it is important to recognize that there is no virtue in getting high scores for their own sake. The only purpose in using tests at all is to enable predictions of job or training success, and there are seldom grounds for supposing the higher retest score to be a better predictor than the original. On the contrary, there usually is reason to suspect its validity. Therefore, retesting should be practiced sparingly. It is justified only when the record shows a discrepancy which seems to merit investigation. A man's score on a test may be markedly inconsistent with his abilities as inferred from other information (education and occupation, for example). Or a man's records may show obvious discrepancies as a result of errors in recording scores. In such cases, steps should be taken to see that a correct score is obtained and recorded.

Section V
SUMMARY

84. Standard Scores Aid Interpretation

a. Army standard scores offer a meaningful way of describing test performance. They compare the performance of the individual with that of other Army men on the particular skill or aptitude which the test measures.

b. The Army provides *conversion tables* which enable field personnel to translate raw scores into standard scores.

c. The *reliability* or consistency of a test as a measuring instrument is carefully checked prior to the test's use. Field personnel can contribute much to maintaining the reliability of a test by strict adherence to regulations governing its use.

84B. Title not used.

Paragraph not used.

Chapter 5
THE PRACTICAL VALUE OF SCORES

Section I
THE PRACTICAL SIGNIFICANCE OF SCORES

85. General

As indicated in the preceding chapter, a raw score on any personnel measuring instrument has no meaning by itself. It begins to acquire meaning when it has been converted to a standard score or some other means of permitting its comparison with other scores. However, even the standard score does not represent the full significance of a test score. That a soldier has a standard score which puts him in the top 10 percent of all men in some particular respect is factual information. But it will remain an isolated useless fact until its practical significance is determined. In general, personnel measuring instruments are only a means to an end. Their main justification in the Army is to insure that a man assigned to a given job can be expected to perform well in that job.

86. Ambiguity of Instrument Titles

The name of an instrument tells only in a general way what the instrument measures. "Clerical aptitude" tests measure abilities related to proficiency in clerical work and the "automotive information" test gives indication of the probable proficiency of an automobile mechanic. These titles, however, give only rough indication of what the tests are supposed to measure and predict. Full knowledge of how the tests can be used comes only from continued research in which relationship of test scores to performance in a considerable variety of jobs is studied.

87. Factors Affecting the Use of Scores

It is the purpose of this chapter to consider the various factors or concepts which bear upon the most profitable use of test scores. One of these factors, reliability, was discussed in chapter 4. Others are validity, the selection ratio, minimum qualifying scores, and the supply and demand for persons with various skills, knowledge's, and aptitudes. General considerations having to do with sampling and the statistical significance of personnel research results will also be discussed.

Section II
VALIDITY AND SCORES

88. The Correlation Coefficient

a. General. The correlation coefficient is a statistical measure which, because of its usefulness, is encountered frequently in personnel research. It may be defined as a number, between the limiting values of +1.00 and –1.00, which expresses the degree of relationship between two variables. These variables may be two sets of test scores on a group of men, test scores and criterion measures, scores on the same instrument from two different periods of time, or measures on any other traits, qualities, or characteristics which exhibit differences in magnitude.

b. Graphic Illustration of Correlation. A correlation coefficient can be illustrated simply by means of a graph. In the example chosen (fig. 12), the two variables compared are a test and a criterion, for each of which scores are available for 5 men. Scores of the 5 men are given in the first box of the figure. In the second box the manner of plotting the point representing the paired values for the first man is shown. The horizontal axis is chosen here to represent test scores and the vertical axis the criterion. In the third box, all 5 points have been plotted for the example, and in the fourth box the result of computation (not shown) is given—the correlation coefficient of +.54 for these values. Plotting of such a graph is similar to plotting a distribution of scores, but since there are two scores for each person, the distribution may be termed a double distribution (also called "scatter diagram" or "scatterplot"). As can be seen, it closely resembles the plotting of points on a map through the use of coordinates.

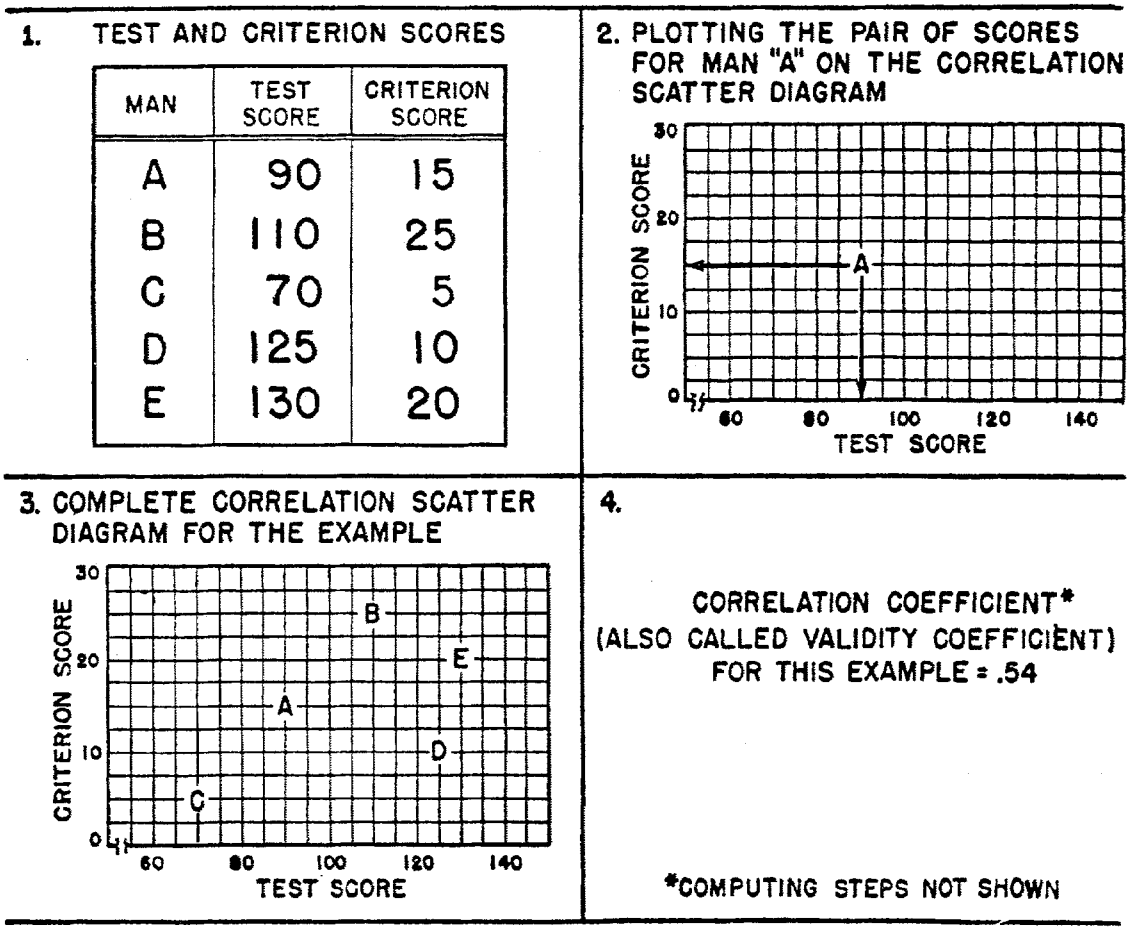


Figure 12. Illustration of the relationship between test and criterion scores for 5 men

c. Interpretation. If there is no relationship between the, two variables, the correlation coefficient is zero. If exact or perfect relationship between the variables exists, a coefficient of either +1.00 or –1.00 is obtained; if high scores on one variable are associated with high scores on the other while low scores on one are associated with low scores on the other, the coefficient is positive; if high scores on one variable are associated with low on the other while low scores on one are associated with high on the other, a negative coefficient is obtained. Figure 13 shows the kinds of graphs associated with correlation coefficients of .00, +1.00, –1.00, and +.55.

(1) The correlation coefficient should be interpreted in terms of a "line of relationship." The closer the points to this line, the higher the correlation between the two variables. In the second and third graphs of figure 13, where correlation is perfect, all points would fall exactly on this line.

(2) Accuracy of prediction of one variable from another is a function of the degree of relationship between them. In the case of a correlation coefficient of 1.00, prediction is perfect; knowing a score on one variable, one can predict the score of the same individual on the other variable without error. When the correlation is zero, it is impossible to predict one score from knowledge of another.

(3) Any thorough–going discussion of the correlation coefficient should introduce the concept of regression, which may be defined as the tendency of a predicted value to be nearer the average than is the value from which the prediction is made. The lower the correlation, the greater is this tendency. When there is no correlation between two variables, the best estimate that can be made of the value of any variable is the average score on that variable. For fuller discussion of these ideas, a statistics textbook should be consulted.

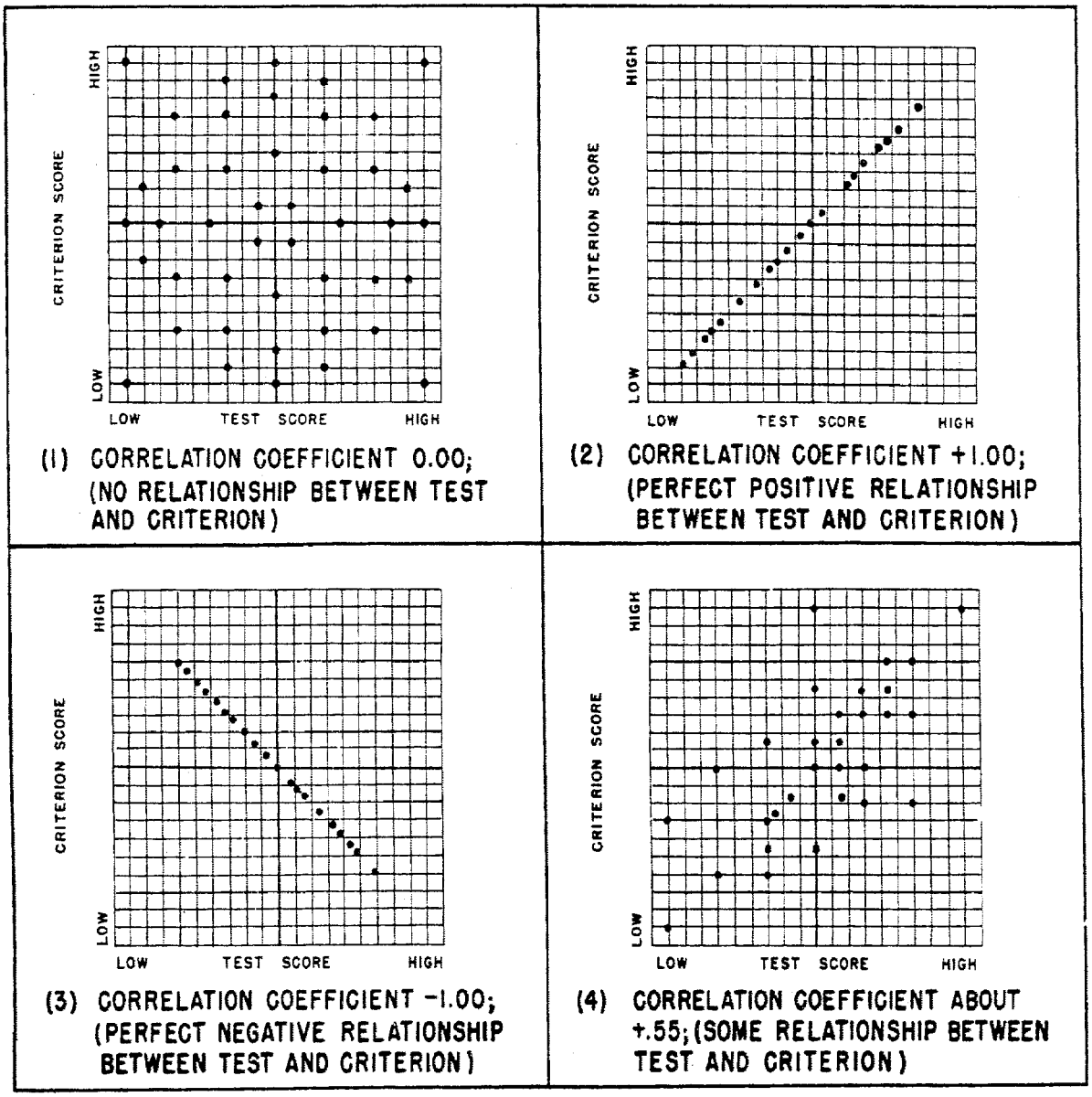


Figure 13. Correlation scatter diagrams illustrating different degrees of relationship

(4) Correlation coefficients are sometimes given special names according to the purpose for which they are used. A correlation coefficient used to express validity is called a validity coefficient; used to express reliability it is called a reliability coefficient.

89. The Validity Coefficient

a. Meaning. The validity coefficient for a test is the coefficient of correlation between the scores on the test and the scores on a criterion representing what the test is supposed to measure. It indicates the extent to which the individuals who score high on the test also tend to score high on the criterion, and the extent to which individuals who score low on the test also tend to score low on the criterion. For example, a very high validity coefficient, say .90, would mean a high degree of relationship between the test and performance on the criterion, such that a man with a high test score would be extremely likely to exhibit excellent performance on the criterion; there would be some chance of error in estimating his criterion performance from his test score, but it would be slight. Similarly, men with test scores at other points on the scale would be very likely to perform on the criterion measure at the corresponding levels. On the other

hand, if the correlation between test and criterion is low or close to zero, a man with a given test score might achieve criterion performance anywhere along the scale–high, low, or average.

b. Use With Army Tests. Much Army personnel research is directed toward obtaining tests which will predict job performance or training success with the highest possible accuracy. When scores on Army tests are correlated with later performance, the obtained validity coefficient may range from zero to .75, or higher. It is obvious that the higher the validity coefficient, the greater the predictive value of the test. Yet, even in cases where research has been able to produce predictors of only low validity, their contribution to selection and classification may be a sufficient improvement over purely random placement of men to make their use worth while. To predict training success, the Army generally requires high coefficients, i.e., .50 to .80. On–the–job performance is more difficult to predict and validity coefficients may more typically range from .30 to .50. This leads to an important point in the interpretation of validity coefficients—the size of the coefficient which is deemed useful is relative to the type of instrument and the situation in which it is to be applied. In some situations, a predictor with a very low validity may be far better than no predictor; in other situations, the validity must be much higher before the instrument is worth introducing.

c. Specificity of Validity. From the foregoing it can be seen that the validity coefficient shows the degree of relationship between scores on a test and performance in some particular job. Validity, then, has meaning only in relation to a particular test and a particular job. A test may have high validity in predicting performance on some jobs, moderate validity for certain other jobs, and no validity at all for certain additional jobs. It may have validity in predicting success in training but little or no validity in predicting success on the job for which the training is conducted. It is meaningless to speak of the validity of a test; it is necessary to specify what it is valid for.

d. Use With Test Batteries. So far, the discussion of validity has been limited to the predictive value of scores from a single test. Actually, as the discussion in chapter 2 indicated, several test scores are often weighted and combined to give a single composite score for the battery. Performance on a job can often be much better predicted if measures of a variety of aptitudes and traits are used in estimating later job performance. The discussion of the validity of a test score applies equally well to the validity of a composite score resulting from a combination of several test scores.

e. Expectancy Charts. In the use of a test for selecting men who will perform successfully on a given job, the meaning of the validity coefficient is sometimes expressed in the form of an expectancy chart. Such a chart shows what the probabilities are that men making various tog scores will achieve better than average success on the particular job. The validity coefficient can thus be translated into terms easy to apply in a practical situation.

90. Non–Empirical Validity

a. The foregoing discussion of validity is based on empirical determination of degree of relationship between a set of test scores and a set of criterion scores. Empirical validity is expressed as a validity coefficient computed from the two sets of scores. But frequently it is necessary to resort to the use of non–empirical validity when it is not possible to determine a computed relationship. Non–empirical validity—sometimes termed "face validity"—is applied to the usefulness of tests developed in several different ways. In each case, however, validity is inferred rather than established empirically.

b. Validity by definition means that a test or a measurement bears a common–sense relationship to the measurement objective. In the construction of an automotive information test, for example, the test may set such, a task that the universe of possible automotive information facts (of which the test is a representative sample) is the only practicable criterion and the test may therefore be considered a valid measure of the universe defined in terms of the sample. In practical terms, one would first outline the concepts to be covered in a particular course of study, and then sample these concepts systematically and comprehensively until an adequate representation of all the questions which might be asked on the course is achieved.

c. "Built–in" validity, a similar type of inferred validity, refers to the test in which subject–matter experts determine the items to be included in terms of requirements of the job for which the test is being constructed. In the construction of achievement tests, where questions are written to cover essential aspects of each job, the subject matter expert provides items which in his judgment are likely to be answered correctly by men who know the subject and incorrectly by those who don't know the subject.

d. Validity by hypothesis refers to the construction of a new test on the basis of prior knowledge; items are selected from other tests on the basis of statistics computed in previous studies. In the absence of contrary information, the validity of the items in the new test may be inferred.

e. The appearance of validity is directed at obtaining general acceptability of the instrument among persons who will have contact with it; the user of the test, not having access to empirical validity data, may react unfavorably to an item which, although empirically valid, does not have the appearance of relevance or practicality.

f. A final point—empirical validity and non–empirical validity are both to be distinguished from internal consistency, which refers to correlation between each item and the other items in the test (ch. 2). Internal consistency may be determined either by subjective estimate or by statistical computation from empirical data, but it is usually not thought of in terms of test validity.

g. It should be clear that non–empirical validity is no substitute for empirical validity. In the case of measurement instruments which involve right and wrong answers (achievement tests, for example), the "built–in" validity may be

defensible. However, for items in instruments which do not involve right and wrong answers (for example, self–report forms), empirical validity must be determined, As a general rule, attempts should be made to determine empirical validity for all types of instruments. Sometimes items which appear to be acceptably valid are found to have little or no empirical validity; conversely, items that do not appear to be at all valid are found to have empirical validity.

Section III
SELECTION RATIO

91. Definition of Selection Ratio

While the validity coefficient is most important in deciding whether a test score will be useful in selecting men for some job or training course, other factors are also important. The selection ratio is one such additional factor. Specifically, the selection ratio is the number of men needed (demand) divided by the total number of men available (supply).

92. Significance of the Ratio

In general, if the selection ratio is small, that is, if there is need to select only a few from a large number of available men, test scores are highly useful. If this is the case, the cream of the crop can be selected. If, on the other hand, the need exceeds the number of men available, tests will be of no value whatever since all available men will have to be used. It will be seen below (par. 94) that the selection ratio has an important bearing on the establishment of minimum qualifying scores.

93. Application in the Army

a. The selection ratio is a highly important factor in determining the quality of men finally selected for some given job. A low numerical value of the validity coefficient may actually indicate very useful validity if the selection ratio is small. To take an extreme case, an instrument with a validity coefficient of .15 may be very effective in selecting men who will be successful on the job if only 10 men out of 1,000 available are to be selected.

b. The Army uses the basic ideas underlying the selection ratio to good advantage in selecting men for important assignments. A small selection ratio is used in selecting officer candidates. Selection of candidates for the United States Military Academy is a second instance in which the factor plays an important role. In jobs less important and involving less responsibility, lower standards and larger selection ratios are employed. Of course, it is not always possible to control the selection ratio and keep it stable since the manpower picture changes from time to time. This means that careful staff planning is essential to obtain reasonably accurate estimates of the demand for manpower and the supply—an important illustration of the close relation that exists between Army personnel policy and Army personnel research.

Section IV
MINIMUM QUALIFYING SCORES

94. Factors in Determining Minimum Qualifying Scores

In order that assignments will be made effectively on a uniform basis throughout the Army, minimum qualifying scores are set for almost all possible assignments. It can be seen from the discussion of the selection ratio that the particular standard score designated as the minimum acceptable score is to a considerable degree determined by the importance of the job, the number of men who are available, and the number who are needed. Generally, then, a minimum qualifying score for a job is set by considering the distribution of scores for the test or battery found to be most valid for that, job and deciding on the portion of those available needed to fill the vacancies that exist.

95. Conflicting Demands for Men

The minimum qualifying score cannot always be set for one Army job without taking into account the needs for other possible jobs. When feasible, the more important assignments should be filled first. Thus, while high minimum qualifying scores will usually be used when the job is important, less important jobs must be filled from those remaining. Often, fortunately, the type of man needed for one job is not the same as that needed for another. In selecting clerks, for example, most of the men selected would not be expected to do well in various jobs calling for mechanical skill and ingenuity.

96. Minimum Standards of Performance

While minimum qualifying scores are usually set as a function of the expected supply in relation to the needs for a given job, an additional factor is sometimes important. Some men are so inept at some jobs that they would contribute nothing or make errors costly enough to overbalance what they did accomplish. In other words, minimum standards must be used to avoid placing men in jobs where their probable performance is such that they will be a detriment rather than an asset. Generally, minimum qualifying scores should be set as high as the supply–demand situation allows. They should also be sufficiently high to insure that minimum standards of performance are met. Caution must be exercised

that the minimum standards are not set arbitrarily high—they should be based on careful study and not on mere desire to select nothing but the best.

97. Need for Uniform Standards

If a minimum qualifying score set on an Army wide basis is not adhered to, confusion will result. If one installation uses a minimum qualifying score of 125 for Officer Candidate School while another uses a score of 115, men scoring between 115 and 120 will be admitted to Officer Candidate School from the second installation, while men with scores of 123 are being rejected in the first installation. The quality of men entering the school will depend on the installation they came from, rather than on the instruments. Not only would the quality be uneven, but it would make it extremely difficult to maintain an effective selection system.

Section V
SOME GENERAL CONSIDERATIONS INVOLVING STATISTICAL SIGNIFICANCE

98. Necessity for Sampling

a. Data collected for study in personnel research—whether it be test scores, course grades, ratings, or personal history information—usually represent only a portion of all of the data available. In most cases, it would be expensive, difficult, and unnecessary to utilize a total population when a sample adequately representative of the total could be drawn and used. For example, if the average score on a new test for the Army population is to be determined, it might be possible to administer it to every soldier and to compute the mean of this tremendously large number of scores. However, such a procedure would obviously be time–consuming and costly; practical considerations would make it impossible as well. It usually suffices to obtain a sample of men from the total population and to use the mean of their scores as the mean that would be obtained in the total population.

b. Reasons other than time and expense sometimes make sampling necessary. For example, if in arsenal manufacturing hand grenades wants to be sure that they will explode properly, a perfect test would be to explode them. But it is obvious that after this test there would be no grenades left! In this type of testing situation, it is necessary to select a sample and to apply the test to the sample only

99. Errors Due to Sampling

Whenever a sample is drawn from a total population, it is to be expected that any statistic, such as a mean or a percentage, that is based on that sample will vary from the "true" value of that statistic. A "true" value here is used to mean the value that would be obtained from results based on the entire population. Consider again the example of computing a statistic such as the mean test score for the Army population. A sample of men might be drawn and the mean of their test scores computed. Then another sample of men might be drawn, independently of the first selection, and the mean of their test scores determined. If this process were to be repeated many times, the means from the successive samples would be observed to fluctuate. Most means would fall at, or very close to, the true mean, but a few would, by chance, deviate more. The likelihood of obtaining a mean very different from the true mean would be small, but such a mean might occur in one of a large number of samples. If a large number of these means were obtained, they would be found to be distributed around the true mean according to the normal probability law. These successive means (obtained by sampling) could themselves be treated like scores, and the average and standard deviation of the means computed; about 68 percent of them would be found to lie within one standard deviation on either side of their population mean. This standard deviation of sample means is not the same as the standard deviation of the original test scores; it is referred to as the standard error of the mean. Estimates of the standard error of a statistic usually can be made from statistics based on one sample. The standard error of the mean, thus, can be shown to be equivalent to the standard deviation of the individual scores from one sample divided by the square root of the number of scores. Discussion of standard errors of other statistics can be found in any statistics textbook.

100. Factors Affecting the Adequacy of a Sample

a. Manner of Selecting a Sample. First and foremost, statistical results based upon a sample will be affected by the manner in which the sample is selected. Bias entering into selection of a sample will be reflected in results based on that sample; the extent to which statistics derived from a biased sample can be generalized to the population is usually indeterminate. Discussion here refers to the situation where samples were randomly drawn from the total population. The conditions for simple random sampling are that each individual in the sample has an equal chance of being drawn and that the selection of one individual will in no way affect the selection of another.

b. Size of Sample. Statistics based on large samples are more likely to represent population values than those based on small samples. For example, if only two men were selected from the total Army population, the chance of a large difference between the average of their scores on a test and the total Army average would be great; however, if 2,000 men were randomly selected, the chance of an equally large difference would be very, very small.

101. Statistical Results as Estimates

Whenever any statistic, such as a mean, is computed on the basis of a sample, the question arises—how good an

estimate is this statistic of the value for the entire population? Answers to this question can be made in terms of the "standard error," described in paragraph 99. If the mean of a sample of 100 test scores is found to be 50, and the standard error of the mean is 1.5, these data may be interpreted as follows: If a number of random samples of 100 cases each were to be drawn from the original population, and the mean test score computed for each sample, 68 percent of these means might be expected to lie within one standard error of the true or population mean, 98 percent within two standard errors. Thus if an obtained sample mean differs by as much as three standard errors from the population mean, we can reject the notion that the sample was drawn from the population.

102. Test Scores as Estimates

In basing decisions on a man's test score, a practical question arises—how much meaning can be attached to a given score? If the man were to be retested a number of times, what fluctuations might occur in his score? How likely is a given test score of 105 to become 106 on retesting, or 103, or even 120? For What reasons might these variations occur? Fluctuations of this type, although they are called "errors," axe somewhat different from the fluctuations or errors discussed in paragraph 101. Rather than being due to the effects of simple random sampling as in the previous examples, variations in individual test scores are due to lack of refinement in the measuring instrument. The error of measurement here is of the same type that exists in physical measurements—length can be measured by a yardstick, by a foot–ruler, or by a micrometer, but with each type of measuring stick, successive readings will vary, the degree of variation depending on the refinement of the measurement made. In the personnel testing situation, another way of looking at these errors of measurement is that they are errors of a "mistake" type, where the examiner did not happen to time the test precisely, where unusual distractions might be present, where the examinee's motivation was exceptionally poor, or various similar circumstances existed. The "standard error of measurement" is used to represent the variation in measurements obtained with a particular test. The extent of this error is a function of the reliability of the test as a whole and also of the total expected variation in the test. From this discussion of types of errors of measurement, a man whose true score on a test is an Army standard score of 90 might well obtain a score somewhere between 85 and 95.

103. Significance of Differences

In personnel research studies, it is often necessary to decide whether observed differences reflect true differences or merely chance errors in the values being compared. Need for evaluating the differences may arise in the following situations: comparing average grades of successive classes in schools, comparing validity coefficients of tests in order to select the best predictor, comparing reliability coefficients of tests, selecting the best method of job training by comparison of average performance scores of the groups taking training. For example, consider two groups, A and B, with mean scores on a test of 105 and 115, respectively. Assuming that groups A and B are random samples from their respective populations, it is assumed that there is no difference between the means of these populations. Then, using the samples A and B to estimate the standard errors of the sample means, it is possible to compute how often the observed difference in means would occur by chance if in fact the assumption is true. Suppose in the example that the standard error of each sample mean is determined to be 2.1. If the population means are the same, then the probability that random sampling fluctuations would produce a difference of 10 points or more between the sample means turns out to be .0001. In other words, the odds are quite small—one in ten thousand–that chance alone led to the observed difference. The assumption of no difference is then not tenable, and it is concluded that the difference between the means of groups A and B is "statistically significant." Such significance depends essentially upon the fact that the standard error of the mean is small compared to the difference between the two group means. Thus either a large difference or a small standard error will lead to significance. It is worth noting that a large number of scores will lower the standard error, so that, in a large enough sample, even a very small difference might be found to be statistically significant. The topic of practical versus statistical significance is discussed further in paragraph 104. Elementary statistics textbooks should be consulted for further development of these ideas and for explicit formulas. It should be kept in mind that if a particular statistical comparison indicates no significant difference, it does not follow that the true measures are identical. It may well be that there is a true difference which the samples that were drawn failed to reflect on this occasion.

104. Practical Significance as Compared With Statistical Significance

Test scores, as well as research results, must be interpreted not only in terms of their significance as computed by the best statistical tools available, but also in terms of elements in the real situation in which they are being used. Sometimes a difference between two values which has been shown to be "statistically significant" may have to be ignored on practical grounds. For example, the average performance rating of men who have undergone training method A may be superior to that of men under method B; the probability may be almost zero that the difference could have arisen by chance; yet, factors such as cost, time, convenience, or availability of equipment may more than offset the gain in performance. A test may be built to yield extremely refined differentiation's in possible scores, with a very small standard error of measurement; yet this refinement of measurement will be useless if all that is needed in a given situation is to lump examinees into two general categories, such as those acceptable for training and those not acceptable. Further, it is sometimes true that the real situation does not permit obtaining data on a sample of men

drawn from the population according to desired sampling plans; allowance must be made for this in the interpretation of research results based on the data. Statistical results must always be applied in terms of the real situation. Statistical techniques are tools to aid in dealing with practical problems and can be used only as such. Their usefulness must be related to the practical aspects of personnel problems.

Section VI
SUMMARY

105. The Interpretation of Test Scores

a. The validity of an instrument must be known before the scores can be properly interpreted. Validity is, however, specific to a given job or assignment; a test may have different validates for different purposes.

b. The selection ratio—the number of applicants needed divided by the number available—is an important consideration. When the selection ratio is small, men with high scores can be selected and a low failure rate on the job is to be expected. When the ratio is large, selection is less efficient and the on–the–job performance of those selected approaches what would be obtained if men were selected at random.

c. A minimum qualifying score is used to set Army–wide standards for particular assignments. The number needed by the Army for that assignment and the number available were shown to be the two most important considerations in deciding upon a critical score. The need to avoid selection of men whose performance will probably be completely inadequate is an additional factor in setting minimum qualifying scores. Great care must be exercised in determining what completely inadequate performance is.

105B. Title not used.

Paragraph not used.

Chapter 6
USE OF APTITUDE MEASURES IN INITIAL CLASSIFICATION

Section I
INITIAL CLASSIFICATION

106. Purpose

In the course of his first days in the Army, an enlisted man goes through the process of initial classification. At its close, a recommendation is made as to the military occupation in which he will be assigned. The purpose of Army initial classification is to identify the abilities of the men in the Army manpower pool available for assignment. These skills and abilities must be matched with requirements of Army jobs. Because requirements of Army jobs today comprise a complex manpower structure, more careful classification of men is needed. Not only is it necessary to place a man in a job he can do well or learn to do well, but it is important that his abilities and skills not be thrown away, i.e., they must be used in jobs for which the Army has a definite need. And, to insure optimum utilization of manpower today, not only must the Army consider what a man can do, but also what a man will do. Thus, personal characteristics as well as individual aptitudes must be measured and taken into account in the classification process. Approaches to this problem are discussed in chapter 9. Fortunately, as a result of the Army's sponsorship of a research and development in–service program during the past two decades, more is known today about what makes a soldier successful on the job, what abilities are needed to meet the requirements of Army jobs, and what mental abilities are measurable.

107. Basis for Initial Classification

a. Before the choice of assignment for an enlisted man is made, two kinds of inventories are necessary. First, there must be an array of facts about jobs in the Army: What jobs have been set up as Army jobs? How many jobs of each type are to be filled at this time? Which jobs must be filled immediately? How many men have to be trained for the various jobs within a certain time? In regard to each job, such questions as these must be answered: What kind of physical effort must a man be capable of to do it? What skills must he have or will he have to acquire? How much will he have to learn? How difficult are the problems he will have to solve?

b. The man, too, must be inventoried. This inventory has one main objective—to obtain advance estimates of his probable success in different kinds of Army jobs. The classification officer who finally makes a recommendation for the man's assignment will have at hand all the information that has been collected about the man in his few days of processing. The classification officer knows what the man's physical status is. He knows how much formal education the man has completed, what games or amusements he likes, what jobs he has held, how long he worked in each one, as well as what kind of work the man thinks he would like to do. And, most important, the classification officer has at

hand estimates of how well the man can be expected to do in a number of occupational areas, these in the form of scores on the battery of aptitude tests the man has just taken.

c. The process by which all of this information is obtained from the enlisted man and applied in initial as well as in later phases of classification is described in a series of Army publications including AR 611–203. It is the objective of this chapter to explain how personnel measurement techniques are employed to increase the effectiveness of this classification.

Section II
APTITUDE AREAS

108. What Differential Classification Is

a. Army jobs cover a wide enough variety of functions that it is possible to take advantage of the fact that people have more potential to perform some jobs than others. When the skills required in an occupation are among the things a man does best or can learn to do most easily, the job is likely to be well done. In addition, other jobs which are of the same level but which require different abilities can be done better by other men whom skills and abilities, attained or potential, fit those jobs. The principles of differential classification are practiced by the Army as the means by which a man's strong and weak points are considered in recommending him for assignment to a certain job or training for that job.

b. Differential classification usually employs results on a battery of tests which provide estimates of a man's probable success in a number of different occupational areas. This makes it possible to we in which fields he is likely to do best and in which fields his prospects of success are relatively poor.

109. Definition of Aptitude Areas

a. The Army Classification Battery. The set of tests used to evaluate the capacities of enlisted men during initial classification is the Army Classification Battery. Each of the tests in the battery is intended to measure a different aptitude or skill important in one or more types of Army jobs. The battery is typically made up of a number of comparatively short tests that measure verbal, mechanical, radio code, and clerical aptitude tests of arithmetic reasoning, pattern analysis, and shop mechanics, tests of general, electronics and automotive information, and a personal inventory.

b. What an Aptitude Area Is. In use, the Army Classification Battery follows a somewhat different pattern from that of the typical selection battery developed and validated as explained in chapter 2. The tests in the Army Classification Battery are not employed as successive hurdles. Nor are they used as a single test composite to provide an overall measure of the individual. Instead, they are used in a number of different combinations, each made up of two tests. The score on each pair of tests represents a combination of abilities which is considered important for satisfactory performance in a number of Army jobs. The combinations of tests, together with the group of jobs for which each combination predicts success, are the aptitude areas, each of which is associated with varying numbers of Army jobs. For example, an estimate of probable success as a clerk is based on a composite of scores on the verbal and clerical speed tests. Associated with this same combination of tests are other jobs such as postal clerk, cryptographer, and personnel specialist, which at first consideration would seem to have little in common. The method of combining tests to predict performance in a job area is described in paragraph 110c.

c. Changing Character of the System. In thus defining aptitude areas, two points should be emphasized. Both have to do with the developing nature of the system. In the first place, it is recognized that, as research on the tests goes on and as Army jobs change, new tests are profitably added to the battery or substituted for some tests now in the battery. Secondly, changes occur in the clusters of jobs basic to aptitude areas, and jobs may shift from one area to another. Such changes follow upon a continuing analysis of aptitude areas in operational use. Validation studies are undertaken to check on how well the aptitude area scores indicate the relative success in performance or training in the case of specific jobs. At the same time, it is possible to try out as predictors for those jobs combinations of tests other than the one with which a job is now associated. Research is continuing toward the development of tests which are more independent of each other and provide measures of more clearly defined traits. This might mean eventually a battery composed of a greater number of short tests, each of a highly specific ability, so that there will be decreasing overlap among aptitude area test scores.

110. Development of Aptitude Areas

a. Limitations of a Single Overall Score in Classification. In the early days of group testing on a wide scale in the Army, main reliance was placed on a general measure of ability, scores on which were widely used, not only in allocation of men to fill job quotas, but in selection for specialist school and officer candidate training. The test sampled several areas of ability, but only a single overall estimate of the enlisted man's trainability was obtained. Use of a single measure in classification was adopted in response to the pressing need to fill jobs with men who could do them. The simplest way of getting a job done satisfactorily, it was felt, was to put in a job that man who had at, least enough general ability to learn a job of that level. It was understood, however, that in a way this was a wasteful solution. Mental aptitude is not a single ability; it is a collection of abilities. Each man has these abilities in varying

degrees more of some, less of others. A score on a single overall test of a soldier's aptitude is but a general average of all his abilities. All the Army would find out about a man, when using such an overall measure, would be where he stands in relation to the rest of the manpower pool with respect to his job potential in general. To classify a soldier adequately for one of the current host of Army jobs that must be filled, the Army has to know not only the man's general level of aptitude but also his particular pattern of abilities—what he can do best, how many outstanding abilities he has, how he compares with others in each ability, and what he will do best.

b. Steps Toward Differential Classification. Research information on the abilities of Army personnel and a better understanding of the requirements of Army jobs led in 1949 to the implementation of a more comprehensive classification system. Instead of a single classification test, ten tests comprising the Army Classification Battery were introduced. Each of the tests measured a different group of abilities, yielding a more comprehensive pattern of abilities for each soldier. Since 1949, research and development activities have resulted in increased knowledge concerning how each ABC test can make an even more unique contribution to the classification of enlisted men. Since October 1958, eleven tests have comprised the ABC, including two new tests of motivation and interest pertaining to expected success in the combat arms.

c. Formation of Aptitude Areas. A set of test scores identifies abilities. This information has to be matched with job requirements. The combination of abilities required in one job area is different from the combination of abilities required in other job areas. Various combinations of the scores on the eleven tests of the Army Classification Battery are therefore needed to accomplish the classification of soldiers. Combinations of Army Classification Battery test scores are called aptitude areas. An aptitude area score is the combination of test scores which provides the best available prediction of success in a particular military occupational specialty. Today there are eight aptitude areas. When this aptitude area system was introduced in 1949, the amount of research information available on classification procedures argued for a conservative approach—as many aptitude areas were established as then appeared to be necessary for adequate differentiation of the combinations of abilities required by the Army job structure.

d. Continuing Research Keeps Aptitude Areas Current. Since the introduction of the original aptitude areas, the system has been under constant study. Aptitude area scores of Army personnel have been compared with measures of success in a variety of Army military occupational specialties. When enough additional research information is available, the aptitude area system is revised and the effectiveness of the Army classification system increased. Revisions are not based on guesswork or armchair strategy; they are based on years of research by Army research scientists and on years of operational experience with the previous classification system.

111. How Aptitude Areas Are Used in Classification

a. When the classification officer approaches the problem of evaluating the man and deciding upon his initial assignment in the Army, he has at hand a "profile" showing the man's aptitude area scores. This shows at, a glance the area or areas in which the man has the best prospects of success. It can be related directly to the types of job he is likely to do best. By considering this information against the jobs he has to fill, the classification officer can usually narrow the choice of assignment down to a reasonable number of alternatives.

b. Selection of initial assignment comes from a synthesis of all the information collected about a man. With different classification problems, different aspects of the body of information come to the fore. In some cases, a man's education or lack of it will determine whether he is assigned to certain types of training. In every case where an enlisted man has gained competence in some civilian occupation, this experience outweighs other indications of his probable success. Even though his score in that aptitude area indicates that he would require much longer than the average training period to learn a job, if such training has already been effectively accomplished while he was a civilian, the Army can take advantage of it.

c. Even when Army needs do not, interfere, assignment is not always made in the area in which a man made his highest, score. The practice followed, where possible, is to make the assignment in an area in which he made one of his higher scores. The classification officer often finds it necessary to decide between several aptitude areas in which the man has made almost the same score. Such factors as physical profile, school quotas, and work experience then assume relatively greater weight in determining recommendation for assignment. Still another factor must be considered, namely, the level of ability required. A man's best score may be in an aptitude area requiring a high level of ability, and his second best in an area requiring a, lower level of ability. It is possible that he will perform better in the area requiring a lower level of ability.

112. Gain Resulting From Classification by Aptitude Areas

a. General. Initial classification based on aptitude areas provides a systematic approach to estimating likelihood of success in various job areas and makes it less likely that a particular strong point in an individual will be overlooked.

b. How the System Is Used. Each soldier in the Army is given the tests of the Army Classification Battery. From his test scores, his aptitude area scores are computed. Classification of the soldier considers his highest aptitude area scores; he is thus designated for an occupational area in which he will make use of his highest abilities. In figure 14 are graphs of the aptitude area scores of two soldiers. Soldier A's highest aptitude area score is on General Maintenance; his scores on Electronic and on Motor Maintenance axe also relatively high. The occupational arms that are most promising for this pattern of aptitude area scores are the following: Precision Maintenance, Military Crafts, Electronics,

Electrical Equipment Maintenance, and Motor Maintenance. Soldier B has a different pattern of abilities. Soldier B's highest aptitude area score is Clerical; based on this aptitude area score, his most promising occupational area is Clerical. In this way aptitude arms are used to match the specific abilities of the men in the manpower pool with the skills and abilities required by the Army jobs that need to be filled. In the classification procedure, a thorough review is also made of each man's physical profile and work experience before the most appropriate occupational area is specified.

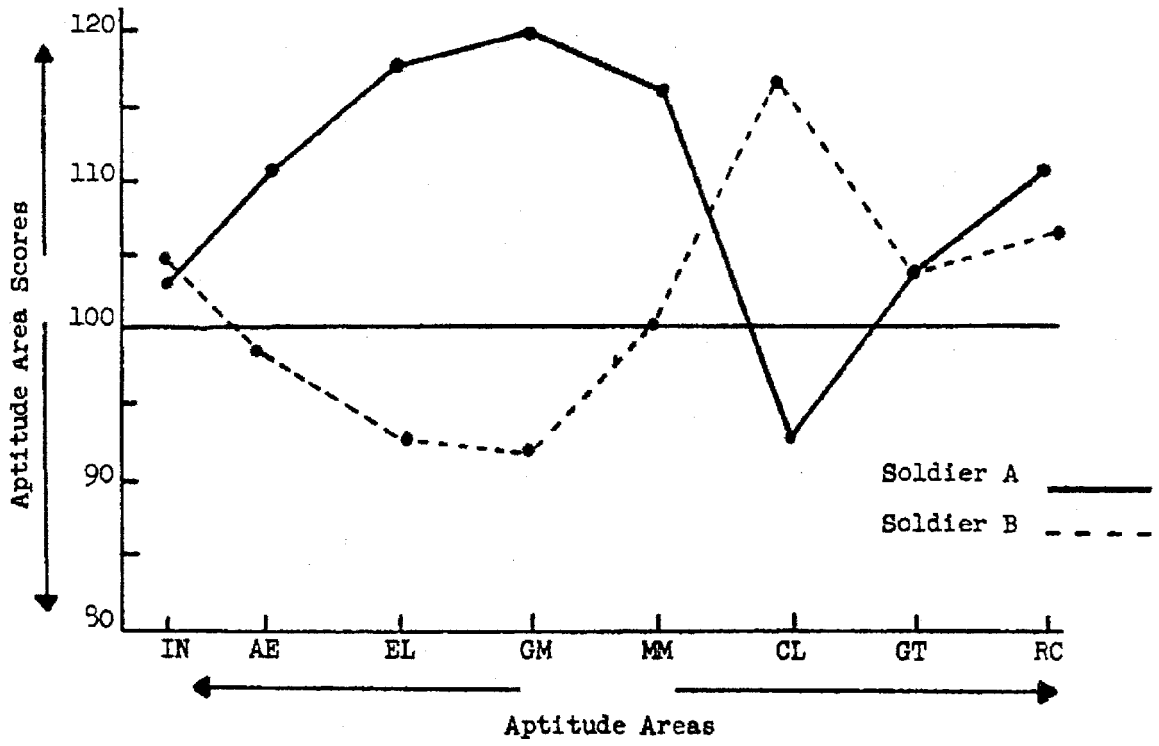


Figure 14. Aptitude Area profiles for two soldiers

c. Advantages to the Army. How well does the aptitude area system do its job of measuring abilities important to Army enlisted MOS? Continuing research over the past decade has followed up the careers in training and on the job of over 40,000 men who have been assigned to more than 80 different MOS representing all occupational areas. These studies have shown the extent to which these aptitude combinations have identified abilities that are relatively unrelated to each other, but which are highly related to training and on–the–job success.

(1) Under the current system, the eight aptitude area scores represent abilities different enough from each other to provide a very good chance that any man will have at least one ability score which is above the average (Army standard score of 100). The more genuinely different the kinds of ability the ABC can measure, the better the chances that each individual's best potential will be high. On the basis of large samples of enlisted input it was determined that, on the average, 82 percent of a typical cross–section of the manpower pool had at least one aptitude area score of 100 or higher. For a similar cross–section, on a single general mental ability measure, the AFQT, only 53 percent had an Army standard score of 100 or higher (fig. 15).

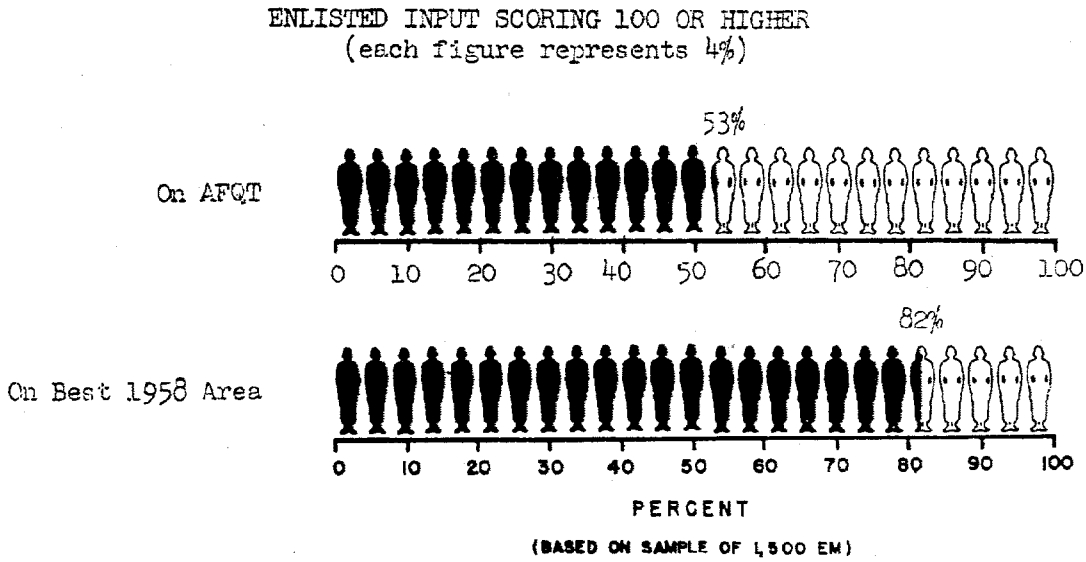


Figure 15. Percentage of enlisted input with standard scores of 100 or higher on AFQT and on best 1958 aptitude area

(2) As the Army develops and puts into operation more complex equipment of all kinds, the need increases for men able to succeed in specialized technical training programs and able to work on highly technical jobs. The payoff on the aptitude area system is the extent to which the system enables the Army to predict how well each man would do in MOS in each occupational area. On this basis, the Army can assign each EM where he is likely to contribute most to the Service. Therefore the important question is: When men are placed in training and in jobs appropriate to their highest aptitude area scores, how well may they be expected to perform in the wide variety of combat and technical MOS training courses? Assuming that a prerequisite of 90 on the appropriate aptitude area is set for admission into training, it has been determined that for combat MOS, approximately 72 percent will perform at levels determined to be "acceptable" or better and that for technical MOS, approximately 79 percent will perform at such levels.

113. Problems in Classification by Aptitude Areas

a. Priority Assignments. For differential classification, it is necessary also to consider the relative importance of the various job areas. In order that certain important jobs may get enough high caliber men, individuals with lower scores may have to be used. Then, there are certain military jobs to which assignment is made on a first priority basis. That is, if a man meets minimum qualifications for the job, he may be assigned to it even though he is more able in other areas.

b. School Quotas. The filling of quotas for training service schools presents an allied problem, since quotas are set on estimated future Army needs. The necessity for making the required distribution of incoming enlisted men to schools sometimes forces assignments which are contrary to the principles underlying the aptitude areas.

c. Problem of Timing. If assignments could be equalized over a period of time, it would be possible to make a more consistent application of the aptitude area system. Unfortunately, it is not possible to control the distribution of abilities of men coming into a classification center at any one time. Nor is it always possible to anticipate those Army's needs which arise suddenly—and often need and qualified supply do not come at the same time. In these circumstances, optimal differential classification is not always possible.

d. Future of the Aptitude Area System. Continuing research on classification instruments and procedures is essential to maintaining and improving the effectiveness of the aptitude area system. Currently, research is under way to develop instruments that will differentiate between certain subdivisions of the present occupational areas, for example, Electrical from Electronics MOS, Construction from other General Maintenance MOS, Medical from other General Technical MOS. Second, research on personality and motivational measures has been undertaken to identify characteristics which predict what a man will do on the job as contrasted to what he can do. Studies tire under way to see if known mathematically optimum systems can be applied to the Army situation. The optimum solution has been known for some time and has influenced the classification described in this chapter. Certain features of this optimum solution cannot be presently used, however, because of computing problems arising from the complexity of formulae involved.

At the present time a system for assignment of individuals to jobs through the use of a large electronic computer is being developed. With this system the possibility is open for considerable increase in the efficiency of classification and assignment. Many practical problems remain in adapting to a computer system, but once this adaptation has been achieved, it is expected that continued improvements in the systems will be feasible. From these and other research endeavors, the aptitude area system may be expected to develop in the direction of increased effectiveness in differential prediction of MOS performance and in increased usefulness for both broad allocation and special selection of EM to meet the changing requirements of Army jobs.

Section III
SUMMARY

114. Aptitude Measures as Aids in Classification

a. Initial classification attempts to make the best fit between the capacities of the enlisted men on one hand and the requirements of the jobs to be filled on the other.

b. The Army Classification Battery of objective tests was developed for use in evaluating the capacities of enlisted men for initial classification. With development of aptitude areas by combining two or possibly more of the tests, and relating them to job areas, differential classification has become possible.

c. Differential classification involves matching the requirements of various job areas with the estimated strengths and weaknesses of men to be assigned. Other factors that must be considered are the relative importance of the job areas and other available personnel information.

114B. Title not used.

Paragraph not used.

Chapter 7
ACHIEVEMENT TESTS

Section I
CONSTRUCTION AND EVALUATION OF ACHIEVEMENT TESTS

115. What Achievement Tests Are

a. Tests whose purpose is to estimate in advance how well a man can learn a job are called aptitude tests. Tests whose purpose is to estimate how much a man knows about a job and how well he can perform the job are called achievement tests.

b. A more detailed discussion of the distinction between aptitude and achievement tests has been presented in chapter 2. A summary of the characteristics and uses if achievement tests will be presented here.

116. Types and Characteristics of Achievement Tests

a. Paper–and–Pencil Objective Tests permit economical coverage of large areas of knowledge; they are economical to administer and score in large numbers; the scoring is not subject to the scorer's varying judgments; they can be standardized.

b. Work Sample Performance Tests are used where paper–and–pencil tests are not available or where it is essential to determine how well a man can manipulate tools and equipment instead of estimating it from the knowledge he shows on a paper–and–pencil test. They may be used as individual tests after previous screening by group administered paper–and–pencil tests, and with illiterates and non–English speaking men. However, performance tests are not adapted for administration to large groups at a time; they are expensive to construct and administer and they are difficult to score objectively (par. 119).

c. Ratings as Achievement Tests. The usual achievement testing procedures may not yield satisfactory measures of acquired work habits and attitudes and skill in dealing with people. Ratings by supervisors and peers are customarily relied upon to measure how men compare with one another in these qualities. Ratings are deceptively simple—unless great care is used, they may be worthless. For use as measures of achievement, ratings should be in terms of concrete situations that are significant for specific purposes, rather than in terms of generalized or abstract qualities. Other requirements for effective rating procedures are discussed in chapter 10.

117. Requirements for Job Proficiency Measures

What is measured to determine proficiency for a job will depend upon an analysis of the job itself. The job may call for not only knowledge and skills but certain work habits and attitudes and certain personal qualities which may be equally important.

a. Possession of Pertinent Information. For any job and for any training program for that job, certain information is

fundamental to adequate performance. In most military jobs, the man needs to know specific facts concerning the job to which he is assigned. But this is not enough. He also needs to know certain facts about jobs related to his own and about the functions and organization of the activity in which he works. Possession of pertinent information is not the only requirement of competence on the job, but unless the man possesses that, he is not likely to be as useful as other men.

b. Accessory Skills. As used in connection with Army achievement tests, skill refers to practical effectiveness in performing a task. In defining the skills required in qualifying for a job, it is important to describe not only the action to be performed or the work product to be achieved, but also the typical situations in which the skill is to be displayed. It makes a great difference, for example, whether the clerk typist who is expected to turn out neat and accurate copy is to work from a typed rough draft with corrections noted, from a longhand manuscript, from shorthand dictation, or from a sound recording. While the action in using the typewriter is the same, the accessory skills are different and the rates at which the transcription can be done may be different.

c. Ability to Apply Knowledge and Skills. Aside from routine operations, practically every job in the Army requires some ability to solve problems. Changing situations, new materials, adaptation of tools and equipment, modified procedures, defective materials or tools—these are some of the conditions presenting problems which the soldier must solve by drawing upon his knowledge and skills. To measure this ability means to determine how well the men solve problems typical of those encountered in the normal course of duty. Making such measurements requires the development of standardized problem situations in which the men to be tested will have considerable freedom of choice in each critical stage. Both the development and administration of such achievement tests may be complex and time–consuming.

d. Possession of Suitable Work Habits and Attitudes. Personnel managers in industry have long recognized that a major factor in employee turnover, both discharges and resignations, is faulty attitude toward work or inability to get along with fellow employees. In the Army, the equivalent of employee turnover is frequency of transfers from one unit to another. Often men who ask for transfer, or who are recommended for transfer in advance of their normal periods of rotation in assignment, are those who fail to adjust to their duties. Examples of questions relating to suitable work habits and attitudes and ability to get along on a given job are—Does the man feel that the job he is doing is worthwhile? Does he take personal responsibility for observing the rules of the job, such as safety precautions and cleaning up his place of work? Does he have the initiative to learn and undertake work related to his assignment when there is need for it? Does he accept suggestions? Does he take over in emergencies? Does he supervise effectively? Can he work well in a team? The answers to such questions cannot be obtained from the usual kind of achievement test. Instead, it is necessary to depend on such techniques as ratings and interviews.

118. Planning Achievement Test Content

a. What is actually selected as test content will depend, as has been stated, upon a careful analysis of the job as well as on the suitability of the measuring instrument. Techniques for analyzing occupations, jobs, tasks, and work operations in terms of the functions performed are applied. These functions are then translated into qualities which the worker must possess in order to perform them satisfactorily and a preliminary estimate is made of the relative significance of each quality.

b. Once it has been decided what the achievement test should measure and whether the need for the test is sufficiently great to justify the expense of construction, the steps described in chapter 2 are followed. A test plan is prepared to make certain that there will be adequate coverage of the important aspects of the job. Items are constructed to sample these aspects reliably and to provide useful levels of difficulty. The test is then ready for field testing.

119. Are Paper–and–Pencil Tests Practical?

The statement is frequently heard that the work sample performance test is a more practical test—that is, that it bears a more apparent relationship to what is being tested—than the paper–and–pencil test. The statement may have some justification. However, the work sample test may not be as practical as it appears. For one thing, a relatively small sample of the job usually makes up the test, although what is examined on the test is examined intensively. In addition, work sample tests are usually harder and more expensive to construct and to administer efficiently on a large scale than are paper–and–pencil tests. Scoring is more difficult. The paper–and–pencil test, on the other hand, can be administered on a group basis fairly easily. It usually emphasizes a broad and extensive sample of the job content rather than an intensive one. Its practicality may not be apparent but, through field testing, its practicality can be determined and improved if necessary.

120. Evaluating Achievement Testing

The construction of achievement tests is not an end in itself. Such measuring instruments must be evaluated to determine whether they suit the purposes for which they are developed. This chapter will consider the characteristics of objectivity, reliability, and validity as they apply to achievement tests.

121. Objectivity of Achievement Tests

a. For an achievement test to be objective it must be possible for anyone to administer and score the test and obtain

the same score that any other scorer would obtain. Objectivity is important in obtaining equivalent scores from examinees who have done equivalent work on a given test, thus enabling comparisons to be made among examinees. To this end, methods of administration and scoring are rigidly prescribed. Objective items for achievement tests are constructed in such a way that there is, among persons who are supposed to know the correct answers, universal agreement as to what is actually the correct answer. Essay tests are relatively easy to construct and the temptation is strong to rely upon them. The correctness of the answers, even if a scoring guide is used, is not always easy to determine. Furthermore, scoring of essay answers is influenced by how well the examinee writes, and not by the correctness of his answer to the question asked. There is some mistrust of the notion of testing recognition of facts and details through objective items rather than recall, reproduction in writing, and the so–called "higher mental processes" of the essay types. This need not be—proper construction of test items can measure the more complex characteristics and provide objective scores. Whatever might be said of objective tests, there is abundant practical evidence that objectivity is basic to the construction of reliable and valid achievement tests.

b. In general, the more chance permitted for the exercise of judgment or discretion on the part of the scorer, the smaller chance there is for obtaining a reliable score. Since objective tests do not require the exercise of judgment in scoring, this is not a source of unreliability for them. One practical point should be kept in mind—large scale scoring would not be possible in the Army if objective tests were not used. To provide enough competent judges to score essay tests would require a small army of judges.

122. Reliability of Achievement Tests

To be reliable, an achievement test should—as should other types of measuring instruments used in personnel actions—yield about the same scores on repeated administrations of the same test or on alternate forms of the test. The techniques for determining and increasing reliability have been discussed in chapter 4.

123. Validity of Achievement Tests

In testing achievement, the test constructor is called upon to select the achievements which are most likely to be significant for the purpose specified, and to develop valid measures of these achievements. Validity of achievement tests may take various forms, and frequently the statistical demonstration is not complete. Some of the recent, thinking on demonstrating validity of achievement tests is given below.

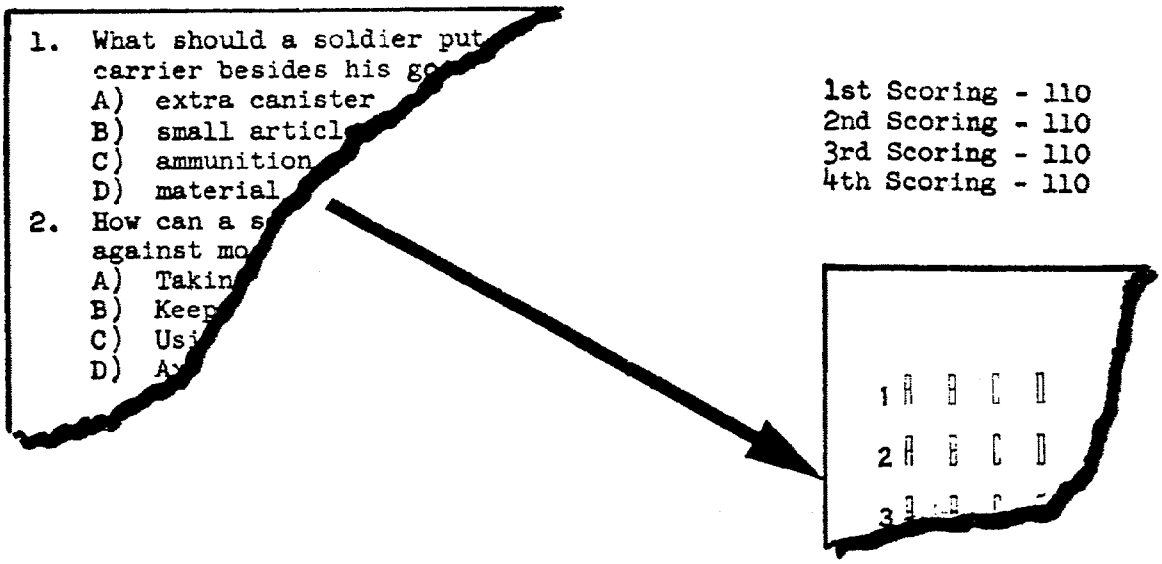
a. Test criteria are difficult to establish for many qualities in achievement measurement. Sometimes statistical or empirical validity is estimated for a new test by correlating scores on it with scores on other tests which experience has shown to be satisfactory. This procedure, how ever, is of uncertain value. The new and the old tests may be correlated, but the criteria they predict may be quite different.

b. The test constructor frequently attempts to build validity into his achievement tests. Constructing achievement tests is a long process and there frequently is not time to conduct a field test to determine validity. For some jobs, there may not be enough men assigned in the occupation to provide an adequate sample on which to validate the items prior to the official use of the test. The test constructor then relies on his seasoned judgment of what will constitute a valid test (ch. 5). This means the test must be designed and developed in such a way that there can be reasonable expectation that it will really measure the knowledge or skill which it is intended to measure. The test is a systematic sampling of the kinds of knowledge and skill the test constructor and experienced operating personnel consider crucial for differentiating among examinees. Much of the case for validity of such tests rests, then, not on a computed coefficient of correlation, but on an inferred correlation based on a description of job knowledge or training requirements, the representatives of the items, and insight with which the items have been selected.

c. Constructing achievement tests may be likened to the development of many kinds of criteria (ch. 3). Subject matter experts provide the basic information on what is right or wrong, or good or bad. In the absence of empirical validity data, it is important that achievement tests have adequate reliability, a proper range of difficulty and adequate differentiation among men at various score levels.

d. When conditions permit, the achievement tests may be tried out in the field. The forms of the test are administered to a sample to determine their range of difficulty and to determine their internal consistency (ch. 2). The results are analyzed and the test is revised, if necessary, to make certain that the difficulty of the test is suited to the population to be tested and that the test has sufficient reliability. With proper difficulty level and sufficient internal consistency, the built–in validity may reasonably be expected to result in an empirically valid test.

A STANDARD ANSWER TEST WILL GIVE A MAN THE SAME SCORE EVERY TIME SCORING IS REPEATED



BUT AN ESSAY TEST MAY GIVE HIM A DIFFERENT SCORE WHEN SCORING IS REPEATED

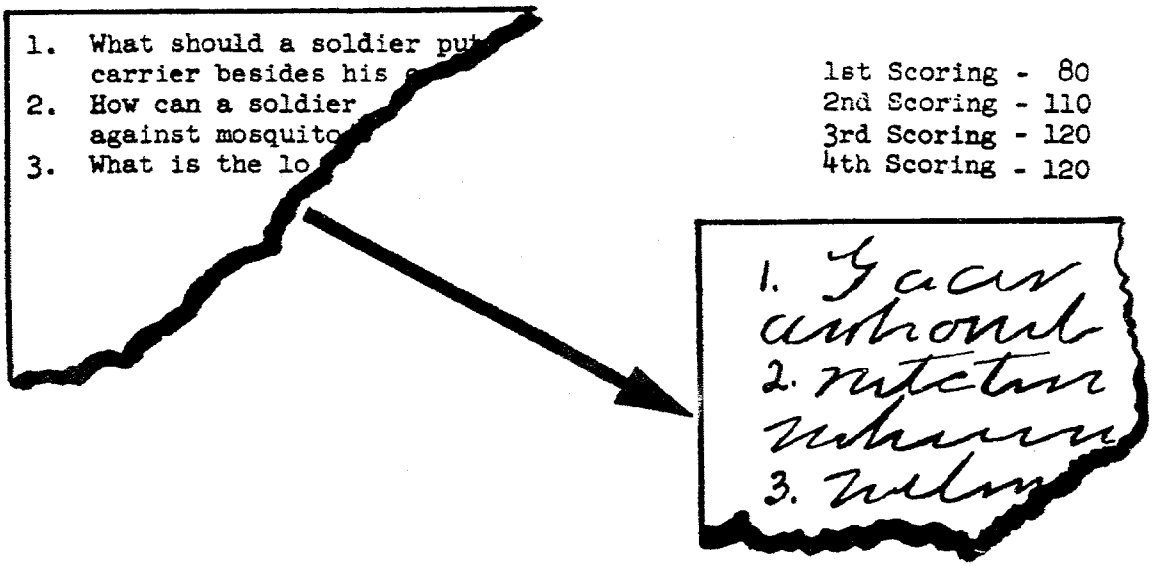


Figure 16. Comparative objectivity of standard answer tests and essay tests

e. In training courses of the classroom type, where there are usually not enough trainees available for a trial administration, it is not uncommon for classroom tests to be designed, developed and issued for use without benefit of any experimental administration. Whatever validity the test may have is primarily dependent on the judgment exercised in choosing questions, problems, and situations which will maximize the likelihood the test will actually be valid in the empirical sense.

f. Other factors relative to the construction and administration of achievement tests may have bearing on its validity. Tests of knowledge, for example, should not be tests of the ability to read. If too much material for the time allowed is included in a test measuring, for example, knowledge of light weapons operation, the test may become, in part, a test of how fast a man can answer questions. Or a test, by including very complicated items, may actually measure, in part, the examinee's ability to understand the complicated directions rather than what he knows about a given subject. Such tests may be completely invalid.

Section II
USES OF ACHIEVEMENT TESTS

124. Classification of Army Personnel

Screening and selection instruments are essential to the most effective assignment and promotion of personnel, and achievement tests can play an important part.

a. Screening is done when the administrative problem is to identify all those in the available supply of personnel who meet the minimum qualification requirements for a given type of duty. As mentioned in the first section, screening tests are frequently tests measuring the amount of knowledge and skill achieved up to that time.

b. Selection techniques may be used in the following situations:

(1) When the problem is to choose from among those available a required number who will be best qualified to perform a duty, or who are most likely to benefit from a training program, selections may be made from a screened group.

(2) In order to assign men who are partially qualified for a specific assignment or type of duty to an accelerated training program. Particularly during mobilization, the services bring in large numbers of men who present varying levels of technical proficiency in civilian occupations. Many of these occupational skills are directly convertible to military jobs. Others can be used to a considerable degree in the operation and maintenance of military equipment that has no counterpart in civilian industry. Achievement testing may be used to help verify an individual's report that he has the knowledge and skills necessary to perform a given job adequately.

c. Promotion is a typical administrative problem in which both screening and selection measures are used. The Enlisted Evaluation System offers an example of Army–wide use of achievement tests for purposes of promotion. The Enlisted Evaluation System is the method for determining the competence of enlisted personnel in their primary military occupational specialty. Evaluation scores attained under the Enlisted Evaluation System indicate relative standing of individuals classified in the same 4–digit MOS. These scores are used as a basis for various personnel actions, including promotions and proficiency pay increases. Under the provisions of the Enlisted Evaluation System, all Regular Army personnel in pay grades E–4 or above are evaluated. A separate MOS evaluation is generally prepared for each skill level of an MOS. Evaluation tests include questions on all duty positions within a skill level and may be paper–and–pencil tests, performance tests, or a combination of the two. Each such test is designed to measure the knowledge and skill a man has already acquired about the MOS in which he is being evaluated. It is assumed that those individuals whose evaluation score indicates high relative standing will be more likely to merit promotion, proficiency pay, or other favorable administrative action. Needless to say, achievement tests are not used as the sole determiners of promotion. For example, the evaluation score is weighted to take into account not only the proficiency test score but the Commander's Evaluation Report, which is a rating based upon the individual's current duty performance.

125. Use of Achievement Tests in Training Programs

Achievement tests may be used for a variety of purposes in training programs. If training programs are not accomplishing their objectives, it may be necessary to administer tests to determine if the trainees are learning what they are supposed to learn.

a. Adjusting Instruction to Different Levels of Achievement. Achievement tests may be used to divide training classes into two or three groupings representing different levels of previous achievement. The instructor can then adapt the instruction to the level of the group.

b. Diagnostic Testing for Training. Sometimes coaching is necessary to provide special background knowledge and skill. For example, a trainee may demonstrate a high degree of background knowledge and skill necessary for an infantryman but may be very weak in the mathematical skills required to use maps effectively. In fact, an entire group

may be lacking in this knowledge. Achievement tests may be used to discover specific areas of background knowledge and skill which require intensive teaching.

c. Measuring Progress in Training.

(1) A common use of achievement tests is to measure progress of trainees. If a well selected group shows a weakness on a test, the instructor may suspect that his instruction in that area has been inadequate. He may discover that the group understands a particular area and can thus spend less time on it. And of course he can discover the weaknesses and strengths of individual trainees.

(2) Not all subject–matter tests should be used as measures of progress or achievement. Sometimes such tests are primarily training aids—they need not be built according to the principles of good achievement testing if they are used primarily to arouse interest or serve as talking points. Short informal quizzes that make no attempt at adequate coverage may be used for this purpose. Generally speaking, they should not receive as much weight in determining achievement as do the larger periodic tests.

(3) As measures of progress, the larger periodic tests should serve to inform the trainee as to his progress and to motivate him to improve if necessary. This means that he should be informed of the results of the test as soon as possible—not only his total score but also his answers to individual questions.

d. Measuring Achievement at End of Course. Final examinations provide a basis for deciding which trainees have attained a satisfactory level of competence for the jobs for which the training was provided. The scores achieved also serve as an indication of how effective the course has been in training men to do the job. Accordingly, final examinations should be directed at the purpose of the training program, rather than at its content. They should show whether the man can use what he has learned in practical situations. Knowledge of facts, principles, or procedures may be essential to competent performance, but more possession of knowledge is not a guarantee that the individual's performance will be competent.

e. Comparing Effectiveness of Various Training Methods.

(1) An achievement test may be used as a criterion to determine the relative effectiveness of various training methods. Merely trying out various methods is of little use unless an acceptable standard is used to evaluate them.

(2) Suppose, for example, it is desired to determine which of several methods is best for teaching recruits the nomenclature and functions of the various parts of an M–14 rifle. Various practical methods might be tried on different groups, such as formal lectures, demonstrations and informal discussion, manipulation of the weapon by the recruits, and so on. If care is exercised that the various groups of trainees are equivalent in background knowledge and aptitude, and if the various instructors are relatively equal in effectiveness, then a test administered at the end could show which methods are the more effective. An achievement test, then, can be used to evaluate the results of an experiment in training methods.

Section III
SUMMARY

126. Achievement Tests as Measures of Proficiency

a. Achievement tests are tests which reveal how proficient an individual has become as a result of experience or training with reference to a particular Army job. Most achievement tests take the form of paper–and–pencil tests or work sample tests. Achievement tests can measure not only information but skills, the ability to apply skills and knowledge, and the suitability of an individual's work habits and work attitudes.

b. In planning the content of achievement tests, attention is paid to economy through proper selection.

c. The chief requirements of achievement tests are objectivity, reliability, and validity. Empirical validity is frequently difficult to establish and recourse must be had to "building in" validity. It is consequently necessary that achievement tests have a proper range of difficulty and adequate differentiation at various score levels.

d. Achievement tests have been used chiefly as aids in classifying and promoting Army personnel and as aids to training. The Enlisted Evaluation System provides for the evaluation of enlisted personnel on an Army–wide basis. Evaluation scores, which consist of a weighted combination of proficiency test scores and ratings by Commanders, are used as a basis for various personnel actions. In the service schools, achievement tests are used in estimating backgrounds of trainees, and in measuring their progress and their achievement at the end of the course. Achievement tests may also be used to measure the effectiveness of various training methods.

126B. Title not used.

Paragraph not used.

Chapter 8
INTERVIEWING AS MEASUREMENT

Section I
PURPOSE OF INTERVIEWS

127. General

a. The interview is one of the oldest personnel techniques and one of the most frequently employed in the Army. It is a technique which a great many people use. However, of all personnel test and measurement techniques, the interview is the most difficult to use successfully.

b. Interviews are but one method of obtaining desired information. This method is used when other sources of information are not as feasible or are inadequate. Particularly where the utility of the information to be obtained can be increased by the personal interaction between the interviewer and the person to be interviewed should use of the interview method be considered. For example, interviews may be prescribed in order to obtain necessary factual information or to assist personnel in making necessary decisions. Such interviews may be prescribed in all aspects of a man's Army career from activities connected with his induction or recruitment to those connected with his separation. This interview may also be an essential tool for communication and leadership at all echelons of the Army.

c. Each person who enters the Army is subject to an initial interview and a classification interview. In addition, he may undergo interviewing for preinduction, for critical assignment for which he may qualify as a specialist, for volunteer duty, for POR (preparation of replacements for oversea movement), for settlement of personal affairs, for appearance before classification boards, for personnel assessment prior to special training (as with OCS training), for personal adjustment purposes and as a part of action separating him from active service. For those occasions on which interviews are prescribed by Army Regulations, specific manuals are provided which set forth the procedures to be used in the conduct of such interviews. A detailed description of the general principles to follow in conducting interviews is contained in DA Pamphlet 611–1, "The Army Interview." Three types of interviews used in the Army will be discussed in this chapter.

d. Proficiency in interviewing can be acquired. The Army provides appropriate instructional materials to those individuals who are required to conduct prescribed interviews.

128. Fact–Finding Interviews

The initial classification interview conducted at reception stations is a fact–finding interview or, more accurately, an information–getting interview, since its primary purpose is that of obtaining information about civilian education, work experience, prior service, avocations, hobbies, extra–curricular school activities, and interests. This information, after being obtained by the classification interviewer, is used along with test score data in making recommendations for training and assignment.

129. Interviews to Survey Attitudes

Interviews to assess attitudes, feelings, or opinions of men or officers is another kind of personnel management tool. For example, a survey of soldiers' attitudes toward some aspect or change in Army policy can be expected to yield data of value to Army personnel officers, although such data must generally be coded and analyzed statistically before they can be put to use. In principle, at least, Army periodic sample surveys are good examples of this type of interview, although frequently the question content to be answered is so well organized and objectified that face–to–face contact is not needed.

130. Assessment Interviews

Standardized measurement interviews conducted for the purpose of assessing specific aspects of the interviewee are used extensively in the Army. The focus is on the behavior of the applicant during the interview rather than on specific information he furnishes. On the basis of interactions which take place during the interview between interviewer and interviewee, certain impressions are formed about an applicant which may frequently be recorded in the form of rating measurements. In general, interviews used as measurement instruments are constructed, standardized, and validated in a fashion similar to other personnel measurement techniques. Such interviews must be applied and scored exactly in accordance with the manual governing the conduct of the interview, if they are to yield results comparable to those obtained when the interview was first standardized.

Section II
THE INTERVIEW AS A MEASURING INSTRUMENT

131. The Value of Interviews as Measuring Instruments

a. The standardized measurement interview, when conducted by qualified personnel who adhere to stated procedures, will yield results which can make a useful contribution to personnel selection and evaluation.

b. One of three instruments constituting the OCS selection battery is a standardized measurement interview conducted by members of an examining board of 3 to 5 officers. Only delimited aspects of the applicant are considered, since the purpose of the interview portion of the examining board procedures is to observe and evaluate personal qualities and social skills required for officer performance. Board members during the interview do not have access to any part of the applicant's previous record. Only after they record their independent evaluations of the candidate on such specifics as self–assurance, appearance, voice control, and ability to organize ideas, do the officers make an appraisal of the applicant in terms of his complete record (including scores on the interview) in determining his overall qualifications for officer training and a subsequent commission.

c. To reemphasize, this kind of interview is not legitimately employed to obtain an estimate of intelligence or information concerning amount of education, or the number of positions of leadership held in school. This type of information is generally available elsewhere as a matter of factual and accurate record. But this type of interview has been demonstrated to be valid for the purpose described—the assessment of a specified aspect of human behavior.

132. Limitations of Interviews as, Measuring Instruments

a. Interviews are quite sensitive to the interviewer's deviations in procedure from that prescribed in the manual. To stray from established procedures in the interview will usually mean that certain aspects of the interviewee, probably already measured by other means, will be given undue weight. The net effect may be to destroy whatever validity can be claimed for the interview in this particular situation.

b. Scores computed as the result of interviews are based on information recorded by the interviewer. Research has demonstrated that special precautions are needed to guard against bias and systematic error in the information recorded. For example, where the interviewer records his impressions in the form of ratings, he must learn to guard against the usual biases or tendencies known to reduce the value of ratings (ch. 10).

c. As a personnel measuring instrument, the interview tends to be a most expensive one. In interviews such as the Board procedure described above, at least, three interviewers must be assembled to obtain reasonably reliable scores. Seldom can more than two interviewees be processed in an hour, with a probable maximum of 12 a day by the same Board.

Section III
SUMMARY

133. Interviews as Personnel Measuring Instruments

a. Interviews are widely used in the Army to obtain factual information and measurement data. They are generally used when other methods of obtaining the desired results are not as feasible or are inadequate.

b. The standardized measurement interview has demonstrated usefulness in assessing certain aspects of the individual. If its validity is to be maintained, it must be conducted according to the prescribed procedure. Administrative considerations limit its use.

133B. Title not used.

Paragraph not used.

Chapter 9
SELF–REPORT FORMS

Section I
THE NATURE OF SELF–REPORT FORMS

134. General

Information about how well a soldier may be expected to perform can be gathered from many sources outside the man himself (efficiency reports, special evaluations, school grades, interviews). Aptitude and achievement test results represent information that comes from the person himself. In this chapter will be discussed a person's own statements regarding his background, his attitudes, beliefs, and personal reactions. One method of gathering such information is by "self–report" forms. Methods of developing such forms are described in this chapter, also their use in selection and classification.

135. Kinds of Self–Report Data

Self–report data fall into two groups. One is objective, factual background information, such as how far a man went in school, what kind of work he has done, what his family's income bracket was. The second group includes more subjective information—what the man thinks about himself, how he feels about his environment, what his attitudes and beliefs Pre regarding other people. Both types of information may be gathered in a single form, variously termed a

"self–description form," a "biographical information blank," a "personality inventory," a "preference form," or some similar term. In this chapter it will be referred to as a "self–report form." In any event, whatever term is used is not necessarily a guide to the specific kind of content of the instrument.

136. Advantages of Self–Report Forms

What are the advantages of self–report forms, as compared with other means of gathering information about a person? When and how may they be used to best advantage?

a. Systematic Data Collection. Self–report forms provide a fairly simple and systematic means of gathering information from a person about himself, in a manner which permits scientific evaluation of the resulting data. The mass of such data that can be gathered is enormous; its very volume presents a serious problem of organization. Therefore, if the information is to be used for prediction measurement purposes, a systematic means of evaluating and using the information about each individual is needed. Self–report forms are devised in such a way that the information gathered can be readily and systematically analyzed and applied in measurement.

b. Uniform Wording. Obviously, uniformity of wording is a crucial requirement in eliciting self–report responses for measurement purposes. It is practically impossible to maintain necessary uniformity of wording in an interview. Even if the wording of interview questions were not varied, such factors as inflection, rate of speech, and word intonation could influence the interpretation of a question. Through such means, the interviewer tends, often unconsciously, to put his own interpretation and cues into his questions. Furthermore, the personal nature of some of the questions may make many people hesitate to reveal their true feelings in an interview.

c. Uniform Appearance. Self–report forms also insure a measure of uniformity in the conditions under which information is furnished by the examinee. The printed instrument is impersonal, uniform in appearance. Special directions may even be used in the attempt to have all examinees adopt a standard attitude toward the instrument and the mental tasks required by the instrument. Even with these precautions, however, there is suspicion and some evidence that important variations remain in the attitudes of individuals responding to the questions, and that these variations have an effect upon the results obtained. Some of the methods developed for controlling the effects of these differences in attitude aid in making self–report forms generally more valid and fairer to the men to whom they are applied.

137. Uses

Since 1958, the Classification Inventory, a self–report form in the Army Classification Battery, has been administered to all incoming enlisted personnel. Other self–report forms are used as aids in screening applicants for certain special assignments where the more intangible personality traits appear to be important. This kind of instrument has been developed to aid in screening for Officer Candidate School training, for helicopter pilot training, for driver training, for assignment of men as recruiters, for selection of Special Forces trainees, and for selection of ROTC honor graduates for commission into the Regular Army. Other instruments have been used successfully in selecting enlisted reserve officers, enlisted personnel, nurses, and medical specialists for Regular Army commissions. Ongoing and future research is expected to add to the number of such instruments used operationally in the Army.

Section II
CONSTRUCTING THE SELF–REPORT FORM

138. General

Self–report forms are more than the series of questions found on the usual questionnaire. Then how is the instrument constructed? What steps are necessary to obtain usable results? How can we tell how good a self report form is after it is constructed? These are questions that have an important bearing on whether or not the report form will be valid.

139. Constructing Items for Self–Report Forms

To begin with, it is necessary to invent or borrow a large pool of self–report items believed or hypothesized to be related to whatever one is trying to predict. Let us suppose the problem is one involving the prediction of leadership. What is important in leadership? What traits of character, what home, family, school, and employment background may possibly contribute to success in leadership? What personality quirks are likely to prevent a person from being a good leader? What personal habits and mannerisms, what patterns of likes and dislikes, what beliefs and attitudes appear to be pertinent? Experience over a number of years indicates that only a small percentage of self–report items written are actually found to have the necessary validity. In view of this fact, many more items are written than are needed for any particular form. Examples of items which might appear on a self–report form are shown in figure 17.

140. Validation of Experimental Self–Report Items

a. The Criterion. After a large number of items have been written, from five to ten times as many as one ultimately expects to use, the items are tried out to determine their validity. The first step in such a tryout, as in the field testing of other instruments, is the development of a criterion. The general problems involved in criterion development have been discussed in chapter 3.

b. Determination of Item Validity. Once the criterion has been obtained, the items are tested against the criterion. This process consists, in essence, of finding what percentage of men who receive high criterion ratings answered "yes" or "high degree" to an item, and comparing it with the percentage of men who received low criterion scores and who answered "no" or "low degree" to that item—in other words, how well the item predicts the criterion ratings. An item responded to one way by men rated high and the opposite way by men rated low would be considered valid and retained. If both good and poor men gave the same kind of answer, the item would not be considered valid for differentiating between them. This process of item analysis is a laborious one, but it is essentially the same process described in chapter 2 in connection with the construction of any test. With self–report instruments, however, it is essential. There is no way of knowing whether a self–report item is valid or not until it is tried out. It may be a desirable quality for an item to appear to be valid, but this apparent validity cannot be counted on in any degree as a substitute for demonstrated validity, since only a minority of items can be expected to have sufficiently high demonstrated validity to warrant their use.

c. Selection of Items. Once the validity of each item has been determined, the instrument can be set up in form for actual use. The items which have proved to be valid are either assembled in a new form or are spotted in the original one. A scoring key is set up to permit counting the number of the valid answers chosen by each person who takes the test. The score obtained from such a key is the revised predictor of the criterion. The problem of determining how accurate a prediction it can give is the last step.

SELF - REPORT ITEMS

CAN REFER TO FACTUAL BACKGROUND

Check one

How many living brothers and sisters have you?

A) none
B) 1
C) 2
D) 3
E) 4 or more

When of high school age, how many evenings a week did you go out?

A) less than 1
B) 1
C) 2
D) 3
E) 4 or more

PERSONALITY CHARACTERISTICS

Mark A or B

A) I believe it is important to get what you want even if you have to fight to get it.

B) I believe it is more important to be well-liked by my employees than to get the work done exactly according to established procedures.

A) I am open-minded.

B) I hold my purpose.

Mark the degree to which the statement applies to you.

Always finish what I start

A) To a low degree.
B) To the usual degree.
C) To a high degree.

Figure 17. Types of items used in self–report forms

d. Cross–Validation. To find out how good the instrument really is, it is necessary to try the revised form out on a new group of men on whom criterion ratings are obtained. If good agreement is found between the criterion ratings and the self–report answers as scored by the key developed from the item analysis, there is reasonable assurance that there is a real relationship between the keyed items and the criterion. This process of second try–out is known as cross–validation. If this second try–out fails to show substantial relationship, it is necessary to discard the key and start over again with either another group of items or a more stable criterion, or both. Cross–validation is a research step which has been strongly emphasized to assure validity of Army measuring instruments.

Section III
A SUPPRESSOR METHOD OF IMPROVING VALIDITY OF SELF–REPORT FORMS

141. Errors of Conscious Distortion

Self–report forms are subject to errors of distortion. Some individuals will deliberately attempt to deceive in order to appear in a more favorable light. These examinees will select responses which they hope will create a favorable impression or which will help them to get a desired assignment. A strong warning to be "honest and objective" will have little effect on such an individual.

142. Errors of Unconscious Distortion

Some examinees may select inaccurate responses, not realizing that they are thus distorting their self–appraisal. Unintentional distortion is a major source of error because individuals react quite differently to self–report forms. Individuals vary in their opinions of themselves. Some people tend to overestimate themselves, honestly regard themselves highly, and select responses that fit this high opinion. Others may be unduly modest and lean over backward to avoid exaggerating their merits. This latter individual may be more frank than the average examinee in revealing his shortcomings.

143. Control of Distortion by a Suppressor Key

To reduce the effect of distortion, whether conscious or unconscious, a measure of the degree to which each person distorts must be obtained. If this measure is also unrelated to the criterion, it can be subtracted from the self–report score and thus reduce the effect of distortion. Note that the outcome of such a subtraction would not be desirable if the measure were related to the criterion since a valid portion of the self–report score would be removed as well as the distortion. A measure of distortion can be obtained when the self–report scores and the criterion scores are both known. This measure cannot be used for initial selection since criterion scores are then unknown. Through detailed statistical analysis, items are identified that are related to the distortion measure, but which have little or no relationship to the criterion. If enough of these items can be obtained, a reliable measure of distortion will result. This measure, known as a suppressor measure, can then be "keyed", i.e., applied to, each test. With the use of the suppressor key, a person who overestimates himself has, in effect, a subtraction from his score. A person who underestimates himself has an increment added to his score. This process seems to have considerable promise in increasing the validity of self–report forms.

Section IV
FORCED–CHOICE METHOD OF IMPROVING VALIDITY

144. General

Thus far two sources of bias or systematic error in self–report forms have been described and means of dealing with them using the suppressor method have been discussed. Another method is known as "forced–choice."

145. What Is a Forced–Choice Instrument?

In a typical forced–choice instrument, the person is required to respond to each item by choosing between two alternatives which appear equally attractive, but only one of which is valid. There might be three alternatives, and the person may be asked to choose the one which least describes him and the one which best describes him. Or there might be a list of a dozen phrases from which the examinee is required to choose the five that best describe him and the five that least describe him. However many alternatives are presented, they are so set up that the valid alternatives appear as attractive as the non–valid alternatives. The attractiveness and the validity will have been previously determined.

146. Grouping Alternatives

For the sake of simplicity, only the grouping of forced–choice, alternatives into pairs will be described. The procedures employed apply to larger groupings as well. The basic procedure for grouping alternatives takes into account the two main characteristics of each alternative—its attractiveness and its validity.

a. Three measures of attractiveness are obtained.

(1) The number of times an alternative has been marked by the men as describing themselves (the "p value"), which might be called its popularity and which is similar to the difficulty value of test items;

(2) The face validity index, (a measure of susceptibility of an item to conscious bias); and

(3) The distortion index.

b. In principle, the alternatives which would be paired would have the same attractiveness, as defined above, but different validity coefficients, one alternative showing high positive validity (the alternative chosen only by competent men) and the other showing zero or negative validity (the alternative chosen equally often by competent and incompetent men or chosen only by the incompetent). In other words, the man should have considerable difficulty in choosing one of the alternatives except on the basis of whether it really describes him.

c. The above grouping of alternatives is not applicable to all kinds of material. Factual background items do not lend themselves to forced choice treatment. It would not make sense to ask a person, for instance, which is more true of him: that he had an eighth grade education; or that his father was American–born. These things are true or not true, and no weighing of evidence or judgment is involved in answering them.

147. Checking Validity of Scoring Key

The construction of a number of forced–choice items on the basis described above does not complete the task of building a forced–choice self–report instrument. Responses to items in pairs are not always the same as to each member of the pair presented by itself. Thus, there is no assurance that an alternative found to be valid when it was not paired will maintain its validity when paired with another alternative. A key scoring the valid alternatives which are selected can be prepared on the basis of the individual alternatives in each of the items. Such a key is known as a "predetermined" key. However, there is no assurance that this key will stand up after the pairing. Before the forced–choice self–report instrument can be said to be ready for operational use, it is necessary to check the validity of the predetermined key.

a. To do a thorough job of building a forced choice self–report instrument, the experimental instrument should be administered to two new groups of men for whom criterion data are available. The first of these groups is used for reanalyzing the validity of the items, that is, for constructing a scoring key based on the items after pairing.

b. After the key has been set up, it is necessary to determine its validity on a second group (cross–validation).

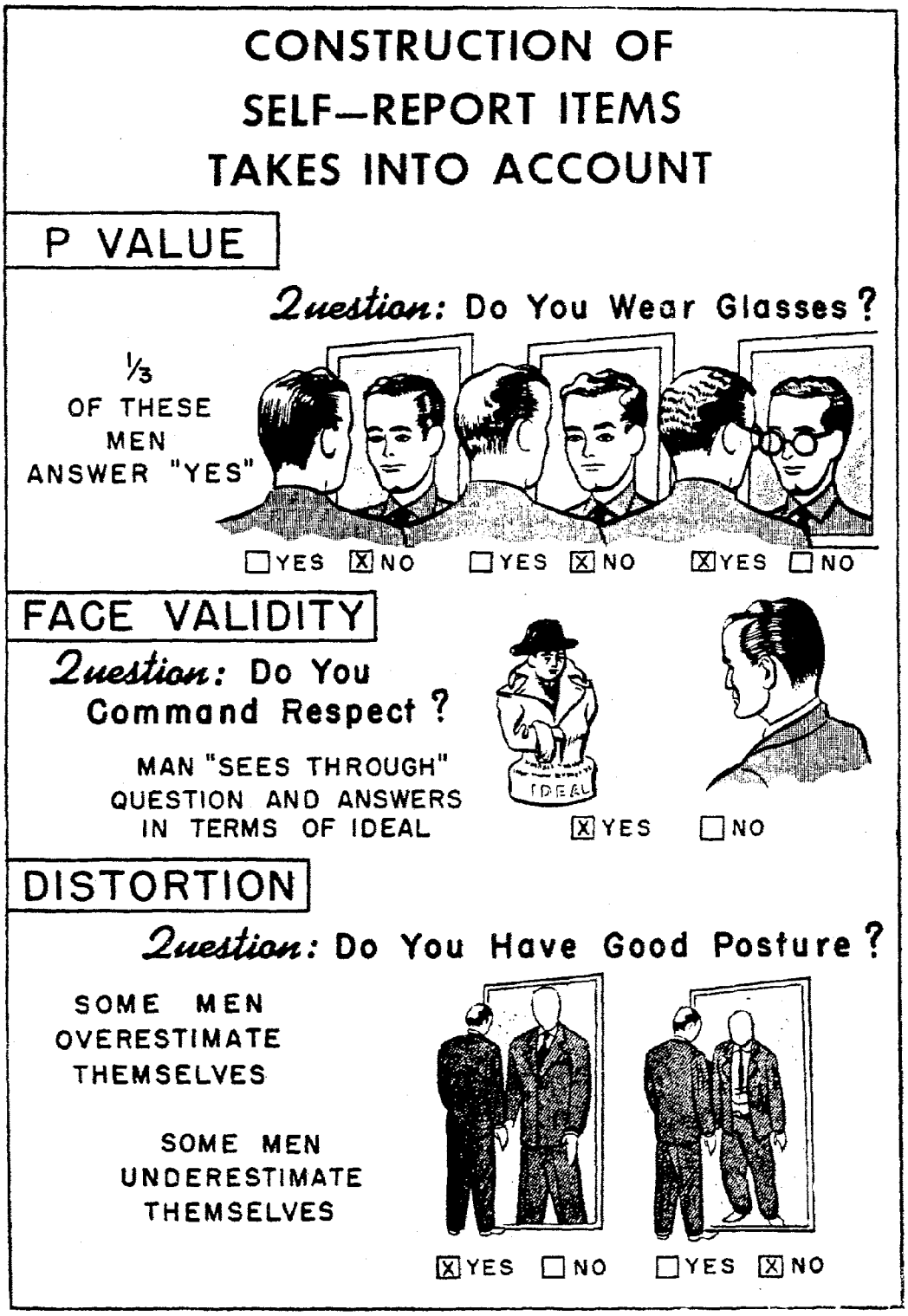


Figure 18. Important characteristics of items considered in developing self–report forms

Section V
SUMMARY

148. The Value of Self–Report Forms

a. Self–report forms are used as a ready means of gathering, analyzing, and interpreting in a standard manner a vast and varied mass of fact which soldiers supply regarding themselves.

b. Successive steps in building such self–description instruments are—

(1) Selection of traits which seem to be related to the criterion.

(2) The writing of a large number of items which appear to be related to the criterion.

(3) Experimental try–out and item analysis.

(4) Cross–validation.

c. A number of technical refinements have proved valuable in reducing bias. These methods include measurement and allowance for the popularity of the items, and their susceptibility to conscious or unconscious distortion.

d. One means of increasing validity of self–report instruments is the use of the forced–choice technique. This procedure requires a person to choose between two or more apparently equally attractive alternatives, one of which is more valid than the other. Where forced–choice procedure is not practicable, special keys are set up to measure and compensate for such sources of error in prediction as conscious bias and susceptibility to distortion.

148B. Title not used.

Paragraph not used.

Chapter 10
RATINGS AS MEASURES OF USEFULNESS

Section I
GENERAL CHARACTERISTICS OF ADMINISTRATIVE RATINGS

149. How Ratings Differ from Tests

An achievement test score is a measure of how much a man knows or how well he can perform on a given task. An aptitude test is an estimate of how well the man can be expected to learn. When a standard interview or a self–report form is used, the resulting score indicates how well the man is motivated, how well he can lead others, or perhaps how suitable his personality characteristics are for a particular job. The more valid the instrument and the more adequate the criterion, the greater the safety with which the information provided by the instrument can be used.

a. Need for a Direct Measure of Usefulness. Still another kind of information is needed for effective utilization of manpower—a more direct measure, a measure based on a man's present job performance as judged by those who know the man's work. What do his superiors and associates think of him? How valuable do they think he is to his unit? How useful to the Army? What does his present usefulness indicate about his future usefulness? It is to provide such an evaluation that rating procedures are used.

b. Problems Involved in Ratings. The prime characteristics desired in any personnel measuring device is validity—is it useful for a specified purpose? Next comes the problem of reliability—is it a consistent measure? Tests of all kinds can be effectively dealt with in terms of these two concepts. With ratings, these concepts are complicated by a number of conditions which are of relatively little concern when dealing with objective tests. A rating involves not only a ratee, but a rater. A person may or may not like a test, but his performance on it is his own. His rating, however, depends not only on what he does but on how the rater values what the ratee does and how this evaluation is reported. Also, the various systems of values involved make the problem of acceptability of a rating procedure to the rater of extreme importance. The rater has definite ideas of where, on a given scale, he is placing the ratee and he expects administrative action to be consistent with this. The problem of validity and reliability is thus cast in a different setting, and a whole set of problems concerned with the characteristics of the rater and the conditions surrounding the rating are added.

149B. Title not used.

Paragraph not used.

Section II
PURPOSES OF RATINGS

150. Ratings Classified According to Purpose

Ratings are used for guidance, administrative, and research purposes. This chapter is primarily concerned with ratings for administrative purposes. But since these purposes are frequently confused, a brief discussion will first be undertaken to clarify the differences. One confusion is especially unfortunate—attempting to use the same procedure for evaluating a person for administrative actions (promotions, selection for special assignment) and for guidance (assisting a subordinate to improve in his work).

a. Ratings for Guidance Purposes. A prime responsibility of an officer or a supervisor is the development and improvement of the work of his subordinates. If a subordinate is to be helped, his strengths and weaknesses need to be analyzed in relation to the demands of a particular job. When this is done, the result will probably be a list of specific requirements, so specific, in fact, that they may sometimes apply only to a single individual. A rating scheme to serve this purpose must, therefore, be flexible and must not be limited to a few characteristics or general job requirements. It usually requires an interview; otherwise the subordinate will not learn of the results of his superior's analysis of his strengths and weaknesses. Furthermore, the scoring of guidance ratings is considered not only unnecessary but undesirable for two reasons, both of which tend to defeat the guidance objective. For one thing, scoring would tend to minimize any diagnostic values to be derived. For another, tendencies to rate on general impressions rather than on specific aspects of behavior, to be lenient, and to render ratings with little spread are more likely to operate when ratings are scored.

b. Ratings for Administrative Purposes. The problems of obtaining a rating on job performance or overall worth are different from those involved in rating for guidance purposes. First, ratings for administrative purposes should reflect a man's standing in his competitive group. A quantitative score is required. Second, their basic purpose involves evaluation of a person. Third, ratings for administrative purposes influence decisions regarding a subordinate's career, a circumstance which has many implications. Fourth, administrative ratings may be used as predictors and, as such, require consideration from the standpoint of reliability and validity exactly as any other predictor instrument. Their value cannot be accepted, but must be established. Fifth, attitudes of the rater toward the rating system and indeed toward the total personnel system are more important. The relationship between superior and subordinate, for example, is involved in the acceptance or non–acceptance of a rating procedure. Sixth, ratings for administrative purposes do not require an interview between rater and ratee. They do not preclude an interview, although the greater leniency that usually accrues under such circumstances tends to reduce the value of the rating.

c. Ratings for Criterion Purposes. Ratings for criterion and administrative purposes have much in common. Both are subject to the same sources of bias or error, and both require a quantitative score to reflect comparative position in some group. Still, there are differences that preclude safe generalizations from one to the other. Criterion ratings are not used for administrative purposes; their basic purpose is to evaluate a test or procedure, not a person. Criterion ratings are not predictors; they are accepted as the best available index or yardstick of a given kind of performance. Since their value is accepted, validity studies need not be conducted. In general, the differences between administrative and criterion ratings stem from the difference in basic purpose—the one, to evaluate a person for administrative purposes; the other, to evaluate a test or procedure.

151. Specificity of Purpose and Effectiveness in Rating

Ratings for administrative purposes may be directed toward promotion at critical levels in the supervisory scale, such as designation of a higher skill level within an enlisted MOS or promotion of an officer from company to field grade. Or the rating may be aimed at selecting men for particular kinds of assignment, such as staff or line duty. In general, the more specific the purpose can be, the greater the likelihood that a better job can be done with a rating scale. The possibility of greater effectiveness must, of course, be balanced against administrative problems involved in having a number of different scales, each for a specific administrative purpose.

Section III
ADMINISTRATIVE RATING METHODS

152. Need for Effective Performance Ratings

In a small business where the owner or manager has ample opportunity to observe all his employees, the problem of performance ratings does not arise. When the time comes to make a promotion, he has little trouble establishing an order of merit. As the business grows and becomes departmentalized, however, the problem of locating the best employees becomes increasingly difficult. And when the business has branch offices scattered all over the world, devising a technique of rating that is valid for the purpose and fair to all employees becomes very difficult indeed. No less difficult is the Army's problem of rating fairly and effectively its great numbers of officers and enlisted men. The Army's need to evaluate the usefulness and potentialities of its personnel makes the search for good methods highly critical. Fairness in rating is critical not only from the standpoint of making the best personnel decisions; it is equally crucial to morale. Methods that reduce "hard–easy" rater differences or minimize the tendency of most raters to

concentrate ratings at the high end of the scale are needed to increase the fairness and effectiveness with which rating systems operate.

153. Methods Unsuitable for Administrative Ratings

a. Essay Ratings. Descriptions and evaluations of the ratee in the rater's own words have the definite advantage of permitting the rater to express himself as he feels about an individual. Such essay ratings also permit the rater to provide detail about a man's performance and capabilities which may be useful in future personnel actions. However, essay ratings are not suitable for large–scale rating operations in that the evaluations do not lend themselves to objective scoring, and therefore do not furnish a reliable means of comparing one ratee with another. Even when the number of ratees is limited, essay ratings are seldom an adequate basis for comparing individuals. The various raters are likely to cover different aspects of the job and different characteristics of the ratees. Bias may also be introduced by differences in the literary skill shown by the raters.

b. Ranking. Placing men in a competitive group in order of merit forces the rater to consider all men in relation to each other. He cannot be lenient and evaluate all his men high—he must discriminate among them. Differences in rater standards do not affect the ratings. However, ranking is not practical for general use. It cannot be used when only one man is being evaluated. The numerical score assigned each man depends on the size of the group. For example, the worst man in a group of five would receive a rank of five, the same as the fifth man in a group of twenty. True, ranks in different groups can be converted to a common scale. However, the converted score, like the original rank order, shows only relative merit; it does not show the amount of difference between individuals. For example, in one group men ranked 1 and 2 may be nearly alike, but in another group the men ranked 1 and 2 may be very different. Thus, rankings are seldom used administratively.

c. Guided Ratings. Having a rating expert assist the official rater, a method described in the discussion of criterion rating methods as a means to more effective rating, is too time–consuming and expensive for application in a large–scale rating system.

154. Rating Methods Suitable for Administrative Reports

An administratively feasible rating system must be applicable in situations where superiors are evaluating a single subordinate or many. Rating systems used in the Army usually are based on one of two types of evaluation: the traditional form on which the rater indicates to what extent the ratee possesses the rated element, and the forced choice method in which the rater is asked hot "how much" but which of two or more descriptions is most or least characteristic of the ratee.

a. Traditional Rating Scales. The traditional type of rating scale has been used in almost numberless variations. This type is exemplified by Section II of the Commander's Evaluation Report, DA Form 2166, used in the Enlisted Evaluation System and reproduced as figure 19. Each question the rater has to answer concerns the amount of an attribute—trait, characteristic, or overall worth—possessed by the rates. The scale may concern general traits such as dependability or relatively specific traits such as the ability to prepare a good report. The scale points may be left somewhat vague, or they may be very carefully defined. The number of points on a scale generally varies from two to not more than seven or eight.

b. The Forced–Choice Method.

(1) In this method, the question posed to the rater is "Which of two or more descriptions applies most to the subordinate?" Two ways of presenting forced–choice descriptive phrases are shown in figure 20. The method can be considered an adaptation of ranking. Instead of placing a group of individuals in order of merit, the rater ranks traits within an individual. The method has been applied to performance evaluation as a means of reducing the bias to which traditional ratings are subject and of lessening the effect of differences in rater standards.

(3) Operationally, the acceptability of forced choice in administrative ratings has not been successfully established. Raters have complained that in many cases none of the phrases from which they had to choose described the ratee accurately. Even less acceptable is the feature that prevents the rater from knowing whether he is giving a high or a low rating. Since he does not know which response would contribute to a favorable rating, he is in the dark as to the final score the ratee will receive. Army officers particularly prefer to know the efficiency report score they are giving their subordinates.

SECTION II *(To be accomplished by Rater and Indorser)*

14. COMPARED TO ALL OTHER MEN/WOMEN THAT YOU HAVE KNOWN IN THIS MOS:	RATER				INDORSER			
	POOR	FAIR	GOOD	BEST	POOR	FAIR	GOOD	BEST
a. HOW WELL DOES HE UNDERSTAND WHAT TO DO WITHOUT DETAILED INSTRUCTIONS?	0	1 2 3	4 5 6	7 8 9	0	1 2 3	4 5 6	7 8 9
b. HOW WELL DOES HE TAKE PROPER ACTION IN THE ABSENCE OF ORDERS?	0	1 2 3	4 5 6	7 8 9	0	1 2 3	4 5 6	7 8 9
c. HOW PERSISTENT IS HE IN OVERCOMING OBSTACLES?	0	1 2 3	4 5 6	7 8 9	0	1 2 3	4 5 6	7 8 9
d. TO WHAT EXTENT DOES HE TRY TO LEARN ABOUT HIS JOB?	0	1 2 3	4 5 6	7 8 9	0	1 2 3	4 5 6	7 8 9
e. HOW WELL DOES HE KNOW ALL ASPECTS OF HIS JOB?	0	1 2 3	4 5 6	7 8 9	0	1 2 3	4 5 6	7 8 9
f. HOW WELL DOES HE COOPERATE WITH OTHERS ON THE JOB?	0	1 2 3	4 5 6	7 8 9	0	1 2 3	4 5 6	7 8 9
g. HOW WELL DOES HE RECEIVE AND CARRY OUT ORDERS?	0	1 2 3	4 5 6	7 8 9	0	1 2 3	4 5 6	7 8 9
h. HOW EFFECTIVE IS HE AS A LEADER?	0	1 2 3	4 5 6	7 8 9	0	1 2 3	4 5 6	7 8 9
i. HOW WELL DOES HE PERFORM THE DUTIES FOR WHICH HE IS BEING RATED?	0	1 2 3	4 5 6	7 8 9	0	1 2 3	4 5 6	7 8 9
j. WHAT IS HIS ADVANCEMENT POTENTIAL?	0	1 2 3	4 5 6	7 8 9	0	1 2 3	4 5 6	7 8 9

Figure 19. Rating scale from Commander's Evaluation Report, DA Form 2166

155. Administrative Rating as Procedure

a. The Report Form. In administrative rating, there has been a strong tendency to emphasize the report form. Actually, rating is a procedure in which the form plays but one part, usually minor. The form, however, can be an aid to more accurate and useful evaluations by requiring the rater to be more careful, more critical in recording his judgments; or it can provide for recording information which might be used in evaluating the judgments of the rater, such as the length of time he has supervised the ratee. However carefully developed, the form cannot make up for inadequate observation and faulty judgment. Far more important aspects of rating are the procedures employed and the competence of the rater to judge.

PAIRS

FOR RATER ONLY. For each pair of words or phases make a heavy X opposite the one that is the MORE DESCRIPTIVE of the rated officer.

1	A. Assigns men properly ☐ B. Keeps his word ☐	8 A. Maintains strict discipline ☐ B. Good educational background ☐
2	A. Courageous ☐ B. Respected by his subordinates ☐	9 A. Can select and define major objectives ☐ B. Thorough knowledge of his own branch ☐
3	A. Willing to take a chance ☐ B. Has physical endurance ☐	10 A. Temperate in his habits ☐ B. Self-confident ☐
4	A. Conscientious ☐ B. Takes action to correct faulty performance ☐	11 A. Is just ☐ B. Can get subordinates to attempt the impossible ☐
5	A. People seek his advice in personal matters ☐ B. Knows his subordinates ☐	12 A. Has a broad grasp of the problems ☐ B. Has foresight ☐
6	A. Alert ☐ B. Has full knowledge of his job ☐	13 A. Vigorous ☐ B. Truthful ☐
7	A. Thoughtful planner ☐ B. Supports actions of subordinates ☐	14 A. Tenacious ☐ B. Makes practicable suggestions ☐

OR

TETRADS

This section consists of sets of phrases which describe characteristics related to proficiency on the job. Consider each set and decide which phrase is MOST DESCRIPTIVE of the officer being rated, and which phrase is LEAST DESCRIPTIVE of the officer. Judge the sets independently -- it is not necessary to be consistent, as you are describing the officer, not evaluating him. Blacken in the space to the right of the MOST DESCRIPTIVE and LEAST DESCRIPTIVE phrases in the appropriate column.

A. Becomes dogmatic about his authority.			A. Always criticizes, never praises.		
B. Careless & slipshod in attention to duty.		1	B. Carries out orders by "passing the buck."		4
C. No one ever doubts his ability.			C. Knows his job and performs it well.		
D. Well-grounded in all phases of Army life.			D. Plays no favorites.		
A. Follows closely directions of higher echelons.			A. Constantly striving for new knowledge and ideas.		
B. Inclined to "gold-brick."		2	B. Businesslike.		5
C. Criticizes unnecessarily.			C. Apparently not physically fit.		
D. Willing to accept responsibility.			D. Fails to use good judgment.		
A. A go-getter who always does a good job.			A. Cannot assume responsibility.		
B. Cool under all circumstances.		3	B. Knows how and when to delegate authority.		6
C. Doesn't listen to suggestions.			C. Offers suggestions.		
D. Drives instead of leads.			D. Too easily changes his ideas		

Figure 20. Forced–choice phrases may be assembled in pairs or tetrads

b. Clarifying the Purpose of the Rating. Administrative policies controlling the purpose—and use—of the rating have major influence on the attitude of the raters. Confusion of purpose can result in ineffective rating. A clear concept of the purpose for which a rating is to be used and of the policies controlling these uses can go a long way toward improving an administrative rating system. If the rater has confidence that rating scores are used with great care and due respect for their limitations, his desire to control the results of his rating can perhaps be lessened.

c. Administrative Support of the Rating System. Administrative conditions can influence the quality of official ratings by creating an atmosphere conducive to careful rating. Management can reinforce its position that ratings are an essential part of personnel administration by allotting adequate time for the rating process and by providing a clear requirement that ratings be rendered at stated intervals or under given conditions. If adequate time is not allowed for careful completion of the report, if the rater has not been in a position to observe his subordinates adequately, or if he is required to rate an excessive number of subordinates, it will be in exceptional cases that careful, accurate ratings will be rendered. It is extremely important, then, that when management is convinced of the need for a rating procedure, this conviction be expressed in the administrative conditions surrounding the rating.

Section IV
MAJOR PROBLEMS IN ADMINISTRATIVE RATING

156. Acceptance of the Rating Procedure

Many problems in achieving fair and valid ratings stem from rater attitudes toward the rating procedure. At best, rating is a difficult and disliked task. The more the raters—and management—accept the system, the greater the likelihood that better reports will be rendered. The rater is more likely to perform the rating task competently if he is convinced of at least three points—the need for ratings, the ability of the rating system to provide the kind of ratings needed, and the justice of the policies which govern the use of the rating. Higher echelons, by providing the conditions and the time that make good reporting possible, can emphasize the value of carefully executed rating procedures.

157. Rater Tendencies in Army Administrative Rating

Three outstanding findings have consistently emerged from Army research studies on ratings: First, raters characteristically display inability to rate specific aspects of an individual's performance without being influenced by their general impressions of the individual. Second, agreement among raters tends to be unsatisfactorily low. And third, raters show a strong tendency toward leniency in official ratings, a tendency which results in a pre–ponderance of high ratings.

a. Difficulty of Rating Independently Different Aspects of Performance. Raters tend to consider at a man who is superior in one respect is superior in all respects. In other words, their ratings on different aspects of performance turn out to be highly correlated, even when, the various aspects appear to be logically independent. This "halo" tendency operates in administrative rating probably in even greater degree than in rating for research purposes. The tendency may not be of great consequence when the rating is one of overall worth. However, differentiation on the part of the rater is important if ratings are to be used in assigning men to training or jobs with different requirements. Moreover, if specific characteristics are accurately rated, a total score combining the specific ratings would be expected to provide a better measure of overall worth than would a single general estimate.

b. Low Agreement Among Raters. Different raters are likely to observe the ratees in different situations and hence base the ratings on different samples of behavior. Raters may interpret the scale points differently; they may have different professional, personal, and social relations with the ratee; or they may observe the ratee in the same situations but have different standards or even different values for the significance of the behavior for overall worth. Much of the lack of agreement can be attributed to rater bias of this sort. Error of an unsystematic nature, of course, still accounts for a certain amount of the disagreement. The implication of the low agreement for administrative rating is that it is advisable to use as many ratings as possible prior to taking personnel action. If it is impractical to get a number of ratings simultaneously, successive ratings can be averaged.

c. Leniency of Raters. Raters rendering official evaluations tend to be more lenient than raters accomplishing non–administrative evaluations. In giving a rating that is to be used for a personnel decision, it is natural to give a subordinate the benefit of the doubt. Also, when raters show the completed ratings to their subordinates—a frequent practice when official ratings serve as occasions for counseling—raters tend to be more lenient and to discriminate less among ratees. The problem of leniency is complicated by the fact that all raters are not lenient to the same degree. Differences among raters may be larger in administrative than in criterion ratings and may account in part for the difficulty of increasing the validity of official reports.

d. Concentration of Ratings.

(1) Another way of describing leniency is to say that raters tend to give a disproportionate number of high ratings. The lack of dispersion can seriously limit the administrative usefulness of ratings. Army–wide ratings are of little value in personnel actions if the scores for a large number of men are so close together that relative degrees of competence cannot be distinguished.

(2) One reason for lack of adequate dispersion may be a confusion over reference Points. Raters may be applying the scale, not to a given competitive group, but to the general population. Concretely, an individual, when compared with others of the same grade or rank, may be the poorest of the lot and still be considerably better than the average of the general population of the country. In administrative rating, the reference point should be the average of the competitive group, not the average of the general population.

158. Validating Administrative Rating Procedures

The problems involved in validating ratings to be used in personnel actions differ in many respects from those involved in validating other instruments. The need for validation, however, remains the same. Early studies of efficiency reporting dealt with problems of distribution of scores, reduction of halo, rater standards, and reliability. Toward the close of World War II, the Army began intensive research on efficiency reporting methods with emphasis on the validation of the rating system. The accepted approach was to obtain a number of carefully collected anonymous ratings by associates and superiors and to use the average of these ratings as a criterion of effectiveness. The rating procedure producing scores which best agreed with the criterion measure was considered the least biased, that is, the most valid.

a. Technique Contamination in Validating Ratings. While the standard approach to the validation of efficiency reports is sound, several problems were soon encountered. When a number of types of rating are being evaluated, the type most like the type used to obtain the criterion ratings tends to show the highest validity. In other words, like scales tend to correlate with like scales. The tendency is referred to as "technique contamination" and is particularly serious when two widely different rating techniques—traditional scales and forced–choice methods, for example—are being compared.

b. Rater Contamination. If some or all of the criterion ratings are obtained from raters who also complete the ratings to be validated, the two sets of ratings will show agreement attributable to "rater contamination." Rater contamination is difficult to avoid because under most circumstances it is not possible to find two independent sets of raters who are sufficiently informed about the ratee. An adequate validation procedure dictates that the rating being validated and the ratings used as the criterion should not be made by the same raters.

c. Because of the differences noted between administrative ratings and ratings obtained for research, it is essential that the validity of a rating procedure be measured under conditions that approximate official operating conditions.

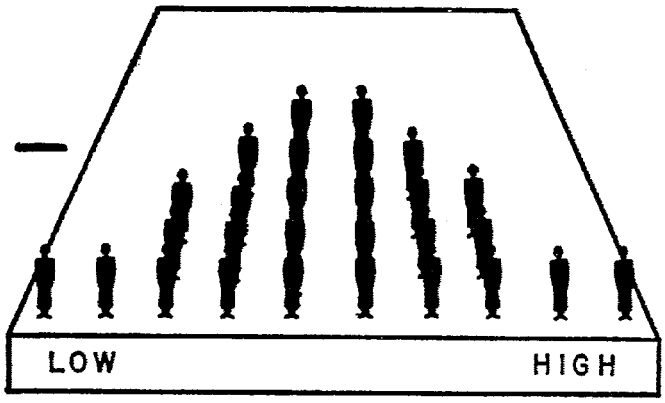
159. Standardizing Administrative Ratings

a. The Standard Score Scale. Use of a standard scale for official ratings permits the rating to serve a number of different purposes. In addition to enabling comparisons of individual scores with an Army–wide reference population, the standard scale provides a means of comparing evaluations made at different times in an individual's Army career. Even more important is the desirability of averaging reports from different sources or over a period of time, a procedure which would not be possible unless the same scale were used with all reports. However, use of standard scores for administrative ratings such as the Commander's Evaluation Report used in the Enlisted Evaluation System, or the Army–wide, Officer Efficiency Report, presents problems not encountered in the operational use of standard scores on tests.

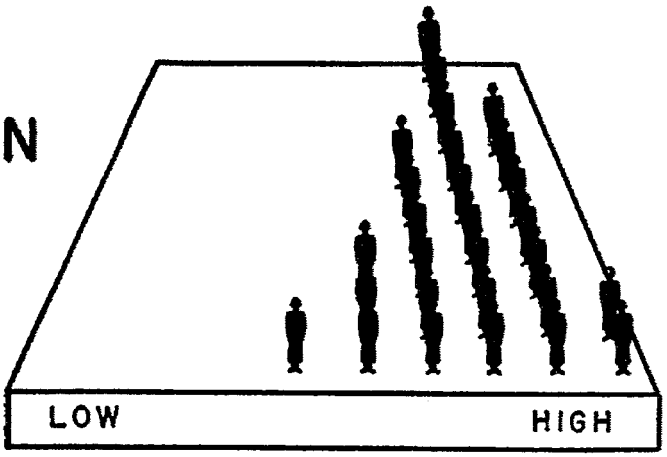
RATINGS

SHOULD BE DISTRIBUTED—

LIKE THIS—



BUT OFTEN LOOK LIKE THIS—



BECAUSE RATERS TEND TO

1. CONCENTRATE THEIR RATINGS
2. BE LENIENT

Figure 21. Two important difficulties in obtaining valid ratings

b. Rater Acceptance. The problem of acceptance by the rater is peculiar to administrative ratings. When the raw score on an evaluation is converted into a standard score, the rater often feels that the resulting score does not represent his intentions with regard to the ratee, particularly when the score falls below average, that is, below 100. Then, too, raters in the Army generally prefer to know where on a scale they are placing a subordinate. However, if the basis of conversion is made available to all raters, they may be influenced by the knowledge to place ratees at desired points on the scale. The usual concentration of ratings results, and, the standardization no longer holds. Further, if the rater rates in terms of relative position, he is not rating in terms of specified rating scale content, and any rating system reduces to an estimate of position in a group. Although ratings are generally more valid when the rater does not know how favorably he is rating a subordinate, use of standard scores centrally determined is limited in view of low degree of acceptability to Army raters, particularly rating officers.

c. Difficulty of Maintaining Standardization. All Army administrative ratings for which the score has been determined centrally by a conversion table to which the rater did not have access have tended to yield higher average scores from year to years raters learn the relative positions their subordinate attain as a result of their ratings, they then tend to become more lenient. Ratings become confined more and more to a narrow range in the upper portion of the scale and cease to provide a means of discriminating among ratees. The standardization becomes obsolete, as does the report form itself.

d. Operational Standardization. In view of the known differences between administrative and non–administrative ratings, the standard scale cannot be established on an experimental sample basis. Thus the standardization of administrative rating procedures—like their validation—must be in terms of the operational situation.

160. Improving Administrative Ratings

a. Averaging Reports. An average of ratings by a number of individuals is generally a more valid and stable measure than a single rating. Averaging reduces effects of individual bias and differing frames of reference among raters. The principle has been consistently applied in obtaining ratings for criterion purposes. The principle was also recognized in the Over–all Efficiency Index (OEI), derived by combining an officer's successive efficiency report score into a long–term composite score. With current Army emphasis upon evaluation of an officer's performance during a given rating period, the OEI is no longer computed. However, a weighted Annual Score is computed from reports received by the officer over one year.

b. Separating Rating from Counseling. When raters are required to show completed ratings to subordinates, the raters tend to be more lenient, and to discriminate less among ratees. Provision for counseling to take place apart from rating for official purposes obviates the need for the rater to show the completed rating to the ratee. In the present officer efficiency rating system, counseling well in advance of the date of rating is mandatory; even optional showing of reports to the rated officer has been discontinued.

c. Educating the Rater. Raters need education and training in the rating process. They need training in how to prepare themselves for the rating task, particularly with respect to the important factors in an individual's performance and behavior they should observe before making a rating. They need guidance in the specific evaluation procedures they are asked to go through as a part of making the rating. One practical way of indoctrinating raters would be to include appropriate training materials in the curricula of various Army schools. However, this form of education would need to be supplemented by supervision in making actual ratings. Such supervision could be finished by the rater's superior. Provision of in expert to work directly with the rater is an effective but time–consuming and expensive means of improving ratings. It is feasible only when a rating system applies to a limited number of ratees.

Section V
SUMMARY

161. Ratings for Administrative Purposes

a. Administrative ratings are needed in the Army as estimates of the overall value of men in large competitive groups.

b. Two methods of rating have been found suitable for administrative ratings: the traditional rating scale and the forced–choice rating. The traditional rating has found greater acceptance among Army raters.

c. Rater tendencies that reduce the usefulness of administrative ratings are: tendency to rate on general impression, to have varying standards of performance, and to concentrate ratings at the favorable end of the scale.

d. More useful administrative ratings can be achieved by:

(1) Improving techniques for obtaining ratings;

(2) Educating the raters in effective rating procedures;

(3) Assuring the rater adequate opportunity to observe ratee performance;

(4) Clarifying the purpose of the rating;

(5) Allowing the rater time and opportunity for careful rating; and

(6) Averaging a number of ratings of an individual.

e. Because administrative ratings and ratings for research show wide differences, an official rating system can be validated and the standard scale established only under conditions approximating those under which operational reports are rendered.

161B. Title not used.

Paragraph not used.

Chapter 11
THE ADMINISTRATION OF ARMY TESTS

Section I
INTRODUCTION

162. General

The scientist develops tools; the technician puts them to use. The instruments devised by personnel psychologists are constructed on the basis of extensive knowledge, and carefully standardized. As such they are valuable means of increasing the accuracy of observations or personnel measurements, and of revealing important and useful information. Once an instrument is constructed, great care must be taken that scores are obtained with these instruments exactly as specified by the research which built the instrument. The aim of this chapter and chapter 12 is to clarify the work of the technician in administering and scoring personnel evaluation instruments so that results of the greatest possible value to the Army will be produced.

163. Authorized Test Instructions

a. Specific directions for administering and scoring are set forth in the manuals which accompany each Army personnel test. In addition, AR 611–5 provides general instructions which must be adhered to in administering all authorized Army personnel tests, except MOS evaluation tests, USAFI test materials, and clinical psychology diagnostic tests.

b. The general instructions contained in AR 611–5 and the specific directions for administering and scoring set forth in the manuals accompanying each Army personnel test are as much an integral part of every test as the test questions themselves. These directions must be adhered to strictly, since any deviation can adversely affect the accuracy of measurement.

164. Testing Situation Must Be Standard

Since it is the function of every evaluation instrument to compare each individual with others in the Army population, it follows that the conditions under which tests are administered and scored must be the same for every soldier, regardless of when or where the test is given. The scores of men who are tested in noisy surroundings or by slipshod methods in all probability are not comparable to those of men examined under favorable circumstances. Nor are such scores necessarily accurate indications of the real abilities of those men. The use of such scores can only result in incorrect evaluation with attendant loss of efficiency to the Army.

a. Testing conditions and procedures should be so standardized that if it were possible to find two individuals exactly alike, both would achieve the same scores, though tested at different times and in different places. Every effort should be made to insure that all men perform to the best of their ability. Standard conditions should therefore be optimal conditions.

b. Tests should be administered and scored in a manner identical with that employed in their standardization. Standardization (ch. 2) involves the administration of each test to a sample group of men with known characteristics in order to obtain norms by means of which each subsequent score may be evaluated and interpreted. In other words, test performance in the field is evaluated by comparing it with the performance of the men in the standard reference population. If this comparison is to be a valid one, the administration and scoring should be identical in the two instances. Army tests are always standardized under conditions which can be duplicated in the operational situation. The principles set forth in this chapter should be followed in order to duplicate the standardization conditions as closely as possible.

Section II
PRINCIPLES AND PROCEDURES FOR ADMINISTERING GROUP TESTS

165. General

It has already been stated that the procedures for administering tests should be such as to call forth the best

performance of which the individual is capable under standard conditions. Each individual will tend to do his best if his environment is reasonably free from distracting influence, if he understands what he is to do, and if he considers it worthwhile to do his best. The first of these conditions depends upon the physical aspects of the testing situation; the others upon the techniques employed by the examiner in controlling the testing situation.

166. Physical Surroundings

All behavior, including test performances, takes place in an environment. Since it is impossible to administer tests in a "vacuum," the next best thing is to take steps to insure that the environment provided is standard for all administrations of the test, and that it does not impede or hamper the performance of the examinee. While it is recognized that ideal testing conditions cannot always be achieved with the limited facilities available in field installations, attention to the following factors should provide conditions that are adequate in most cases.

a. So far as possible, the testing room should be quiet. Excessive noise is one of the principal sources of distraction. Noise which continues steadily at a moderate and fairly even level of intensity can be considered normal for testing conditions. Such noise might include the steady hum of indistinguishable voices from another part of the building, the drone of machines, or the continuous but muted clatter of typewriters. But a sudden shout outside a window, a bell, the clatter of unloading a truck, the blare of a radio, or the sound of persons passing through the room are very distracting to examinees.

b. Consideration should be given to the acoustics of the testing room. The examiner's voice must be clearly audible to all men being tested. The public address systems now found in most Army testing rooms have solved this problem; however, care should be exercised in placing loudspeakers and in locating microphones. The level of amplification should also be controlled. Loud directions booming forth above one's head can be very disconcerting.

c. The testing room should be well lighted and ventilated. There must be sufficient illumination on the working surface to prevent eye strain. If a light meter can be obtained, the illumination in various parts of the room should be checked. A light meter laid on the working space should register 6 to 10 foot–candles. Special care should be exercised to avoid glare spots and shadows; there is perhaps nothing as annoying as having part of the test paper intensely illuminated with the rest in the shadow cast by a pillar, a partition, or the examinee himself. Conditions of temperature, humidity, and ventilation are sometimes difficult to control, yet every effort must be made to do so. No one can perform at his maximal efficiency in a room where the air is hot, sticky, or stale.

d. Among the foremost factors of good test administration is the physical arrangement of the testing room. The examiner should be provided with a raised platform or rostrum in a part of the room where he can see, and be seen, by all men being tested. Desks or tables for the examinees should be arranged to leave aisles for the proctors' use in distributing and collecting test materials and to permit circulating about the room during the test. If possible, there should also be enough space between rows to allow passage. The writing surface itself should be flat and smooth and free from cracks. If the only available tables are rough, a tight covering of linoleum or press board (masonite) should be used. The space allotted to each individual must be wide enough to accommodate both a test booklet and a separate answer sheet. Chairs with writing arms should not be used for testing since the writing surface usually provided is far too narrow. Large tables with vertical partitions separating the surface into booths approximately 30 inches, wide and 18 inches deep are highly desirable. If tables like these are not available and cannot be constructed, mess tables make an adequate substitute.

e. The temptation to give or to receive aid always seems to be present wherever people are examined in groups. The use of partitioned booths or of alternate seating will help to prevent collaboration. Despite all precautions, the proctors will still have to prevent cheating during the examination. For this reason, proctors should circulate (as quietly as possible) rather than remain at a fixed post. The mere nearness of the proctor on his rounds is often a sufficient deterrent to cheating.

f. Not all distracting influences are in the external surroundings. The physical and mental condition of the individual also affects his test performance. The man who has just had disturbing news from home, or is in physical distress is in no condition to do his best on an examination. In individual cases, these factors cannot always be foreseen, but for the group as a whole, much can be done by scheduling testing sessions at a time of day when fatigue or physical or emotional discomfort can be expected to be at a minimum. Normally, morning is the best time to schedule an examination and the end of a long day the poorest. Where possible, activities should be controlled so as not to interfere with testing schedules. In the reception station, for example, processing should be so regulated that testing does not follow hard exercise, long hours of waiting in line, or immunization shots. In all cases the test officer, examiner, and proctors should be alert to signs of genuine distress, and the affected persons should be excused until a more propitious occasion.

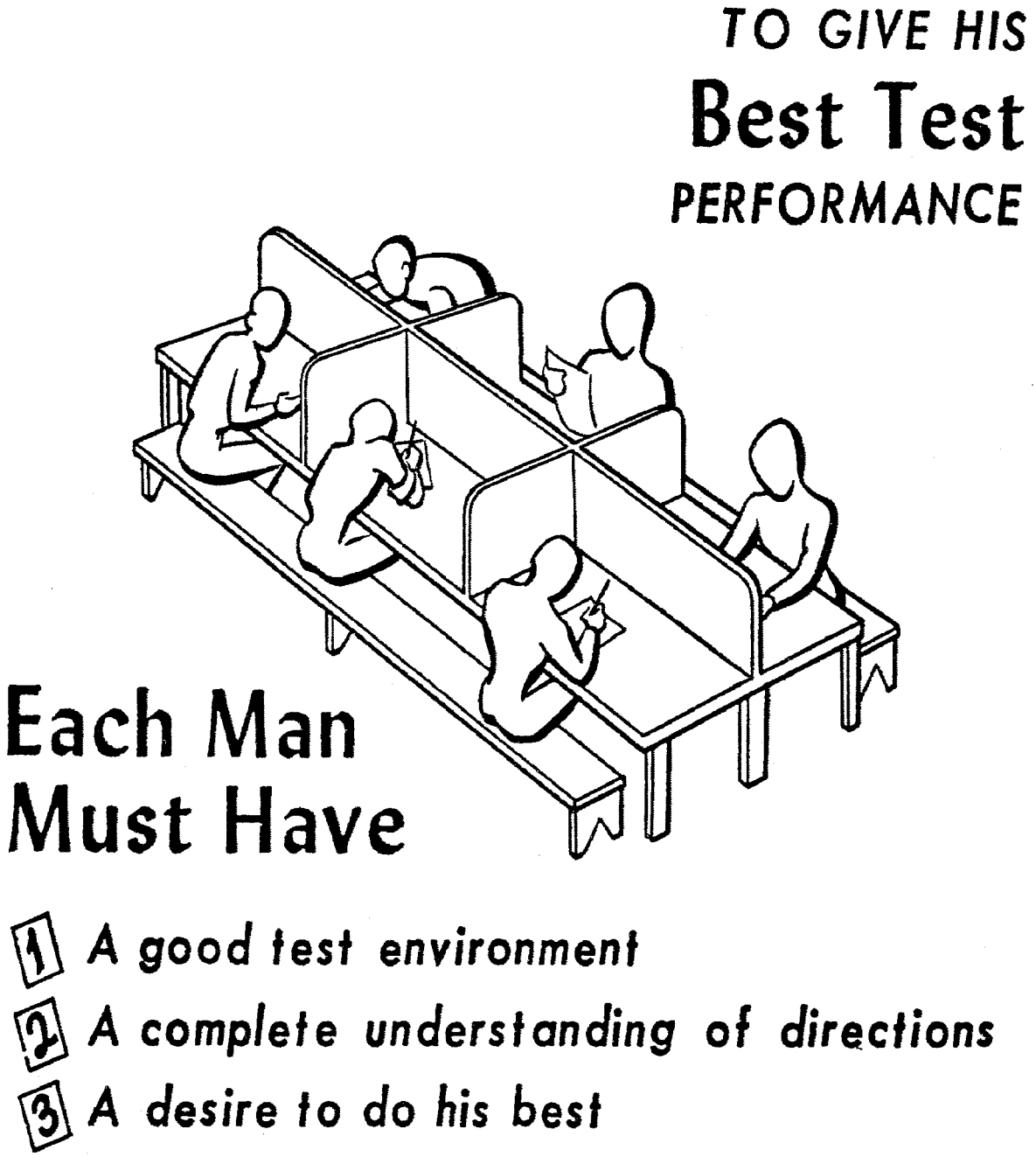


Figure 22. Factors affecting test performance

167. Testing Session

Although the major portion of the time is allotted to taking the test itself, all other activities, such as assembling and seating the men, distributing materials, giving preliminary directions, and collecting materials must receive their share of attention. Yet it is the management of then details which can make of the session either a smooth running efficient operation or chaotic confusion. Control is best achieved by careful preparation and practice in all phases of the process—preliminary arrangements, test administration, and the collection and disposition of materials. The discussion that follows will cover general principles and specific suggestions for making the testing session an orderly, systematic affair.

168. Preparation for the Testing Session

a. General. Preliminary planning for the testing session involves the careful selection of the testing team, the instruction of all members, and practice drill in all required testing procedures. The examiner is selected for the quality of his speaking voice and for his ability to handle groups of men. While no one demands that he have precise diction, he should speak so that he can be readily understood. It is also desirable that the examiner be capable of controlling the testing situation—by virtue of his personality, authority, and prestige.

b. The Examiner's Preparation. The examiner should make a careful study of the manual to make sure that he knows the purpose of the test, the materials needed to give it, the directions to be read, and the problems which are likely to arise. He should study those directions which are to be read aloud until he can read them smoothly. Familiarity with the contents of the test itself is also invaluable. It is excellent practice for both the examiner and all proctors to take each test in the prescribed fashion before attempting to administer it; this procedure should be standard whenever a new test is installed or new examining personnel are trained. In this way, the examiner gains an appreciation of the men's viewpoint on the test and learns how to anticipate, and thus be prepared for, the common questions which may arise.

c. Duties of the Proctors. The examiner is responsible for instructing the proctors thoroughly in their specific duties. The common practice of snatching any man not at the moment occupied, and making him a proctor, then and there, should be discouraged. It is far more efficient to designate regular testing teams responsible for the administering, proctoring, and searing of tests. Each proctor should be assigned a certain section of the room for which he will be responsible. Before the testing period, he should check the materials to be used to make sure that they are in good condition and order, and in sufficient quantities. He should know the order in which these materials are to be distributed and collected, SO that, when the time comes, he can execute this phase of his assignment efficiently. With the administration of the test itself, his real job begins. While directions are being read and while the test is being taken, he should patrol his assigned area. Within this arm, he is responsible for—

(1) Seeing that each examinee has all the necessary materials for taking the test, and furnishing these, where needed.

(2) Insuring that each examinee is following the directions correctly and Understands what he is to do and how he is to do it, The proctor should be alert to detect incorrect methods of marking answers where separate answer sheets are employed.

(3) Seeing that each examinee is doing his own work, independent of his neighbors.

(4) Excusing from the examination any person who is or becomes too ill to continue without discomfort.

169. Administering the Test

The value of any test score will depend on the extent to which the examinee understands just what he is to do and the degree to which he considers it worthwhile to do his best. The examiner's primary responsibility, in fact his main function, is to elicit this willingness to work and to provide the proper instruction. The first is a matter of the appropriate stage setting and of creating favorable attitudes. The second is handled by proper utilization of the oral directions contained in the manual for administering the test.

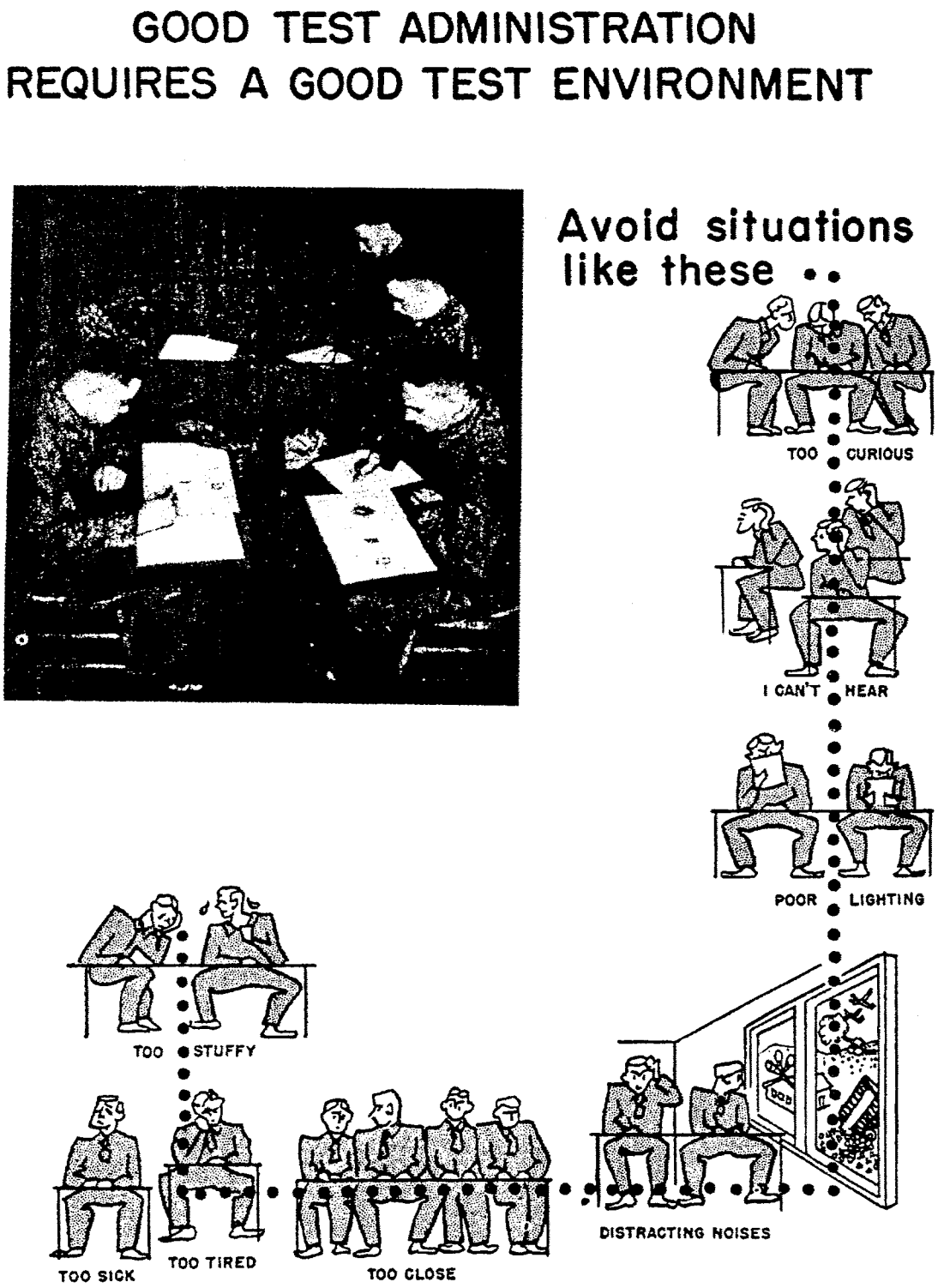


Figure 23. Some conditions detracting from good test environment

a. The proper stage–setting is a matter of appropriate physical surroundings and thoroughness of advance preparation. Equally important is the climate created by the behavior of the examiner and proctors during the testing session. Their dress, demeanor, bearing, and speech should tend to create a climate with just the right degree of seriousness, formality, and order. Examining personnel should be both courteous and in full control of the situation at all times. When the examinees enter the testing room they should be seated in an orderly fashion. Before the examiner starts giving the instructions for taking the test, any housekeeping details that could possibly be distracting should be disposed of for example, announcements by troop commanders or instructions for the storage of excess clothing and canteens.

b. When all preliminary matters have been disposed of and the examinees are ready for the test instructions, the examiner greets the examinees and introduces himself. He then makes a brief informal statement explaining the test to be given, how the results will be used, and why it is important for each person to do his best on it. The aim of these remarks is to dispel anxiety and release tension, and at the same time, to stress the necessity for maximum effort and output. A careless presentation may create the impression in the minds of the men that the test they are about to take is of no consequence and in no way related to their future Army careers. A too emphatic presentation may so impress them with the seriousness of the situation as to give rise to disturbing tensions. On the whole, the best, results will be achieved through a brief, straightforward but nontechnical statement of facts, delivered in a straightforward manner.

c. Having set the stage and gained the necessary cooperation, and having distributed the test material, the examiner next informs the examinees what they are to do. There is only one way to do this—by reading aloud the directions provided in the manual. And it means reading aloud all the directions that are to be read aloud and no more. The test administrator should not make the mistake of reading aloud to the examinees any of the explanatory statements in the manual which are intended solely for the test administrator. Directions should be read in a natural voice, in a smooth, coherent fashion. Hence, they must have been thoroughly practiced. Notice, however, that they should be read, not paraphrased, given from notes or memory, or adapted to someone's idea of what is more appropriate for local conditions.

d. Many Army tests are given with certain time limits which must be strictly observed if testing conditions are to be uniform from session to session, and from place to place (par. 81). These time limits, either for the complete test, or for various parts of the test separately, are always specified in the test manual. They are exact, not approximate; so timing should be handled with care. If a stop watch is available, it should be used. If not, any good watch with a second hand will serve, if used in the following manner:

(1) When giving the signal to start the test, jot down on paper the hour, minute, and the second of starting.

(2) Write below this time the hours, minutes, and seconds of working time for the test as specified in the manual.

(3) Add these two figures to obtain the exact time when the signal to stop work should be given. (If the minutes add up to more than 60, of course, the 60 minutes would be carried as an additional hour and the excess listed as minutes.)

Example:

Starting time	1451:00
Time limit for test	45:00
Stopping time	1496:00
	or
	1536:00

The signal to stop should, therefore, be given promptly at 1536. The timing should always be done in this way. It is unwise to trust to memory or to attempt the necessary computations mentally. And it is good practice to have some of the proctors check the timing independently.

170. Collection and Disposition of Test Materials

After the signal to stop work has been given, the materials should be collected as quickly as possible. The period between tests or at the end of a session can be one of tremendous confusion with everyone talking, comparing notes, reaching for coats and hats, and being anxious to leave. Under these circumstances, the test materials tire apt to be collected in haphazard fashion, with booklets and answer sheets jumbled together. In such confusion test materials may disappear. Since all test materials must be strictly accounted for, it is essential to establish procedures for the orderly collection of materials. The following system achieves a maximum of order and control.

a. As soon as the stop signal is given, the examiner should instruct the men to remain quietly in their seats and to follow directions in order to expedite the collection of materials.

b. He should then direct the examinees to pass these materials to the ends of the rows, specifying which end. The materials should be passed separately, first answer sheets, then test booklets, then supplementary materials such as scratch paper, and finally, pencils.

c. The men at the ends of the rows should be instructed to stack the materials in separate piles, making sure that all booklets are closed, with the cover sheet outside, and that all answer sheets are faced the same way.

d. The proctors can then collect the materials and at the same time make a count of the numbers turned in. Only after all materials have been checked in and accounted, for should the group be dismissed or the next test begun.

e. On some occasions, examinees may be permitted to leave the testing room individually before the time limit is up. Care should be exercised to insure that these men do not leave until they have complied with all directions in the manual for administration and that they have turned in all of their testing materials. If, for any reason, an examinee is permitted to leave the testing room temporarily during a testing session, all materials on his desk should be checked for completeness.

171. Care of Booklets

After each testing session, and certainly before the next session, all test booklets should be carefully scrutinized for answers or marks of any kind. In spite of all warnings, some persons will write answers in the booklets or use them for scratch paper. If the answers or marks can be erased, this should be done. If not, or if the booklet is worn or torn, it should be destroyed in accordance with proper procedures for handling classified material. All used scratch paper should be destroyed. All tests and testing supplies should be kept secure according to the appropriate security classification when not in use.

172. Answering Examinees' Questions

As a rule, questions asked by examinees prior to the actual time of beginning the examination are answerable. These questions usually pertain to testing procedures, time limits, and the purposes and uses of tests. After the examinees have begun the test, test administrators and proctors should be especially careful to avoid revealing information that might influence the proper evaluation of an individual, as in the following circumstances:

a. If an examinee asks the test administrator or proctor to tell him the answer to a specific test question, the answer to him should indicate that the examiner is not permitted to tell the examine the correct response to a test question since that would not allow the test to do an accurate job.

b. If an examinee states that he does not understand what a question means and asks the test administrator or proctor to explain it to him, he should be told that to do so would influence his responses and thus not permit the test to give a true evaluation. The examinee should, however, be encouraged to do the best he can.

c. Even though some of the questions presented by the examinee seem "stupid" or appear to be repetitious, impatience or contempt should not be displayed by the test administrator or proctor. In fact, an unusual number of questions may be an indication that the test administrator is not following directions, is not speaking loudly or clearly, or that acoustics are unsatisfactory. If such is the case, corrective action should, of course, be taken.

d. In general, the examiner should try to answer as many questions as possible without giving away answers or providing clues to answers.

Section III
ADMINISTERING INDIVIDUAL TESTS

173. General

For the most part, evaluation instruments employed by the Army are of the paper–and–pencil type administered to groups of examinees. In certain special circumstances, proper evaluation of the soldier will necessitate the administration of an individual test. Individual tests, not to be confused with group tests which can be administered individually, are administered in a more personal and less formal manner than are group tests. The expenditure of time and effort is greatly increased with the employment of such instruments, and they should be used only when specified by appropriate regulations. In general, the individual type of test is recommended for cases in the following categories:

a. Where the paper–and–pencil type test is inappropriate because the examinee is lacking in the educational skills of reading and writing.

b. Where more personal contact is needed to insure that the examinee is at ease, is properly motivated and encouraged, and knows just what lie is supposed to do.

c. Where it is desired not only to determine the individual's overall score, but also to give the examiner the opportunity to observe him at work and to estimate his specific strengths and weaknesses.

174. Individual Testing Session

The individual testing session is a more personal and somewhat less formal affair than the group testing session. This does not make it easier to manage; on the contrary, the individual test administrator needs much more than average training and experience to achieve results which can be accepted with any confidence. Men assigned to this job must be selected with care. The ideal examiner is a man with a knowledge of the principles of psychological measurement and an appreciation of the needs for exactness and precision. He is personable, friendly, patient and tolerant, never given to a show of arrogance, flippancy, or sarcasm, no matter how absurd the response of the subject might be.

a. Individual tests, because they are essentially personal interviews, should be given in an atmosphere of privacy. Special rooms are recommended, but separate booths divided by partitions will serve where facilities are limited. Examiner and examinee should be provided with comfortable chairs facing each other across a table or desk. This table or other working surface should be large enough to accommodate all of the test materials and provide room for the examiner to jot down responses on the record sheet. The field table is of the proper dimensions for most individual testing.

b. Before undertaking the administration of an individual test, the examiner should make a careful study of the manual of directions and of the test materials. The exact wording to be used in presenting the materials will be specified, and should be rehearsed until it can be read in a normal conversational manner. The examiner should also practice the things he is to do—the placement of the materials, movements, pointing, demonstrations, etc.—until these are smoothly coordinated with the verbal directions. Finally, he should give practice administrations under the supervision of a qualified examiner until he can maintain the examinee's interest and confidence, select and use necessary materials and instructions without fumbling, and make of the whole procedure a smooth and effective performance.

c. The administration of an individual test begins as soon as the examinee enters the room. The first step is to get him into the proper frame of mind for taking the test, i.e., to remove fear and tension which may conceal qualities valuable to the Army. The man reporting for the test is quite apt to be afraid, discouraged, misinformed, or antagonistic and in no condition to perform in a fashion that can be considered a trustworthy indication of his true ability. The skillful examiner should greet the examinee in an affable manner, ask him questions about himself, about his work, listen to his complaints, and give every indication of being genuinely interested. He will often have to call upon all his skill and patience to carry this off without creating an air of playacting.

d. The transition from this informal chat to the presentation of the test materials should be gradual and natural. The test manual will contain suggestions for bridging this gap, or such statements as "I have some problems here I would like you to try," or "Let's see what you can do with these questions," will serve the purpose. From this point on, the presentation of the problems, the questions, and all directions to the examinee must follow the exact wording of the manual. Moreover, any performance materials, such as blocks or pictures, must be placed on the table or exposed precisely as specified, and the different parts of the test must be given in order, without skipping around. It cannot be overstressed that any departure from the manner of administration in which the test was constructed and standardized will make the test score unreliable.

e. The examiner should speak distinctly and slowly while administering the test so that the examinee may hear and understand, for the examiner may not repeat any question (unless some unexpected disturbance has prevented his being heard the first time). He must guard against gestures, words, or inflections of the voice that may suggest an answer. Throughout the course of the examination, the examiner will have to motivate the examinee to do his best. He will have to make appropriate remarks of approval or praise after each success, and he will have to console and encourage him when he fails. Suggestions for such appropriate remarks will usually be included in the manual for administering the test.

175. Timing

With the individual test, the problem of timing is usually somewhat more difficult than with group tests. This is so because the problems and questions which comprise the individual type of test often have separate time limits, and these limits are often specified in terms of seconds rather than large intervals. Furthermore, many of the items of an individual test are scored in terms of the time required to arrive at the correct solution. This means that the examiner will have to look at his watch frequently. Yet, because of the close personal nature of the situation, the examiner cannot be too obviously engrossed in the time problem without creating a disturbing and distracting tension in the examinee. Everyone has experienced the feeling of nervous strain when working against time and the maddening tendency of fingers to become all thumbs as the seconds tick off. Some of this tension is natural, of course, when one is told to work as rapidly as he can, but it may be tremendously heightened if the examiner is a nervous clock watcher. So, the timing should be done unobtrusively and with the appearance of casualness. This does not mean, however, that it can be slipshod. It is of utmost importance that the timing be precise and that the exact limits specified in the manual be observed. Where tests are scored in terms of time, an error of a few seconds may account for a difference of several score points. It is essential that the examiner should be thoroughly practiced in timing. If possible he should use a stop watch, because this will always start at zero, and because the timing can be done with one hand, leaving the attention of the examiner focused on the test itself. If a stop watch is unobtainable, an ordinary watch may be used. It should be of the type equipped with a large sweep second hand for ease of reading, and the examiner should always give the starting signal for any item when the second hand is at zero.

Section IV
SUMMARY

176. Importance of Proper Administration

a. The development of an instrument includes the method of administering the instrument. The value of the instrument may be lost if its administration is varied.

b. Proper administration requires careful attention to the following:

(1) Physical surroundings.

(2) Organization and conduct of the testing session.

(3) Answering questions.

c. Administration of individual tests requires more competent administrators than administration of group tests. Because individual testing must seem informal, special precautions are needed to maintain standard conditions.

176B. Title not used.

Paragraph not used.

Chapter 12
SCORING ARMY TESTS

Section I
SOME GENERAL CONSIDERATIONS

177. Importance of Accurate Scoring

Technicians scoring an instrument must understand the procedures and be careful to follow them exactly at all times. Long experience has shown that wherever people work with numbers, mistakes can occur very easily. Even the best workers must be constantly "on their toes" to avoid errors in numbers. Test scores may be used to make recommendations and decisions important to individual soldiers. Sometimes a man's entire military career and his usefulness to the Army may be changed because of a high or low score on a test. Therefore, knowing that most of the scores on a test are correct is not enough; we must make sure that every score is correct. Accurate scoring involves—

a. Complete understanding of scoring procedures.

b. Conscientious following of scoring procedures.

c. Checking the scoring and recording the scores.

177B. Title not used.

Paragraph not used.

Section II
SCORING PROCEDURES

178. What Is "Scoring"

Basically, "scoring" means comparing the answers which a man makes to the questions on a test with a certain specific pattern of answers, and counting the number of instances of agreement. The number of instances of agreement is called the "score." The specific pattern of answers is called the "scoring key." Usually the scoring key is merely a way of indicating the correct answer to each question, so that the "score" is then merely the number of correct answers. Exceptions to this are scoring keys for personality and interest tests, where there are no right or wrong answers; the scoring key, in these cases, indicates the answers which are related to the criterion used in validating the instrument.

179. Meaning of the Scoring Formula

a. When the score on a test is the simple count of correct answers (or of the number of instances of agreement with the scoring key), the score may be referred to as a "rights" score. Some tests, for reasons discussed in chapter 2, are scored by counting the number of correct answers and then subtracting from this figure the number of wrong answers, or some fraction of the number of wrong answers. Such scores would be called "rights minus wrongs" scores, or "rights minus one–fourth wrongs" scores, or whatever the case may be. This statement of how the final score is arrived at is called the "scoring formula" for the test. It is stated in the test manual in the section oil scoring instructions. It appears in abbreviated form on the scoring key as—

"R" for rights scoring

---

* The fraction is always 1/(n-1) where n is the number of alternative to each item.

"R–W" for rights minus wrongs scoring
"R–?W" for rights minus one–third wrong scoring and so on.

b. Applying the third formula above to the case of the four–alternative multiple–choice items in a test of 100 items, let us see how correction for guessing operates. Let it be assumed that two examinees each know the answers to 50 items of a test, but that whereas one of them stops at this point, the other goes on to make pure guesses on the next 20 items and, by chance, gets five of them right and fifteen wrong. The "Rights" score of the first examinee will be 50 and that of the second, 55. Application of the scoring formula to both cases, however, will give the first man (50 minus 0) or 50 and the second man (55 minus ? of 15) or also 50. It is important to note that the fraction in the formula depends upon the number of alternatives to each question.* For a test composed of items having five answer choices, the formula would be "rights minus one–fourth wrongs."

c. The scorer must be careful to use the proper scoring formula and to compute it correctly.

Caution: **He must be especially careful when certain questions are omitted. The number of wrong answers cannot be computed by subtracting number right from total number of test questions; number wrong, in this case, is the sum of number right and number of omits subtracted from total number of test questions. Omitted questions must not be counted as wrong.**

180. Hand Scoring and Machine Scoring

Some tests must be hand scored, and others, when special answer sheets and special pencils are used, may be scored by means of the International Test Scoring Machine. The latter procedure is referred to as "machine scoring." Both methods make use of a scoring formula, but with the scoring machine any deduction for proportion of wrong answers can be made automatically. Not all tests can be machine scored; nature and format, of the test sometimes necessitate hand scoring. However, all machine–scorable tests may be hand scored, if a scoring machine is not available.

181. The Scoring Team

a. Whether scoring is done by hand or by machine, a number of steps must be performed in orderly sequence. When several tests are to be scored at one time, efficient scoring requires a team of scorers with a different individual assigned to each of the successive steps in the process. Members of the team complete only the operations assigned, passing the papers along to others who perform succeeding operations.

b. With hand–scored tests of the type with answers oil the test page itself, it is efficient to have each scorer handle a single page, writing the score at the bottom of the page. Other members of the team should be designated to add together the scores for each page and change total scores into converted scores; still others should check each step in the process.

c. When tests are machine scored, the following steps can conveniently be done by different members of a machine scoring team:

(1) Scan answer sheets.
(2) Operate the scoring machine.
(3) Spot check scoring by hand.
(4) Hand score papers rejected in the scanning process.
(5) Change raw scores to standard scores.
(6) Check conversion to standard scores.

182. Checking

The importance of checking at all points cannot be overemphasized. A test that merits an hour of the examinee's time is worth the additional seconds to insure that the score is an accurate one.

Section III
HAND SCORING

183. Hand Scoring Test Booklets

Tests in which the answers are made directly on the booklet itself are expendable, as contrasted with nonexpendable tests in which separate answer sheets are used. With expendable tests scoring keys, scoring stencils, or a list of correct answers on a strip may be provided. The test manual will contain specific instructions for use of these keys. These instructions must be followed exactly, and all scoring steps checked.

184. Hand Scoring Separate Answer Sheets

Many Army tests make use of separate answer sheets oil which all answers are indicated. These test booklets are nonexpendable since they can be reused with new answer sheets. On most of these answer sheets, spaces are provided for possible answers by means of boxes, letter blocks, or pairs of dotted lines. The examinee then records his answers by marks in these specified spaces. With tests of this type, scoring keys are usually in the form of stencils with holes

punched in the positions of the correct answers. Marks showing through the holes when the stencil is placed over an answer sheet are correct answers. Care must be taken to line up the stencil with the edges of the answer sheet, or preferably with two fixed "landmarks" on opposite corners of the sheet. The procedures outlined below for hand scoring separate answer sheets have proved valuable over a period of years and are strongly recommended.

a. Scan (that is, look over) each answer sheet and draw a red pencil mark horizontally through all response positions for each question to which the examinee has indicated more than one choice or no choice. When an examinee has not completed the test, draw a red line through all response positions for every question following the last one marked on the answer sheet. The sum of the red marks will be the number of omissions.

b. Place the punch–hole stencil, right side up, over the answer sheet, and line it up with the edges of the paper or the landmarks. Count the number of marks made by the examinee which appear through the holes of the key, excluding all marks with a red line running through them. The sum of these marks will be the number of *rights*.

c. With the key still in place, count the holes where no mark appears (neither black nor red). The sum of these unmarked positions will be the number of *wrongs*.

d. If the scoring formula is merely rights, only steps a and b need be performed (stop c will, however, provide a check, as shown in the following paragraph). If the scoring formula calls for subtraction of a proportion of the number of wrongs from the number of rights, the figures obtained in b and c will be entered into the proper formula to obtain the raw score.

e. Make the following check: the number of *omission*, plus the number of *rights*, plus the number of *wrongs* should equal the total number of items on the test (exclusive of practice items).

Caution: **Sometimes when a standard answer sheet is used, the test will not have as many questions as the answer sheet provides for; in these cases the scorer must be careful to count only the spaces allotted to the test questions.**

Section IV
MACHINE SCORING

185. Requisites for Machine Scoring

a. Machine scoring is generally more accurate than hand scoring. In order to achieve the advantages of speed and accuracy in machine–scored tests, three requisites must be met:

(1) Examinees must properly record their responses on the answer sheet.

(2) The scoring machine must be in proper functioning order.

(3) The operator must set up and manipulate the machine correctly.

b. All three requisites can be readily fulfilled— the first by care in administration and proctoring when the test is given; the second, by systematic checking of the machine; the third, by using trained, conscientious operators to do the scoring.

186. How the Machine Obtains a Score

The underlying principle of the International Test Scoring Machine is simple. The graphite deposited by a special lead pencil in making a mark on paper will conduct an electric current. If two wires from a source of power are pressed against such a mark, the circuit is completed. The current is carried from one wire through the mark to the other wire and causes a deflection of the needle of a galvanometer connected in series in the current. If there are hundreds of these simple circuits, all connected to the same galvanometer, all of those which are closed by means of pencil marks add to the current flowing through the galvanometer. In other words, the amount of the deflection tells how many of the circuits are completed. In a sense, therefore, the galvanometer reading is a count of the number of pencil marks. If the answers to a test are indicated by pencil marks with graphite in them in a specified place on an answer sheet, and if this answer sheet is then pressed up against a mass of open end circuits (or electrodes), the dial of the galvanometer will register the number of such marks. If a punched–hole scoring stencil is inserted between the answer sheet and the electrodes, the current carried by the "right" pencil marks (represented by the punched holes) can be routed one way, and the current carried by the remaining marks routed another way. Thus, the meter dial can be made to register the number of right answers, the number of wrong answers, the number right plus the number wrong, and finally the number right minus any portion of the number wrong.

187. IBM Manual of Instruction

It is urgently recommended that all personnel involved in machine scoring of tests become thoroughly familiar with the manual entitled, "Principles of Operation, IBM Test Scoring Machine, Type 805". Detailed explanations of the principles and operation of the machine need not be repeated here. Particular attention is called to the chapters of the IBM manual on "Principles," "Scoring Set Up Procedures," and "Recommendations for Supplies."

188. Scanning Answer Sheets Before Scoring

a. Completed answer sheets must meet certain standards (figure 24 illustrates some of the difficulties encountered when answers are improperly marked).

(1) Marks must be made on the answer sheet with a special pencil, that is, one with high graphite content.

(2) Marks must be heavy and black enough so that current can flow through them.

(3) There must be no stray marks not intended as answers; any such must be erased since they may also bridge a circuit and produce a score.

(4) Erasures must be clean; remainders of marks may produce a score.

(5) Marks must be placed properly in the designated space (letter blocks, boxes, or pairs of dotted lines). Contacts in the machine line up with these response positions, and marks placed outside the response positions cannot complete the electrical circuits.

(6) Only one response should be marked for each question.* Additional marks may be picked up, incorrectly, by the machine as either right or wrong answers, depending on the scoring procedure.

(7) Answer sheets must not be badly wrinkled or torn. Damaged answer sheets will not fit properly into the position channel of the scoring machine and will not yield accurate scores.

b. Test scorers should not assume that personnel administering the test were able to make sure that every examinee followed the rules perfectly. The first step in the scoring process should be scanning—inspecting the papers with the above points in mind.

c. In most cases, it will be most efficient to lay aside all papers shown by the scanning to be improperly marked; these papers should later be hand scored.

d. Occasionally, a group of papers will fail to meet requirements in only a few instances; then it may be best simply to "fix up" these few papers, as follows:

(1) Re–mark, with the proper pencil, responses made with wrong pencils.

(2) Darken marks that are too light.

(3) Erase stray marks.

(4) Clean up poor erasures.

(5) Erase all answers to questions with more than one answer given.

e. From the explanation of the scoring process, it should be clear that only specially printed answer sheets can be used for machine scoring. The location of the letter blocks or pairs of dotted lines on the paper must be extremely precise so that they will line up with the electrical contacts when the paper is placed in the machine.

---

* Exception:
A few personnel tests call for multiple answers to test questions. Consult specific test manuals for scoring procedures for these tests and keys.

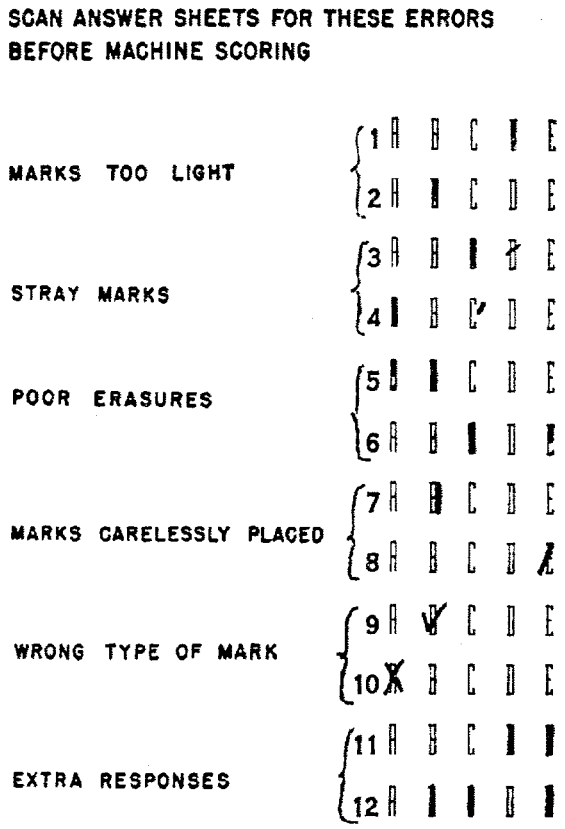


Figure 24. Errors frequently found on answer sheets

189. Setting Up and Balancing the Scoring Machine

a. Inserting Scoring Key.

(1) For most tests, only a Rights scoring key is used. The following procedures apply with tests having only a Rights scoring key:

(a) Remove scoring key frame from the machine. Place the appropriate Rights scoring key, printed side down, on the top leaf of the scoring key frame with the narrow margin of the key toward the hinges of the frame. The holes in the key should line up with the holes in the frame.

(b) Close the frame; insert it into the test scoring machine so that the top leaf is toward the back of the machine. Then move the frame into place by turning the key clamp lever.

(2) For tests with both a Rights scoring key and an Elimination key, the following procedures apply:

(a) Remove scoring key frame from the machine. Place the Elimination key, printed side down, on the top leaf of the scoring key frame with the narrow margin of the key toward the hinges of the frame. Place the Rights key, printed side down, on the bottom leaf of the scoring key frame, with the narrow margin of the key toward the hinges of the frame. All holes in the keys should line up with the holes in the frame.

(b) Proceed as above, closing frame, inserting into machine, and moving frame into place.

b. Complete Machine Check. A complete check of all machine sensing contacts can be made by use of field check sheets described below. The machine should be checked with these field check sheets at least once during each day the machine is in use and oftener if called for by machine difficulties.

(1) Use 10 machine–scored standard answer sheets with 150 5–choice response positions. Mark with a special pencil all 75 response positions for field 1 (answer spaces for questions 1 through 15) on the first answer sheet. Mark all 75 response positions for field 2 (answer spaces for questions 16 through 30) on the second sheet. Continue marking the

75 response positions for each field on separate answer sheets through field 10. Every possible response position will then have been marked on some one of the 10 sheets.

(2) To check the sensing contacts and the Rights circuit on the A, B, or C fields, set the master control switch to either A, B, or C, and the appropriate formula switch to "R". No scoring key is used for this operation, but the scoring key frame is in place and the key clamp lever is wound in. Insert a field check sheet into the machine. The dial reading should be 75. Insert the other 9 field check sheets separately. The dial reading should be 75 for each check sheet. This procedure insures that all of the sensing contacts are correctly sensing the pencil marks and the Rights circuit is operating efficiently.

(3) To check R–W circuit and also adjust for a R–W/3 scoring formula, insert a scoring key which has been punched in all field control positions (A, B, and C) at the top and bottom of the scoring key (no answer positions punched). Set the master control switch to "A" and set the A formula switch to "W". Insert a field check sheet into the machine and the dial reading should be 75. Set the A formula switch to "R–W". The dial needle should drop below zero. Pull the negative score key on and adjust the "A–" potentiometer so that the dial reading is 25 (? of 75 wrongs). Score the other 9 field test sheets, pulling the negative score key on; the "W" score should be 75 and the "R–W" score should be 25 with no further adjustments. The same procedure should be used to check the B and C scoring fields.

(4) To adjust the machine for "R–W/4" scoring formula, a set of check sheets is needed with the number of marks divisible by 4. For example, 60 response positions can be marked in the separate fields. Then, follow the above step (3) and adjust the machine for R–W/4. The dial reading at "W" should be 60, and at "R–W" should be 15, in this example.

(5) To adjust the machine for other scoring formulas, follow the same general procedure.

c. Checking the Machine for a Particular Test.

(1) Scoring directions for each test describe the preparation and use of check sheets for that test. In general, these directions call for using unmarked answer sheets for that test, marking one answer sheet with all correct responses, marking another so that the score will be approximately the typical score for the group of papers to be scored, and marking a third so that the score will be a low score among the group of papers to be scored. For tests where wrongs appear in the scoring formula, the second and third check sheets should be given wrong responses as well as right, and the correct formula scores computed for all check sheets.

(2) Scoring keys are placed in the scoring key frame as described above.

(3) The master control switch is set to the field being used (A, B, or C). The formula switch for the field selected is set to the correct scoring formula.

(4) Insert, in turn, each of the check sheets, prepared as in (1) above. In order to score either "R," "W," or "R+W" adjust the " +" rheostat for the field being used so that dial reading corresponds to the score on the first check sheet. Dial readings should then also be correct for other check sheets. In order to score "R–W," "R–? W," "R–1/4 W," etc., first see that the separate "R" and "W" scores are read correctly by the machine as just described, then turn the formula switch to "R–W"' and adjust the "–" rheostat for the field being used so that the dial reading corresponds to the correct score on the first check sheet. Dial readings should then be correct for the other check sheets. When the machine is set so that it gives correct scores on all check sheets, leave the formula switch set according to the desired formula. The machine is now ready for scoring.

d. Points to Check when the Machine Does Not Yield Accurate Scores on Check Sheets:

(1) Are the scoring keys properly placed in the scoring rack?

(2) Is the scoring key frame in proper position in the machine?

(3) Is the key clamp lever wound in?

(4) Are marks on the check sheets heavy and well placed?

(5) Are marks on the check sheets made with the proper pencil?

(6) Are check sheets free of smudges and stray marks in unmarked positions?

(7) Are check sheets free from tears and frayed edges?

(8) Are the check sheets counted correctly?

(9) Are the answer sheets being inserted correctly?

(10) Is the power switch on?

(11) Is the cord plugged to a 110–volt outlet?

(12) Is the master control switch set to the correct field?

(13) Is the formula switch set to the desired formula?

(14) Is the feed channel clear? (Answer sheets occasionally stick in the feed channel.)

(15) Are the contact blades clean? (These can be cleaned by turning off the power switch and brushing with a long, thin brush provided by IBM for this purpose.)

Caution: **Be sure to turn off the power switch before brushing.**

(16) Are the answer sheets completely dry? (A heating unit is built into the machine to dry out answer sheets on humid days.)

Note. If the answer to all of these questions to "yes" and the machine still does not give accurate scores, an IBM serviceman should be called.

190. Scoring and Checking the Answer Sheets

a. Prior to machine scoring, hand score every 20th answer sheet in the group for the test being scored. Record the scores in the way specified in the test manual, usually in a particular box directly on the answer sheet. Keep these answer sheets in their original order within the group.

b. With the machine properly set, insert each answer sheet, one at a time, and score as specified in the test manual. The interspersed hand–scored answer sheets will be scored as they occur in the group, as a check on the accuracy of the machine. It cannot be assumed that the machine will remain in correct adjustment even though it met all checks at the beginning of the scoring run.

c. When a score has already been entered on the answer sheet and that score differs from the machine reading, the source of the discrepancy must be determined and corrected.

(1) If the score obtained by hand scoring is in error, the scorer will correct it and proceed with the machine scoring.

(2) If the score obtained by hand scoring is correct, adjust the machine according to the procedures given in paragraph 189. When it is necessary to adjust the machine, the preceding 20 answer sheets must be rescored to insure the accuracy of the machine scores.

191. Additional Checking of Machine Scoring

Since the machine can score so rapidly, it is usually possible to have all answer sheets scored a second time by a different operator to check the consistency of the two scoring runs. This is recommended wherever possible.

192. Summary of Steps in Machine Scoring

a. Scan answer sheets.

b. Set up the machine; check the balance of the machine.

c. Hand score every 20th answer sheet.

d. Score the answer sheets.

e. Hand score answer sheets which—

(1) Do not scan properly.

(2) Do not check.

Section V
RECORDING SCORES

193. Importance of Accurate Recording

Accuracy and legibility are the most important factors to be stressed in recording scores. Complete, independent checking of numbers copied from one source to another is recommended. Experience has shown, even among careful workers, a certain percentage of number reversals, omissions, repetitions, "switching" names and scores, and other types of errors. Numbers should be typed or hand–written with care. Corrections should be rewritten, not traced over the old numbers.

193B. Title not used.

Paragraph not used.

Section VI
SUMMARY

194. Accuracy in Scoring

a. The usefulness of an instrument may be seriously reduced unless proper precautions are taken to minimize errors in scoring.

b. The principal precautions are as follows:

(1) Understanding of scoring keys and procedures.

(2) Careful compliance with instructions on use of keys and procedures.

(3) Systematic checking.

194B. Title not used.

Paragraph not used.

Chapter 13
HOW THE ARMY USES PERSONNEL MEASURING INSTRUMENTS

195. Administration and Use of Army Personnel Measuring Instruments

a. The tests and measurements described in this pamphlet are the principal types of tools used by the Army to insure scientific military personnel management. Assume an Army job has been engineered. A man is selected for that job. Then he is trained, if he does not already have the required skills. Next he is put to work (assigned), and in many assignments proper work methods must first be established before the man can be considered "operational." Whatever the personnel operation—selection, classification, assignment, or utilization—and whatever the type of personnel management tool, personnel management seeks to promote maximum efficiency of individual and unit performance on a given job.

b. The starting point for scientific military personnel management for the Army was the creation of a tool for mental screening of manpower during World War I, when requirements led to the development of Army Alpha and Army Bets. These were measurement devices which could be applied rapidly and economically to large masses of manpower. During the last two decades, Army research agencies have been engaged in a variety of research projects to insure continuing progress in the proper use of manpower. The AGCT was an early product of this period; it did the job of mental screening during World War II. The Army Classification Battery came out of post World War II research effort, as did the Armed Forces Qualification Test (AFQT). The AFQT has the distinction of being the first psychological test and mental standard to be legislated by Congress. By 1950 it had been adopted by the Defense Department for use by all services. During this period, also, many devices for special selection and assignment programs have emerged, such as rotary and fixed–wing pilot selection, special warfare training, officer training, Arctic assignment, officer commissioning.

c. Typical personnel measurement instruments used operationally today are shown in figure 25—where and when the instruments are administered, how they are scored, how and for what purposes they are used. Generally these tests are the results of painstaking personnel research. All types of instruments discussed in this pamphlet are represented: mental screening tests, aptitude and ability tests, general culture and other types of achievement tests (including tests of proficiency in 35 foreign languages), self–report forms, ratings, and standardized measurement interviews.

196. The Modern Army a Challenge to Personnel Management

Constantly–changing concepts of present and future warfare have meant more complex involvement of personnel management techniques than formerly. The man on today's military job is required to interact more and more with his equipment. As a direct consequence, more rides on his success or failure than ever before. In the combat and support systems which must achieve greater fire capability, more effective surveillance capability, more rapid identification of enemy targets, and more accurate and rapid communications, the human element is unquestionably the most crucial element because it is the most variable element and hence the most uncertain. Personnel measurement techniques described in this pamphlet not only represent many of the principles and methods whereby much of this human uncertainty can be allayed, but can serve in promoting the overall efficiency of manpower needed for a stronger future Army.

Tests	Units at which normally administered	Program	Examinee or applicant pool	Qualifying standards
Enlistment Screening Test	Recruiting stations	Pre-enlistment screening of male enlistment applicants.	Enlistees who must be tested prior to traveling to AFES for AFQT testing.	Raw score of 28 (31st percentile) for men without dependents, and 33 (50th percentile) for men with dependents.
Armed Forces Qualification Test.	Armed Forces Examining Stations.	Screening of male enlistment applicants.	Enlistees who must be screened for mental acceptability prior to enlistment.	Percentile score of 31 for enlistees without prior service.
Armed Forces Qualification Test.	Armed Forces Examining Stations.	Qualitative distribution of enlistees and inductees.	All male enlistees and inductees who come under draft quotas who must undergo mental acceptability screening prior to entering the services.	Distribution of manpower to each service *AFQT percentile score* — % 93–100 — 11 65– 92 — 34 31– 64 — 43 10– 30 — 12 0– 9 — Rejected
Army Classification Battery.	Reception stations	Initial classification of enlisted male personnel.	Replacement stream enlisted male personnel.	On basis of aptitude area scores and interview information, interviewer recommends MOS for training. MOS may be awarded directly on the basis of school training or job performance.
SDB-Special Assignment	Any installation with classification and testing facilities.	Classification of personnel for Arctic assignment.	Enlisted personnel fully qualified in their MOS, considered for duty at Fort Churchill, Canada.	90 on Aptitude Area IN or a raw score of 65 on Self-Report Form.

Figure 25. Some typical uses of personnel measurement instruments

Tests	Units at which normally administered	Program	Examinee or applicant pool	Qualifying standards
Officer Candidate Selection Battery.	Wherever major commander designates an evaluation board.	Selecting male personnel in the active Army for OCS.	WO's and EM in the active Army; with Basic and Adv Indiv Tng, Apt Area GT 110+, OCT 115+.	Major commander selects on the basis of centrally determined quotas and cut scores and on basis of examining board recommendations.
West Point Selection Battery.	Wherever major commander designates an evaluation board.	Selecting members of regular components for admission to USMA entrance examination.	Enlisted RA Army or Air Force applicants for competitive appointment to USMA.	Selection based on relative score; personnel selected attend Prep Tng facility for 6 months prior to taking USMA examination.
Special Forces Selection Battery.	Any installation with testing facilities.	Selection of personnel for training and assignment in Special Forces.	Qualified EM in active Army and USAR.	Applicants selected by major commanders on basis of relative overall qualifications including composite scores.
ROTC Qualifying Examination.	ROTC institutions during Spring of Sophomore year.	Selecting cadets for senior division advanced ROTC course.	ROTC senior div (college level) cadets successfully completing basic course.	Qualifying raw scores announced annually
Educational Requirements Test, Interviews, and SDB.	Wherever major commander designates an evaluation board.	Appointing male officers on active duty to commissions in the Regular Army.	Male Reserve officers on active duty.	100 or higher on ERT; Scores on Interview and SDB are part of overall qualifications considered in the selection process.
Interview, Biog Info Blank.	Wherever major commander designates an evaluation board.	Appointing nurses to commissions in the Regular Army.	Unmarried nursing school graduates with Reserve commissions.	Scores on interview and biog info blank are part of overall qualifications considered in the selection process.

Figure 25. Some typical uses of personnel measurement instruments—Continued

Appendix A
References

The following selected references are provided for those who wish to increase their knowledge of the technical principles and methods described in this pamphlet. Other books, of course, may be used.

Section I
Required Publications

This section contents no entries.

Section II
Related Publications

PERSONNEL MEASUREMENT METHODS & GENERAL PSYCHOLOGICAL STATISTICS

Psychological Testing
Anastasi, Anne, Macmillan Company, New York, N.Y., 1961.

Essentials of Psychological Testing, second edition
Cronbach, Lee J., Harper and Brothers., New York, N.Y., 1960.

Selected References on Test Construction, Mental Test Theory and Statistics, 1929–1949
Goheen, H. W., and Kavruck, S., U.S. Government Printing Office, Washington, D.C., 1950.

Personnel Selection
Thorndike, R. L., John Wiley and Sons, Inc., New York, N.Y., 1949.

Measurement and Evaluation in Psychology and Education, second edition
Thorndike, R. L, and Hagen, Elizabeth, John Wiley and Sons, Inc., New York, N.Y., 1961.

Fundamental Statistics in Psychology and Education, third edition
Guilford, J. P., McGraw–Hill Book Company, Inc., New York, N.Y., 1956.

Psychological Statistics, second edition
McNemar, Q., John Wiley and Sons, Inc., New York, N.Y., 1955.

Elementary Statistical Methods
Walker, H. M., Henry Holt and Co., New York, N.Y., 1943.

Mathematics Essential for Elementary Statistics, revised edition
Walker, H. M., Henry Holt and Co., New York, N.Y., 1951.

Section III
Prescribed Forms

This section contents no entries.

Section IV
Referenced Forms

This section contents no entries.

Glossary

Section I
Abbreviations

This section contents no entries.

Entry not used.

Paragraph not used.

Entry not used.

Paragraph not used.

Section II
Terms

Note. Explanations of terms given here are intended only as aids in reading the text and not as comprehensive definitions of the psychological or statistical concepts involved.

Achievement test

A test of how much a person has learned about a subject or how skilled he has become in a line of work as a result of training or experience. Provides answers to such questions as—Does he do well enough now to be put to work on the job? Has he learned enough in a basic course to go on with advanced training?

Adjectival score

The qualitative statement of the rating of an individual or his performance, such as letter grades (A,B,C,D,E,) on school achievement, or descriptions of performance—"excellent," "good," "poor."

Administrative rating

A systematic evaluation of performance in an assignment or of other behavior; usually used to influence personnel decisions regarding the evaluee; usually a rating made by a superior of the ratee.

Alternate forms

When an instrument is in fairly wide use, it is common practice to have at least two forms, equivalent in content, which have been standardized to yield comparable scores. With such forms, the distribution, means, and standard deviations are almost the same. Often a single conversion table from raw score to standard score is used. Such practice provides a means of retesting men when necessary without their repeating the same test and also helps to prevent the content of a test from becoming familiar to examinees. Alternate forms of an instrument may also be administered to the same group of examinees for the purpose of estimating the reliability of the instrument.

Alternative

One of the possible answers supplied to a multiple–choice question of a personnel instrument. The person taking the test indicates in some prescribed fashion which of the alternatives he selects.

Aptitude

Readiness in acquiring skill or ability; the potentiality of becoming proficient, given the opportunity and appropriate training. The term may refer to the capacity to learn one specific kind of work or to general trainability.

Aptitude area

Term applied to a combination of Army Classification Battery tests. Each aptitude area represents a combination of abilities considered important for satisfactory performance in a number of Army jobs.

Aptitude test

A personnel instrument used to obtain an estimate of how well a person can learn to do certain kinds of work or acquire certain skills. This estimate is based on a measure of what the individual's present level is in respect to abilities or skills that have been found to be important in the work for which he is considered. Contrasted with ACHIEVEMENT TEST which measures how well an individual has already learned to perform a given task. The content of aptitude and achievement tests may be identical; the same test can, and frequently does, serve both purposes. An aptitude test helps to answer such questions as—Can the man be trained in a reasonable length of time to do the job?

Arithmetic mean
(See Mean.)

Army standard score
A standard score with a mean set at 100 and a standard deviation of 20. Raw scores on Army personnel measuring instruments are usually converted to Army standard scores which state the individual score in relation to the scores of the standard reference population.

Associate ratings
Evaluations of a man obtained from those who work with him or who are fairly closely associated with him. Such ratings have been found to be useful criteria against which to validate various methods of measuring performance.

Average
A number or value that represents all the values in a series; usually refers to the arithmetic mean of a series, but is actually a general term which covers median, mode, and other means.

Battery Of tests
(See Instrument battery.)

Bias
Error in measurement other than chance or random error. In sampling, if each case in the population has the same opportunity to be included, the sample, except for chance errors, will be representative of the population with which a test is to be tried out. However, some influence other than chance may cause a greater proportion of certain types of cases to be included than would result from chance selection; the resulting sample is said to be biased. Bias also affects the validity of ratings. (See Halo effect.)

Biographical information blank
Same as Self–report form.

Chance
Theoretical probability that an event will occur; variation among samples which causes statistical results obtained on one sample to differ from those on other samples selected on the same basis from the same population.

Checklist
A list of items to be noted or verified. In personnel measurement, a list of items which may serve either as a rating or basis for a rating; they are usually checked to indicate whether or not they apply to the ratee (or his behavior) and sometimes to what degree they apply.

Classification
The process by which personnel are evaluated in terms of what military tasks they are fitted to do or can learn to do with a view to their assignment or reassignment to jobs or training. In the Army, classification rests upon an analysis of the individual which takes into consideration his education, experience, interests, and physical status, as well as his aptitudes and abilities as estimated by personnel measuring instruments. Assignment is made in the light of existing manpower needs of the Army. Classification also pertains to the organization of military occupational specialties into related occupational groupings.

Conversion table
A table for changing scores from one kind to another. In the Army, usually a device for interpreting the raw scores earned on a personnel evaluation instrument by translating them into standard score equivalents. Table usually has two columns—a list of all possible raw scores and the standard score corresponding to each raw score. For a few instruments, raw scores are converted to percentile scores instead of, or in addition to, standard scores.

Converted score
Different personnel measuring instruments are likely to yield scores on different scales, with different scale units. A score obtained in terms of one kind of unit which has been translated into the units of another scale of measurement is known as a converted score.

Correction for chance success (guessing)
Many persons taking an objective test will answer all the questions whether they know the answers or not. A scoring,

formula is sometimes set up to deduct from their scores the estimated increase due to guessing. The formula subtracts from their score a certain proportion of the number they get wrong.

Correction for restriction in range
Data available for validating tests in the Army are often obtained not from the entire population of candidates or recruits, but from selected groups of men who have already been picked out for their estimated ability in the very tasks which the tests are assessing. As a result, the groups on which the tests are validated are bound to be closer together in the abilities concerned than the general mass of recruits from which they have been drawn. The result is usually a smaller validity coefficient than would have been obtained with the entire group. Correction gives an estimate of what the validity would be if the entire group were to be tested.

Correlation
Relationship or correspondence between two sets of measures. Positive correlation means that persons who stand high in one set also tend to stand high in the other set, and that persons low in one set tend to be low in the other. In, an inverse relationship, or negative correlation, persons high in one set of measures tend to be low in the other, and vice versa. Zero correlation means no relationship between two sets of measures.

Correlation coefficient
Numerical expression of the degree of relationship existing between two sets of measures, as, for example, the scores made on two different tests by the same group of people. The coefficient of correlation, abbreviated as "r," cannot be greater than 1 or –1 and is usually expressed as a two–place decimal fraction such as .86 or .08 or –.23. Positive values of r indicate degree of positive correlation; negative values indicate degree of inverse relationship. Values of r cannot be interpreted the same as percentages.

Criterion
The standard of human behavior that a personnel instrument is supposed to measure. In validation, it is the standard against which personnel instruments are correlated to indicate the accuracy with which they predict human performance in some area.

Critical score
(See Minimum qualifying score.)

Cross–validation
Repetition of a validation study, using data from a second group of men, for the purpose of seeing whether the validity previously found is maintained.

Cutting score
(See Minimum qualifying score.)

Descriptive rating
Verbal description of the ratee in the words of the rater. May be limited to designated characteristics or may be an overall estimate of value to an organization. Sometimes used with more objective rating instruments.

Differential classification
Designation of Army personnel for assignment on the basis of a battery of instruments for evaluating differences in aptitudes and abilities within the individual and between individuals. Implies assignment of personnel on basis of their more promising aptitudes, the final decision being also, in part, based upon the military personnel needs of the time and upon the education and experience of the individual.

Difficulty (difficulty index)
Applied to a personnel instrument or an item of a personnel instrument, difficulty or difficulty index refers to the percentage or proportion of examinees in a representative group who answer an item correctly. The smaller the percentage, the more difficult the item is considered. Difficulty of instruments or items is not a matter of a priori judgment but of actual trial, except when used as a rough estimate in selecting items for preliminary tryout.

Discriminative index
Validity of a test item in terms of the relationship between getting the item right and performance on job or training. If, for example, individuals who perform best on the job tend to answer a given item correctly more often than those who do not perform well, an item is considered valid.

Dispersion
The extent to which scores of a distribution spread out from the mean. Dispersion measures are in terms of distances on the scale of measurement and show extent of spread. They are used to supplement measures of central tendency, such as the mean, in describing a distribution of scores. The measure of dispersion most commonly used is the standard deviation.

Distortion
In self–report forms, distortion is the result of an individual's tendency, conscious or unconscious, to report his responses so that his score is unduly high or unduly low.

Distribution of scores
A tabulation, usually representing a group of scores, showing how many individuals made each score. Scores are arranged in order from highest to lowest. When there are a large number of different score values, the scores are grouped into brackets, or intervals. The frequency for each interval then indicates the number of cases with score values within the interval.

Empirically determined key
A scoring key based on the results of field trial and analysis; distinguished from predetermined key.

Equivalent form
(See Alternate forms.)

Essay question
A type of test question which calls for a complete answer in the form of discussion, comparison, explanation, or identification; an answer which must be thought out and expressed, in contrast to one. Which is only to be identified, as in, multiple–choice questions.

Essay rating
(See Descriptive rating.)

Essay test
A test composed of one or more essay questions; contrasted with objective test.

Expectancy table (chart)
A graph or table, based on a validity coefficient, showing the proportion of examinees earning each score on a predictor instrument or group of predictor instruments who may be expected to succeed in a given assignment. With a high validity coefficient a large proportion of candidates who make high scores will be expected to succeed on the job and a large proportion of candidate who make low scores will be expected to fail. With a low validity coefficient, proportions of probable successes and failures on the job tend to be the same for all scores.

External criterion
A criterion that is independent of the personnel measuring instrument designed to predict it. Contrasted with internal criterion commonly used in computing an internal consistency index.

Face validity index
A product of item analysis, the index measures the degree to which responses to an item are typically distorted.

Follow–up study
A means of evaluating the success of a personnel action based upon measuring instruments or procedures in terms of later outcome such as level of performance in training or on the job, career progress, or career status.

Forced–choice item
Usually a type of rating or self–report item in which the examinee or rater is required to choose between two or more alternatives which appear equally favorable or unfavorable, but only one of which is statistically related to the characteristic being measured.

Frequency distribution
(See Distribution.)

Graphic rating scale
The scale on which a rating is to be made may be shown as a scaled line (or other diagram showing various steps from

high to low) on which the rater records his judgment. Verbal descriptions are usually placed at each scale point to indicate the level or quality of performance which a man should show in order to be placed at that point.

Group test
A test which can be administered to more than one examinee at the same time, using the same test directions and duplicate testing materials. Most group tests are paper–and–pencil tests—contrasted to individual tests.

Halo effect
A tendency in raters to let their general impression about a person influence their judgment concerning specific and independent traits of that person. In operation, halo may result in a person's being rated high on all or most traits or low on all or most traits it is one of the chief obstacles to obtaining valid ratings.

Individual differences
The way in which persons vary from one another in the pattern of traits into which their behavior is organized. Individual differences is a matter of the varying degree in which different persons manifest traits common to all or almost all persons, rather than of the presence or lack of certain traits. The term may also refer to variations in level from trait to trait in any one person.

Individual test
A test that can properly be given to only one person at a time, usually by a single examiner. Administration usually requires a highly trained examiner.

Instrument
Any means by which differences among individuals can be measured or the relative standing of the individual in the group determined. Tests, rating forms, inventories, and standard interviews are all personnel measuring instruments.

Instrument battery
A group of tests administered for a common purpose. The scores may be used to present a profile for an individual or combined into a single score or rating, with each score weighted according to its contribution to prediction of the criterion, usually success on the job. Generally, a battery can predict performance on the job more accurately than any one of the instruments used alone.

Intercorrelations
Correlation coefficients existing among two or more sets of measures. Usually refers to coefficients among tests or among items.

Interest blank or inventory
(See Preference blank.)

Internal consistency
Degree to which an item ranks men according to their total score on an instrument. If, for example, the individuals who answer an item correctly tend to answer most other items correctly, while those who answer an item incorrectly receive low total scores, then the item contributes to measuring whatever is measured by the test as a whole, and is said to have high internal consistency. An item answered correctly as often by those who make low total scores as by those who make high total scores does not so contribute and hence is usually taken out of the instrument. The index may take one of several possible statistical forms, but is usually a correlation coefficient.

Interview
Conversation between interviewer and interviewee directed along channels predetermined by the purpose of the interview. Purpose may be to give or to get information or directly or indirectly to help in the solution of vocational or emotional problems of the interviewee. In personnel measurement, the purpose of the interview is the evaluation of some aspect or aspects of the interviewee's behavior. (See also Measurement interview.)

Inventory
(See Self–report form.)

Item (test item)
A test question or problem. Any single element of a test to which response is desired. Questions, problems, or tasks comprising an evaluation instrument.

Item analysis
The statistical study of each item of an instrument to find out the extent to which the item contributes to the effectiveness of the test as a whole. It usually includes finding the difficulty, internal consistency, and validity of each item.

Item analysis key
(See Empirically–determined key.)

Item selection
Using item analysis data, the process of selecting from a pool of items those items of specified difficulty and validity to make up the final form of the instrument.

Job
Term used throughout this manual to denote area of work or employment. In the Army often used to mean MILITARY OCCUPATIONAL SPECIALTY or duty assignment within an MOS.

Job analysis
The process of collecting, evaluating, and presenting detailed information about jobs, including the duties performed, and the knowledges, abilities, skills, and personal qualities required for doing each job satisfactorily. Job analysis is basic to the construction of job proficiency tests. It is helpful in preparing instruments for classification or selection for Army job or assignment.

Job sample test
An achievement test of performance on an actual job. The job sample test usually consists of a single operation or a sequence of related operations in which the examinee is required to employ the usual materials, tools, and techniques of the job. Sometimes used synonymously with WORK SAMPLE test.

Mean
Arithmetic mean. In popular usage is called "average." Computed by adding the value of all scores or measures in a series and dividing the sum by the number of scores or measures in that series.

Measurement interview
An interview conducted according to a uniform procedure so that all the persons interviewed go through, as nearly as possible, the same process. May be conducted by a single interviewer or by a board, members of which observe the interviewee on carefully defined aspects of his behavior and rate him on the basis of their observations. Generally used as one of several instruments in a selection or classification battery. Results in a score which can be used in combination with scores on other instruments.

Military occupational specialty (MOS)
The term used to identify an area of military job activities which require similar skills and abilities for their performance. A military occupational specialty includes duty positions which are sufficiently related with respect to duties and responsibilities, skills and knowledges, and physical and mental requirements to be adequately described in a single job description.

Minimum qualifying score
A score below which candidates are not accepted for assignment. Location of the minimum qualifying score on the scale of measurement depends upon the selection ratio (number of candidates needed for a particular assignment divided by the total number of candidates) and the magnitude of the validity coefficient.

Multiple–choice question
A form of objective test question in which two or more answers, or alternatives, are presented. The examinee is instructed to choose the answer he thinks is right and indicate it in some prescribed way, usually by marking it on a separate answer sheet.

Multiple correlation
A technique for determining how to combine tests to get the best possible prediction of a criterion and for estimating what statistical relationship the composite score will have with that criterion.

Nomination technique
A type of rating in which raters are asked to indicate from among an entire group the persons they consider best and poorest in specified characteristics.

Nonlanguage test
A test that requires little or no speaking, reading, or understanding of language on the part of the examinee either in connection with comprehending directions or making responses. Directions may be given pictorially or in pantomime. Used with illiterates or persons unfamiliar with the language in which tests are given.

Nonverbal test
(See Nonlanguage test.)

Normal distribution
A distribution of scores or measures such that it normal curve is the best fitting curve. Most measurements of personal traits or characteristics are found to be distributed normally or approximately normally when data are taken for a large and unselected group of individuals.

Normal probability curve
A frequency curve based on the laws of chance; the form of curve commonly found when a very large number of values are tabulated. Appears as a symmetrical, bell–shaped figure rising above the base–line measurement scale, reaching a maximum height at the center, and tapering to the base–line at both extremes of the scale.

Normalized standard scores
Standard scores converted so that their resulting distribution is normal. Conversion is usually from raw scores to percentile scores which are then transformed into standard scores.

Norms
Norms describe levels of performance reached by specified groups and provide a means of comparing performance of individuals. Norms for Army personnel instruments are usually expressed as Army standard scores and are based on the performance of a standardization sample representing as nearly as possible the population with which the instrument is to be used. (See Army standard score.)

Objective test
An instrument in which there is little or no disagreement among experts as to the correctness of response and on which the result obtained is almost completely independent of the person doing the testing and scoring, so long as directions are strictly followed.

Occupation
(See Military occupational specialty.)

P value
Percentage of the group to which a personnel measuring instrument is administered marking a given item or item alternative.

Paper–and–pencil test
A personnel measuring instrument, most often verbal, on which the examinee responds to questions by writing or checking his answers, usually on a separate answer sheet. Usually administered to groups, but may be administered to individuals.

Percentile
The score in a distribution of raw scores below which occur a given percentage of the cases. Hence, the 70th percentile is the raw score below which 70 percent of scores of persons in the group fall.

Percentile rank
Same as Percentile score.

Percentile score
Indicates the percent of the group which ranks below the specified percentile rank. Thus, an individual who has a percentile score of 65 exceeds 65 percent of the group and is exceeded by 35 percent of the group.

Performance rating
(See Administrative rating.)

Performance test
A test in which the examinee is required to demonstrate some practical application of knowledge or some degree of an essential skill. Frequently a work sample test, but it may also be in paper–and–pencil test form. More likely to be given individually and to require nonverbal responses. Performance tests are of special value with persons having limited language ability, since the verbal directions may be simple and the response is likely to be a non–language response. Performance tests are also of value where the behavior to be measured involves interactions within complex situations.

Personal inventory
(See Self–report form.)

Personality questionnaire
(See Self–report form.)

Personnel classification
Process of evaluating and continuously re–evaluating the individual's mental and physical abilities, interests, education, aptitudes, physical assignment limitations, occupational history, and military experience, in order that he may be assigned to duty positions which utilize his qualifications to the maximum extent consistent with the needs of the service.

Personnel selection research
The development of psychological and psychometric methods through which the best candidates for successful training or job assignments are identified from a large applicant pool. The appropriateness of a selection approach to a personnel problem is contingent upon such factors as the number of personnel needed for training or assignment, the number of personnel potentially available, the quality of personnel required for the job, and the importance of the job.

Personnel utilization research
Development of human factors knowledge and techniques aimed at improving individual and group personnel performance on the job, taking into account needed balance between man and machine capabilities, psychological and behavioral limits of working demands, and factors of work environment including unusual as well as typical conditions of the job.

Population
All of the cases in the group with which a research study is concerned. In the case of a selection instrument, the population is the total of those available for selection. The term population is sometimes applied, less precisely, to the sample of cases which is taken to be representative of the total group and on which analysis is carried out.

Power test
Test in which items are usually arranged in order of increasing difficulty and in which examinees are given all the time they need to complete as many items as they possibly can. Usually contrasted with speed test. Army tests usually measure both power and speed in that time allowed is ample for most examinees but not unlimited.

Practice effect
When a person takes the same test more than once, his scores on the later trials may be higher because of his familiarity with test procedure or content. To reduce this effect, the Army provides alternate forms of measuring instruments for use when retest is necessary.

Predetermined key
A scoring key in which answers are scored or weighted on a judgment of how the items are expected to predict performance or behavior, rather than on how they do predict when actually tried out. Distinguished from empirically determined key.

Prediction
Estimating criterion performance, such as success on the job, from known performance on another measure or set of measures. The degree of accuracy of prediction can be estimated from the size of the validity coefficient.

Preference blank
A type of self–report form designed to appraise systematically the expressed preferences or interests of individuals usually for specified occupational activities.

Probability
(See Chance.)

Proficiency test
(See Achievement test.)

Profile (profile chart)
A graph or diagram which shows a person's relative place in a group on each of several different traits, throwing into relief the pattern of the qualities in which he excels, is average, or in which he is relatively deficient. Such profile patterns may be considered in connection with the known requirements for success on various jobs in deciding upon a suitable assignment for an individual.

Random sample
A sample chosen from a population so that every individual case has an equal and independent chance of being included. Does not mean a sample chosen haphazardly without a selection plan. In practice, such pure randomness is seldom approached.

Range
The limits within which the actual scores or values on a test or other measurement device fall for a given group. Sometimes expressed as the difference between the lowest and the highest obtained scores, but more often expressed merely in terms of these values, as in the following: "Scores range from 10 to 85."

Rank (rank order)
In personnel evaluation, the relative standing of an individual on a given trait, with reference to other members of the group. When all members of a group of ten are arranged in order from lowest to highest, the number 1 may be assigned to the one who stands highest, 10 to the lowest.

Rating
An evaluation of an individual either as to overall value or competence or in regard to the degree to which he shows some particular ability or trait; may be in comparison with others in the group or against a fixed standard.

Rating form
An instrument on which the rater records his evaluations. Usually combines several rating techniques by means of which an overall estimation of the ratee's value is reached.

Rating scale
A device for recording a rater's estimates of an individual, either in respect to specified traits or abilities, or as to his overall value to an organization. The various levels or degrees which make up the scale may be defined in terms of concrete behavior to be observed. May be accompanied by an actual scale of measurement, as in a graphic rating scale.

Raw score
The score as originally obtained, expressed in the original units of measurement. The total number of right answers, or sometimes the total number of right answers minus a fraction of the number wrong.

Regression
In correlation, the tendency of a predicted value to be nearer the average than to the value from which the prediction is made.

Reliability
The degree to which an instrument can be relied upon to yield the same result upon repeated administrations to the same individual. It differs from validity in that it is concerned only with the dependability or consistency of the measure and not with whether or not the instrument measures what it is supposed to measure.

Reliability coefficient
An estimate of the consistency of results to be expected from a personnel measuring instrument. Usually a correlation

coefficient for two sets of test scores obtained by administering a test twice to the same persons or by administering two equivalent forms of a test to the same group. (See Reliability.)

Sample
Generally refers to a group of individuals taken as representing a larger group; for example, an instrument is standardized on a sample group representative of the operational population with which it will be used. The terms "sample" and "population" are, however, sometimes loosely used interchangeably. "Sample" may also refer to selected items of knowledge or behavior taken as typical of a whole body of knowledge or of many aspects of behavior; for example, a test may "sample" job content.

Sampling
General process of selecting a limited number of cases from a population.

Scatter diagram
A two–dimensional chart showing the relationship between two measures represented by horizontal and vertical scales. Each point on the diagram represents two measures, and the resulting configuration of points shows the extent and nature of the relationship existing between the two measures.

Score (test score)
A general term covering raw scores and converted scores, such as standard scores and percentile ranks.

Scoring formula
Part of the scoring directions which states how the final figure expressing the score is to be arrived at. It may clarify such questions as whether a portion of the number wrong is to be subtracted to correct for possible success in guessing.

Scoring key
The specific pattern of answers to be counted on a measuring instrument to obtain a score.

Screening
Gross selection early in the total selection process to identify those in the available supply of personnel who meet the minimum qualifications for a given assignment. Selection in which those below the minimum qualifying score on a preliminary evaluating instrument are rejected from further consideration for a particular assignment. Usually takes place early in the selection procedure to avoid subsequent unprofitable testing.

Selection
The process of choosing a required number of individuals who have the necessary and desirable qualifications for entering upon a certain job or training when the number meeting the minimum qualifications is in excess of the number required. Usually, a broad set of qualities or traits are assessed, and a variety of personnel instruments are employed to do the selecting.

Selection ratio
The number of applicants needed for a given job divided by the total number of applicants available. Along with the validity coefficient, the selection ratio helps to determine a minimum qualifying score on the selection instrument used in selecting personnel for a given assignment. For example, where the selection ratio is low, that is, when only a small number of applicants are to be selected from many, even a low validity coefficient will help in the selection of candidates who are more likely than not to be successful in the assignment.

Self–report form
A technique whereby information is furnished by the individual concerning his background, attitudes, beliefs, and personality reactions. The scoring key may be empirically determined. Score may be useful in predicting his on–the–job success. Usually is in the form of a checklist, inventory, or questionnaire.

Significance (statistical significance)
Statistical significance refers to the numerical probability that variations in scores and other measurements are the result of chance or random errors alone. A variation that is considered "significant" is one which has occurred as a result of factors or circumstances in addition to those of chance.

Speed test
An instrument in which the time limit is set so that almost no one can finish all the items or tasks making up the test. Opposite of power test.

Spread
(See Dispersion.)

Standard deviation
A measure of the spread of scores or how much members of the total group vary among themselves. Computed by squaring the deviations from the mean, averaging them, and then taking the square root of the result.

Standard score
A score expressed in terms of the number of standard deviation units by which it differs from the group average. It enables any individual's score to be compared to the performance of the whole group and thus be given meaning. (See Army standard score.)

Standardization
The administration of a test or other personnel evaluation instrument to a sample representative of the population with which it is to be used to determine the proportion of the group that will reach or exceed each test score. It is then possible to compare scores made by individuals on subsequent administrations of the test with the performance of the standardization population.

Subject–matter specialist (subject–matter expert)
An individual expert in some occupational or job area in which tests are constructed. He may act in an advisory capacity regarding achievement test content or may be trained to construct the actual test items.

Suppressor key
A key used in scoring self–report forms to adjust an individual's score for any constant distortion tendency.

Test battery
(See Instrument battery.)

Test item
(See Item.)

Test plan
An organized outline of the knowledges, skills, abilities and attitudes needed to perform a particular job successfully. Each part is weighted in proportion to its importance to the job. It is from such a content outline that an achievement test should be constructed.

Test score
(See Score.)

Utilization research
(See personnel utilization research.)

Validation
The process of trying out a personnel measuring instrument to determine its effectiveness in predicting performance on an assignment. This effectiveness is usually expressed in terms of the correlation coefficient between scores on the instrument and scores on a criterion of proficiency in job or training.

Validity
In general, the extent to which a measuring instrument really measures the skill, area of knowledge, aptitude or other characteristic it is supposed to measure. The extent to which a personnel evaluation instrument is able to select in advance persons who will do well in actual assignments, and to detect those who are likely to do poorly. A test valid for one purpose may not be valid for another purpose. Validity, hence, is not an intrinsic quality of a measuring instrument but is relative to the purpose of the instrument.

Validity coefficient
A statistic that shows the degree of relationship between scores on an instrument and performance in a particular job or training program. It is usually a coefficient of correlation between a test and a criterion the test is intended to predict. Indicates the extent to which individuals who score high on the test also score high on the criterion and those who score low on the test score low on the criterion.

Variability
(See Dispersion.)

Variable
Any measure which can take on different or graduated numerical values, such as age or scores on an evaluation instrument.

Verbal test
Theoretically any test in which language is involved. In general usage the term is restricted to those tests in which the questions and responses are mainly expressed in language or which use language to a substantial degree.

Weighting
The process of determining, either by judgment or by statistical means, the relative importance each test in a battery or each item in a test should carry in the overall or composite score.

Work sample
A small problem representative of the job as a whole, chosen and adapted for the purpose of testing performance on important operations of the job as nearly under normal conditions as possible apart from an actual tryout on the job. Performance on a work sample is frequently used as a criterion against which prediction devices in personnel evaluation are validated.

Section III
Special Abbreviations and Terms
This section contents no entries.