4-1-1949

Testing as a tool in the effective selection and placement of employees

Edward Sprunt Hamilton

TESTING AS A TOOL IN THE EFFECTIVE SELECTION AND PLACEMENT OF EMPLOYEES

BY

EDWARD SPRUNT HAMILTON

A THESIS SUBMITTED TO THE GRADUATE FACULTY
OF THE UNIVERSITY OF RICHMOND
IN CANDIDACY
FOR THE DEGREE OF
MASTER OF SCIENCE IN BUSINESS ADMINISTRATION

JUNE, 1949

Approved:
J. Byrd Mills

PREFACE

The day has passed when a company is willing to assign to a job any person that comes into its employment offices. With the labor market easing slightly, competition increasing, and production costs at an extremely high level, there is a growing pressure on those responsible for employing labor to improve their selection methods. To such persons, testing represents a necessary addition to present employment practices.

The most likely users of tests, personnel managers, have long been aware of the value of tests and of the large quantity of written information concerning tests. But the essential fundamentals of testing - its background, its classifications, and its technical usage - have never been brought together in simplified form in a way that such users could have convenient reference to it.

For the most part, information about tests is usually found in single chapters of books dealing with personnel management or in studies reported in periodicals. In some cases, entire texts can be found which deal with testing or with special phases of testing. However, these latter books are usually written by psychologists and are written on a professional level. Personnel managers need information of a more practical sort and have not the time to study many texts in order to absorb enough background about tests to understand them. There is a need for written material in the vocabulary of the personnel managers, short enough for them to study.

What is the background of tests and how have they developed? What is the nature of the various factors tested, such as intelligence and personality? What is the relationship of these factors to job success? And, how can test results be interpreted?

These general ideas led to the development of this thesis.

Edward Sprunt Hamilton

TABLE OF CONTENTS

CHAPTER I

GENERAL BACKGROUND

A. Introduction

All humans have measurable differences in physical characteristics, abilities, attitudes, intelligence, personality, and interests. Because of this two individuals cannot be expected to perform a job in the same way.

In a routine check-listing operation on adding machines in one bank an analysis of the records showed a variation in the number of checks listed per error from 758 to 2,788. The best operator listed 1,578 checks per hour while the poorest listed only 1,076. When these operators were employed, without the benefit of testing, they were considered as equally good risks.

In another department where employees count currency by hand, the volume per day ranges from 14,000 to 21,000. In other words, the best currency counter counts 50 percent more than the poorest. Here too, from the standpoint of the bank's existing employment practices, there was no way to predict this distinction.

Obviously, if personnel testing can help a company select more employees who fall near the high end of such records and fewer who fall at the low end it will be worth while.

The application of testing to the solution of such a problem is not as easy as it sounds. It must be determined what aspects of intelligence, motor skills, visual skills, personality, and interest are present in the superior adding machine operators and currency counters. And it must be determined how well tests can actually identify the superior employees.

Not only does attention have to be given to basic differences such as intelligence and skill but to their finer components. One person might show extreme ability in making rapid arithmetical calculations and little ability in recalling recently learned material. Another might show speed and endurance in major arm movements and yet lack finger dexterity.

Among applicants might be found one with a cooperative nature and a tendency for calmness and another who was quarrelsome and easily irritated.

As for interests, refer again to the adding machine operators. What particular interests has the top operator which distinguish her from the low one? Do the upper half of the operators have a distinctive interest profile? And if tests can reveal this profile, can't the ratio of selecting the better operators be improved?

The importance of placing each individual in a job at which he can best demonstrate his abilities and for which he is best suited from a personality standpoint is receiving more attention from industries because it can mean lower costs and fewer supervisory problems. Despite the fact that industry has been previously aware of certain differences in training

and skill as far as individuals are concerned, it cannot be said that industry as a whole is ready to accept the help of testing in meeting this problem as subsequent data will indicate.

Industries which rely on the interview as the most common means of selecting and placing employees have recently begun to realize that few of an individual's basic capacities and interests can actually be detected by the interviewing process.

Typical Methods of Judging Men

A common technique of choosing and placing men is the use of what is called "judgment." This includes the use of an individual's opinion from observations in determining the existence of particular abilities in a person. Responsible supervisors remark "I can usually tell by looking at a girl whether she will make a good adding machine operator."

In addition to applying "judgment" men have fallen back on any number of so-called sciences to aid them in selecting others, including:

> "Astrology, chance as manifested in the drawing of straws, casting of lots, or the flipping of a coin, chiromacy, character analysis, divination, fortune telling, graphology, horoscopes, intuition, magic, mediums, mind reading, omens, occultism, oracles, phrenology, physiognomy, premonitions, sorcery, racial superiority, and sub-conscious hunches."[1]

Several experiments have shown the unsoundness of judgment as a basis for selection of employees. Cook mentions one of these experiments in referring to the study by Stuart Rice in 1929 in which 2,000 homeless

---

[1]Scott, W. D., Clothier, R. C., Mathewson, S. B., Spriegel, W. R., Personnel Management, McGraw-Hill Book Co., New York, 1941, pp. 147-149.

men were being interviewed.[2] In this experiment one of the interviewers was a Prohibitionist and another was a Socialist, although both were well trained and conscientious. The personal bias of each man obviously entered into his formation of opinions during the interviews because the Prohibitionist found that 62 percent of the men attributed their downfall to liquor and the Socialist found that only 22 percent did. Conversely, the Socialist found that 39 percent were in need because of the existing industrial situation while the Prohibitionist found that this reason was applicable to only 7 percent. Apparently, people believe about others whatever they want to believe.

Another experiment mentioned by Scott shows the glaring differences of opinion between men who consider themselves good judges of men.

> "Arrangements were made for 13 industrial executives of major rank, each of whom prided himself on his ability to choose men, from as many different companies, to meet and select the best salesmen from a group of 12 applicants. In doing so each was directed to interview each of the 12 applicants privately, use whatever procedure or method he wished, then rank them from 1 to 12 in the order of his preference.
>
> "Glaring differences of opinion preclude the possibility of all these gentlemen being good judges of men! Messrs. A and B, for instance, pronounced applicant IV the best man of the group, whereas Mr. G ranked him 11. Mr. K chose applicant V as the best man, but Mr. H ranked him 10 and Mr. B ranked him 12. Applicant VIII was ranked 1 by Mr. E, 10 by Mr. G, and 12 by Mr. F."[3]

It is not too astounding that such differences result when we recall that humans in general have an amazing range of prejudices, such

---

[2]Cook, David W., "Psychology Challenges Industry," Personnel Series No. 107, American Management Association.

[3]Scott and others, op. cit., pp. 145-147.

as that all redheads lose their tempers easily or that individuals raised in poverty are industrious.

In another experiment of the same type similar instances were found:

> "Six district managers of one company were called together and instructed to select the best men from 36 individuals and to rank them in order of superiority. The managers' standards were identical; their industrial environment was also the same.
>
> "Excellent agreement was reached on applicant IV but applicant I on the other hand was ranked 1.5 (tied for first place) to 11th; applicant V was ranked from 3rd to 28th place; applicant XIII from 3rd to 23rd; etc."[4]

Other experiments of this type would show that relying on the judgment of several persons results in very little agreement between them as to which person is the best applicant.

An industry is primarily interested in the presence of the differences between individuals rather than the basic causes of them. It should be noted, however, that both heredity and environment contribute separately to the development of these differences. Furthermore, since environment continues to mold what has been left by heredity and brings about continuing changes in these differences, the differences do not necessarily remain constant in an individual.

"The concept of individual differences is concerned with basic differences in capacity which are of importance in every phase of industrial placement."[5] These basic differences determine productivity, promotability,

---

[4]Ibid., p. 147.
[5]Tiffin, Joseph, Industrial Psychology, Prentice-Hall, Inc., 1946, p. 1

versatility, and accuracy. One of this paper's contentions is that the use of psychological tests provides a more accurate method of predicting the existence or absence of different traits among individuals.

In the title of this paper attention has been focused on certain limitations of testing by the phrase "testing as a tool." It is apparent that testing alone cannot supply an adequate answer to the question of whether an individual is well suited for a job. Satisfactory selection and placement require that testing be combined with other techniques of selection such as interviews, application blanks, and experience. Tests should serve more as an addition to the regular employment practices and possibly as a point of departure for certain aspects of the ordinary employment interview. Tests, along with experience, merit ratings, and interviews, will merely improve selection and placement.

It is desirable to limit what is meant by the purpose of tests in this paper to the following general statement: "Discovering the extent to which an applicant possesses the abilities and qualities which organizational needs are to impose upon him if he is placed in a particular work situation."[6]

In a recent speech before a group of personnel administrators Dr. David W. Cook pointed out that industries are "reluctant to use improved

---

[6]National Industrial Conference Board, Inc., New York, "Experience With Employment Tests," Studies in Personnel Policy, No. 32, 1941, p. 5.

methods of selecting and placing employees,"[7] although reports show that there is a slight increase in the number of industries now beginning to set up testing programs. In 1940 one writer listed only 8 companies which, to his knowledge, were making systematic and extensive use of tests in the selection of employees.[8] In an actual survey in 1935, the National Industrial Conference Board reported that of 2,452 companies employing over 4½ million persons only 7.3 percent of the companies were using tests. Four years later the Conference Board polled 2,700 firms employing nearly 5 million persons and found that at least 14 percent of the firms had testing programs.[9]

Apparently a good many personnel managers feel that testing can be used successfully only by firms employing a large number of workers. The relationship between the size of the firms and the use of tests is indicated in the following table. The question asked in this instance was "Are employment or placement tests used as a regular feature of your employment program?"[10]

---

[7]Cook, David W., "Psychology Challenges Industry," Personnel Series No. 107, American Management Association.

[8]Jucius, Michael J., Personnel Management, Richard D. Irwin, Inc., Chicago, 1947, p. 201.

[9]Lishan, John M., "The Use of Tests in American Industry: A Survey," Personnel, January, 1948, p. 305.

[10]Lishan, op. cit., p. 306.

	NO TESTS		UNQUALIFIED YES		QUALIFIED YES	
	Number	Percentage of "no" replies	Number	Percentage of unqualified "yes" replies	Number	Percentage of qualified "yes" replies
Less than 250 . .	45	14	0	0	2	5
250-749	88	27	5	11	7	16
750-1,499	58	18	10	22	6	13
1,500-2,999 . . .	54	17	8	17	10	22
3,000-7,999 . . .	50	16	11	24	7	15
8,000-14,999 . .	14	5	6	13	5	11
15,000-29,999 . .	4	1	2	4	4	9
30,000 and over .	7	2	4	9	4	9

Inasmuch as studies referred to by Mr. Lishan were not made on a valid statistical basis to obtain a sample representative of American industry, it is impossible to draw any conclusions applicable to industry as a whole and as he points out, it can hardly be said that a trend of any kind is discernible.

As a matter of fact, the above information is the least important point which should be considered in discussing the question of testing as an effective tool. The decision should not be based on how many firms do or do not employ tests but rather on whether the particular test in question will give a valid measurement of the qualities being tested and whether those qualities are needed in a particular employee situation.

In the field of clerical employment there has been a more recent

survey conducted by the Office Management Association of Chicago.[11] In this survey 59 companies stated that they now have a testing program; 60 do not have and never had a testing program; and 5 firms have abandoned testing. Most frequently mentioned positions for which tests are given are: stenographic, 47; typing, 45; clerical, 43; machine operative, 28; filing, 23; messenger, 16. Seven of the reporting firms gave tests for all positions.

It is probably safe to say that attempts to increase unit productivity through the use of psychological tests in the selection of employees have not kept pace with the increasing significance of unit productivity in our present day economy. Personnel managers still seem prone to overlook the usefulness of tests to augment a personnel program although a good testing program can accomplish the following things:

A. Measure the extent to which an applicant's abilities and disabilities fit him for job demands.

B. Check on reported experience.

C. Objectively compare applicants.

D. Aid in setting up standards for employment procedure in such a way that prediction can be traced, checked, and progressively improved.[12]

There seems to be adequate evidence that tests are helpful in employment, placement, and in training. In the employment phase, tests

---

[11] The Management Review, American Management Association, New York, September 1948, p. 458.

[12] Pigors, Paul, and Myers, Charles A., Personnel Administration, A Point of View and A Method, McGraw Hill Book Co., Inc., New York, 1947, p. 141.

are particularly useful in the screening of applicants. The following table[13] shows a typical result of testing:

Number	Class of Employee	Wages % of Average	Average Length of Service (Months)
13	Old employees, passed test	116.1	32
10	Old employees, failed test	93.6	24
16	New employees, passed test (not trained)	97.4	2
9	New employees, passed test (trained)	113.3	1/2

It can be seen that within two months the employees who were selected by tests but who had not been trained by the company had earned more than older employees who failed the tests but had been with the company an average of 24 months. Those who had been trained and tested had earnings in 1/2 month close to those who had been with the company an average of 32 months.

Sometimes even an extremely simple test can be helpful in the screening phase of employment. For example, a number checking test given to 12 applicants for jobs with duties involving the listing of checks on an adding machine produced scores ranging from 87 to 154. The test scores were not disclosed to the manager who supervised the training and work of all 12 girls. Of those scoring below 110, 3 were released on the recommendation of the manager for unsatisfactory work and a fourth is about to be

[13]Drake, C. A., and Oleen, H. D., "The Technique of Testing," Factory Management and Maintenance, March 1938, p. 78.

released. Those with scores over 130 have been commended for their progress. Even on such a small scale it would appear advisable to set a score from 110 to 130 as the minimum for individuals going into such work and in this way screen out most of those with a limited chance of success.

B. History

Perhaps the earliest instance of the use of tests in measuring individual differences was carried out by Professor J. M. Cattell at Columbia University in 1896 when tests were given to 100 freshmen. From that point until World War I the development of testing can best be described as experimental with extensive contributions from the Civil Service Commission. This Commission began to give examinations in terms of school subjects during the period prior to 1900 to indicate degrees of merit on which basis the Commission was empowered to select employees. As objections grew to these academic tests on the grounds that they were not related to the jobs to be filled, the questions on the tests were rewritten so that they had reference to a particular job. Practically no other types of tests appeared before the development of psychological test methods during World War I.

There were a few requests from industries to colleges for help in solving selection problems dealing with students, switchboard operators, salesmen, bank employees, and others which served as a stimulus to the various experimental programs. One of the first experiments carried out in answer to one of these requests was performed by Professor Munsterberg

of Harvard in developing a psychological test for selecting streetcar motormen.[14]

As these experiments began to show successful results, interest and activity in this direction increased and business men as well as psychologists became struck with its potentialities. As Hunt stated, this was the starting point "for the development of a number of similar tests and for the beginning of a wide spread cooperation between the science of psychology and industry in the solution of industrial problems."[15]

From this type of test there was very little change until the impetus to psychological methods given by the first World War. The development of these types of measurement for recruits represented the largest scale application of psychological methods to a selection problem that had ever been undertaken.

The psychologists had been asked by the army to prepare mental alertness tests which would point out those recruits capable of learning new skills quickly and to weed out those recruits whose intelligence was

---

[14]Hunt, Thelma, Measurement in Psychology, Prentice Hall, Inc., 1936, p. 265. Hunt describes Munsterberg's early experiments in testing motormen for streetcars. The test consisted of setting up several cards with lines and numerals on them representing a street with the various conditions of pedestrians and vehicles. The numbers represented the number of steps required to bring the object into the path of the streetcar and the color of the numbers indicated whether the objects were moving parallel to the car or at right angles. The cards were exposed on a belt turned by a crank and the subject had to pick out those objects which were potentially dangerous because of proximity and speed and direction. The score was measured by the number of errors and the time.

[15]Ibid., p. 265.

so low they would be dangerous to the army, and to set up methods to select candidates for commissions.

In addition to these tests, psychologists were called on to conduct investigations in the development of specialized tests for aviation and for other more technical occupations. Most of these investigations later served as starting points for testing developments in the field of selection in various occupations after the war.

Following the war, the Civil Service Commission established a Research Division in 1922 which began to work on a series of mental alertness tests for use in selecting public personnel. Later, as it began to be clear that the relationship between abstract intelligence and success on the job was not high enough to rely on that factor alone, adaptations were made in the general intelligence test by the use of specialized terminology. These adaptations brought about a test with a better selective value. An outstanding example of this stage of testing history is the test developed by Telford and Moss for the selection of city policemen:

> "(1) Observation - Show applicant picture of collision between streetcar and an automobile, require him to study it for a limited time, ask him several questions about it without letting him refer to the picture.
> (2) Memory - Show the applicant a large number of photographs with names under them and later require him to identify the unnamed photographs.
> (3) Comprehension - Require applicant to answer questions based on printed selections from laws, ordinances, and police regulations.
> (4) Judgment - Give applicant multiple choice questions dealing with situations requiring judgment. For example, 'If a policeman considers himself unfairly treated by his sergeant and gets no satisfaction

when he explains to the sergeant that he is not treated fairly, he should:

____Refuse to obey any orders given by the sergeant.
____Invite the sergeant to meet him when both are off duty so they can settle the matter themselves.
____At the first opportunity report the matter to his lieutenant and captain.
____Immediately hand in his resignation.' "[16]

From this phase of specialized tests it was a logical step to the next phase in the trend of testing: achievement or trade tests in objective form. These tests were designed for testing applicants for positions in which high intelligence and special aptitude were not enough. Principally, they were short answer tests for selecting workers for such jobs as chemists, bacteriologists, hospital workers, etc. The Research Division of Civil Service continued its efforts along these lines.

The use of tests began to increase after World War I and it is probable that this expansion was one of the reasons why testing fell into bad repute in later years. It has been said that many types of people with little or no experience in testing or in the interpretation of test results made use of the tests and expected too much from them. As a consequence, many firms wrote off testing as another useless fad.

As pointed out by Scott and others, from 1919 to 1933, the scientific testing program went through a period of consolidation.[17]

---

[16]Telford, Fred, and Moss, F. A., "Suggested Tests for Patrolmen," Public Personnel Studies, 1924, Volume 2, p. 112.

[17]Scott and others, op. cit., p. 152.

Those firms that "toyed" with tests after 1919 discontinued them by 1933. In 1931, 1932, and 1933 studies showed tests were used in only 4 to 8 percent of the organizations studied.[18]

[18] Ibid, p. 153

C. Limitations and Classification of Tests

As industry and psychologists began to devote more attention to studying the human element, it was an admission that machines could be used best by properly selected men - an admission that subsequently became a by-word of advanced personnel administration.

In defining a test as "a process of measurement by which it is hoped to determine how well a person has done on something or may do on something in the future,"[19] Jucius calls attention to several other pertinent aspects of tests: first, that they are merely samples; second, that they are measures of past efforts or predictors of future events; and third, that decisions should not be based on the results of tests alone.

Psychological methods are not infallible due to the very variability of the subject tested - humans. Because of this caution must be used in relying on tests for effective selection. Generally speaking, the usefulness of a test will depend on two things: whether the test measures the particular thing that it is designed to measure and whether it measures it accurately and consistently. In psychological terms, these elements are called validity and reliability and will be discussed more fully in a later section of the paper.

Another factor to be considered in the use of tests is the percentage of the present employees who are satisfactory. If, for example, one-half of the present employees are satisfactory, any increase which can be

---

[19]Jucius, op. cit., p. 201.

achieved in the percentage of satisfactory employees by the use of the test will result in improvement. And the amount by which the percentage of satisfactory employees hired is above 50 percent will be indicative of the functional value of the test.[20]

Classification of Tests

One danger in attempting to classify tests is that a test may be thought of as falling within a particular pigeon hole and as having little relationship to tests in the next classification. Another hazard in studying types of tests is to assume that the title is exclusively descriptive. It should be remembered that "the label attached to a test gives no guarantee of its adequacy as a selection tool for any particular type worker."[21]

Examples of this last statement are the McQuarrie Mechanical Aptitude Test, parts of which have been found effective in selecting office workers, and the Minnesota Test for Clerical Workers which has some value in selecting many types of factory workers.

Nevertheless, it is easier to study the use of tests by setting up some sort of classification framework. Tests may be classified according to the manner in which they are given: group or individual; instrumental or paper and pencil. Or they can be classed on the basis of what they attempt to measure, for example: "aptitude, capacity, or latent

---

20 Discussed more fully in Tiffin, op. cit., pp. 37-40
21 National Industrial Conference Board, Inc., op. cit., p. 6.

ability to learn if given training; achievement, how well one can do a job or what he knows about it."[22]

There are almost as many classifications of tests as there are books on the subject of testing. Cleeton and Mason divide tests into general ability, special ability, trade, interest, and personality.[23] Pigors and Myers set up a classification of performance, aptitude (includes general level of intelligence), and temperament or personality.[24] Beaumont makes a more ambitious attempt at establishing a classification by the following:

a. Tests of existing traits
   1. Ability
      (a) Information
      (b) Skill
   2. Adjustment
   3. Attitude
   4. Interests
b. Tests of potential traits
   1. General capacity
   2. Specific capacity (secretarial, mechanical, etc.)[25]

Most writers include a place in their classifications for the personality tests despite Yoder's opinion that their reliability is unsatisfactory since adults adapt answers to what they think they should

---

[22]Tiffin, op. cit., p. 23

[23]Cleeton, Glen U., and Mason, Charles W., Executive Ability - In Discovery and Development, the Antioch Press, Yellow Springs, Ohio, 1946, Chap. V, p. 154.

[24]Pigors, Paul, and Myers, Charles A., Personnel Administration, A Point of View and A Method, McGraw Hill Book Co., Inc., New York, 1947.

[25]Beaumont, Henry, The Psychology of Personnel, Longmans, Green, and Co., New York, 1945, Chap. III, p. 51.

be.[26] And there is consistent agreement among writers that the tests measure either acquired skills or the capacity for acquiring skills.

An adequate classification by Scott and others is along the following lines:

(1) Tests to measure the ability to understand and use ideas.
(2) Tests to measure the ability to understand and operate things and mechanisms.
(3) Tests to measure the ability to understand and manage men.[27]

However, a division of tests into intelligence, manual abilities, visual acuity and skill, personality, and interests is satisfactory for the purposes of this paper.

---

[26]Yoder, Dale, Personnel Management and Industrial Relations, Prentice-Hall, Inc., 1946, p. 184.

[27]Scott and others, op. cit., pp. 155-156. This classification is expanded in the text along the following lines: "(1) - ability to understand and to use ideas. These ideas include: the material out of which are made all abstract and general thinking; all chemical formulas, legal decisions, and scientific principles; all the higher forms of thinking that differentiate the human from the animal; all plans, programs, values, and logical thinking such as similar and dissimilar, profit and loss, cause and effect, and right and wrong. It is spoken of as common sense, verbal intelligence, ingenuity, practicability, and educability.
(2) -ability to understand and to manage things and mechanisms includes: the motor control essential for skilled handwork; the mechanical imagination essential for the creating and for using complicated blueprints; coordination of hand and eye essential for running a machine whether it be a typewriter, a printing press, a bicycle, or an airplane.
(3) - the ability to understand and manage men includes: that which is ordinarily called 'tact' and 'diplomacy'; a winning and inspiring personality; interest in people and eagerness to serve them; and insight into motives and ability to appeal to the sympathies and dominating motives in others."

Test Administration

There are several important precepts concerning tests which it is convenient to group under the title of "Administration." Some of these should be examined briefly in order that the elementary language and principles of tests can be understood.

Testing, per se, cannot insure adequate selection. It must be preceded by job analysis; the tests must be accompanied by good administration; and the results must be recorded properly and in a manner easy to interpret. Furthermore, the tests themselves must be tested.

As for the job analysis, it is essential that it precede the testing program to find what is being done on a job before designing a test for the job. The skills, aptitude, and other characteristics that are used on the job and that the test will attempt to measure must be noted. The tests which are selected or developed should have reference not only to a particular job but to a particular company as well because of differences in job variations and in local conditions.

Since any job requires more than one ability to perform it well and since no single test can be expected to measure all of the needed capacities or abilities, it is better to use a combination of tests. At this point, the question will arise of just which tests to combine in the test battery. For example, a grouping of tests might be made to obtain a picture of general aptitude as follows:

(a) Finger dexterity or manipulative skill
(b) Accounting aptitude - clerical aptitude
(c) Ability to visualize structure
(d) Tweezer dexterity
(e) Inductive reasoning
(f) Creative imagination
(g) Visual memory
(h) Observation
(i) Personality
(j) Tonal memory[28]

---

[28]Jucius, op. cit., pp. 212-213. The author also mentions the following as typical batteries:

For inspectors -
(a) The Minnesota rate of manipulation test to measure hand and finger dexterity.
(b) The Purdue hand precision test to check coordination of hand and eye.
(c) Tests of reaction time and strength of grip.
(d) Steroscopical vision tests.
(e) Measures of height, weight, and age were also recorded.

For supervisors -
(a) Tests of mental ability
    (1) English vocabulary
    (2) Otis Self Administering Test of Mental Ability
(b) Test of personality
    (1) Personality inventory
    (2) Vocational interest bland
(c) Tests of visual perception
    (1) Minnesota Vocational Test for Clerical Workers
        Test 1 (Number checking)
        Test 2 (Name checking)
    (2) Revised Minnesota Paper Form Board Test

To measure vocational aptitude for sales, technical, and executive ability -
Grand information test
Arithmetical reasoning test
Judgment in estimating test
Symbolic relationships tests
Reading comprehension tests
Vocabulary test
Interest inventory
Dominance-submission
Independence-dependence
Extroversion-introversion
Sociability
Judgment of human nature

It should be remembered that the personality of the tester has a great deal to do with the reaction of the applicant. The tester must be able to approach the test with a completely objective technique but with an understanding of humans. It is equally as important that the person taking the tests be placed completely at ease and that every trace of nervousness be dispelled. The person giving the test must "have human sympathy and self-control. There must be sincere appreciation of human weaknesses and a capacity for quiet and keen observation of behavior."[29]

When the program is first introduced, either during the experimental phase or as a completed package, the worker will, as a rule, want to be reassured about the function of the program as it applies to him as an individual. He should be convinced that all the test results will be confidential and that the tests will not be used to his detriment.

The manner in which the results are presented is important. Briefly, the results can be expressed in arithmetic terms or in verbal terms such as percentiles, ranks, norms, etc. In the case of verbal terms, the accompanying disadvantages are that many people are unfamiliar with these terms and that the scores are hard to visualize.

An alternate means is to portray the results graphically as in the "profile" chart. This is prepared by drawing a connecting line

---

[29]Leake, M. Martin and Smith, Thyra, The Scientific Selection and Training of Workers in Industry, Isaac Pitman and Sons, Ltd., London, 1932, p. 12.

between the various test scores and enables one to see at a glance just where the testee scored with respect to the critical limits.

Testing for Executive Abilities

The field of applying testing techniques to the choosing of executive talent is being explored more intensely. As far as the companies are concerned, it is being done as a matter of efficiency. Why select a man and spend years in training him and expend money on his development only to find that he does not have the qualities for successful administration of executive duties? Testing in this area has not been too encouraging and Cleeton and Mason suggest that it is probably due to lack of understanding of the experimental procedures necessary in developing valid measuring devices.[30] However, Social Research, Inc., headed by Burleigh Gardner, reports that it has the answer to spotting executive failures in advance. As reported in Newsweek magazine, Social Research measures the traits which may cause executive failure by administering a "thematic-apperception test" developed at Harvard and the University of Chicago. The subject is shown ten pictures and asked to tell a story about each, describing what has happened, what is happening, and what the outcome will be.[31] (Those familiar with testing will note a similarity to both the Rorshach technique and the psychodramas of diagnostic psychiatry.)

---

[30] Cleeton and Mason, op. cit., p. 131.
[31] Newsweek, Vol. XXXII, No. 3, July 19, 1948, pp. 57-58.

Testing as an Aid in Training

Testing has helped strengthen training programs by indicating who should be trained, where training should begin, and whether training has been adequate.[32]

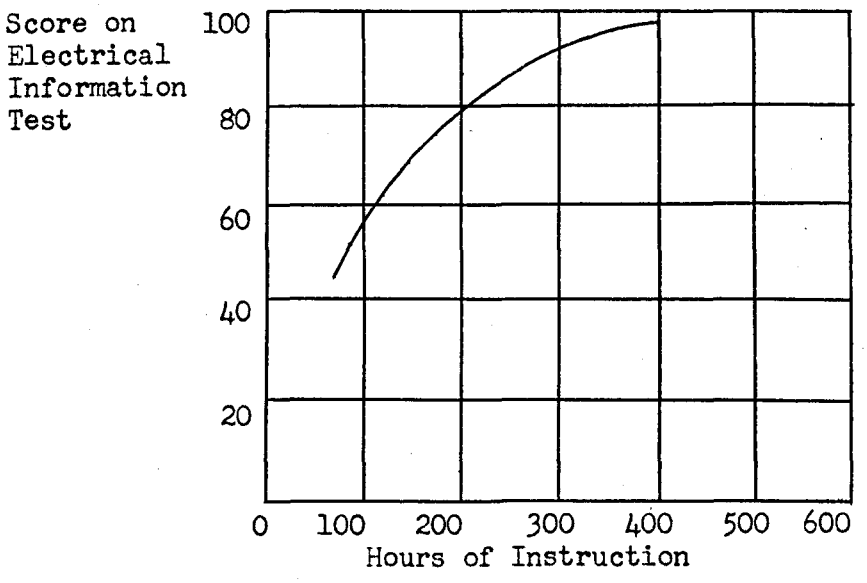
---

[32] In Jucius, op. cit., pp. 205-209 we find several good examples of specific use of tests in answering those questions about training:

" (1) - Learning costs of employees who had scored lowest on a finger dexterity test, as measured by a simple peg board, averaged $59 before they earned the minimum hourly rate on a piecework basis. Those with highest finger dexterity incurred a learning cost of $36.40 before making the rate.

(2) - On a simple measurement question asked about an illustration showing some blocks adjacent to a scale, it was found that 70% of 650 applicants were unable to read 1/32 of an inch. Obviously, training would be wasted in these cases unless the training was started at a low enough level to teach measuring fundamentals.

(3) - By pre-establishing a measure to indicate when a person has sufficient knowledge to handle a particular job it is possible to determine how many hours of training are ordinarily required to attain the desired score. For example, if anyone with a score of 80 can be considered as ready to be turned loose on the job, then training would be needed for about 200 hours."



CHAPTER II

THE VARIOUS TYPES OF TESTS

A. Introduction

The Five Divisions of Testing

The development of tests for use in selection and placement of employees has been in five general fields, each field concerned with a human characteristic which has been found to have a relationship to vocational success: manual abilities, mental abilities, keenness and skill of vision, personality, and interests. Of these, the most extensive development has been in the field of testing intelligence, or mental abilities, probably because this type of test was developed earliest and first held promise of being serviceable. Currently, the field of personality testing is undergoing the most rapid development and psychologists are trying out many new tests to measure personal adjustment in hope that they may be adaptable later to industry.

Testing for Existing Traits and Potential Traits

The field of testing may be more easily pictured if it is divided into two broad categories, testing for existing traits and testing for potential traits. The testing of an individual for existing traits is the less difficult task for the measurement is to determine what has been gained from previous experience. A problematical situation is presented to which the testee responds by drawing on his experience for

a solution. Presumably, in a field in which he is experienced he will possess not only more information but more skill as well. An attempt to measure the existence of personality traits is made by determining how a person reacts to circumstances and events and comparing it to the reactions of individuals whose personality adjustment has been proven. The presence or absence of attitudes and interests could be ascertained in a similar fashion through the use of written questions.

But the measurement of potential traits is a more complex problem since the assumption must be made that if a number of basic capacities exist, under certain conditions, other abilities will develop. Since the development of these other abilities is dependent to some degree on factors which are not subject to control, the measurement process is not a sure one and has made this area of testing less reliable on the whole than the other. Furthermore, tests of this latter type consist usually of selected situations to which the testee is exposed and cannot be expected to include all possible situations which a person might meet.

In measuring the acquired characteristics, a test developer knows in what terms the measurement is to be made - for example, to measure typing ability, one can measure samples of typing. But in the case of potential traits, such as potential typing ability, it may be necessary to measure functions which have no resemblance to typing. The fundamental process is, in the latter instance, the measurement of capacities and is well expressed by Beaumont, "whether an individual may someday master a complex skill can be predicted only by determining whether he

possesses those skills that are basic requisites and which in time will enable him to acquire the more complex skill."[1] It is clear from this discussion why a pre-typing student might be tested with a manual test such as the Purdue Peg Board Test in order to measure his arm and hand movements and his finger dexterity.

Construction of Typical Tests

Before we discuss the types of tests it will be helpful to review some of the methods by which tests are constructed and some examples of what they require of a testee.

A test may be either broken down into its component parts, such as arithmetical, reasoning, memory, and others in the case of an intelligence test; or the subtests can be grouped together and the items mixed in what is called an omnibus test. If the easiest parts of each test are grouped in the first section of the test and the hardest items saved until nearer the end, it is called a spiral omnibus, a scaled test, or a cycle test.

A convenient method of studying typical test construction is to study examples of the methods used. In the field of mental ability there are first the verbal and performance questions. These usually require the writing in of words in blanks, the naming of missing parts in pictures, or the marking of something wrong in a picture. An example would be:

---

[1]Beaumont, Henry, The Psychology of Personnel, Longmans Green and Company, New York, 1945, p. 75.

One's_____in life depends upon so_____factors _____it is not_____to state any single_____ for failure.

Another form of question makes use of "opposites," "similarities," and "analogies" such as these below:

1.	cold-hot	same-opposite
2.	long-short	same-opposite
3.	clothed-naked	same-opposite

1.	red white green	rose, paper, grass, soft, blue
2.	apple peach pear	seed, tree, plum, juice, peel
3.	pan bowl basket	pail, handle, knife, fork, spoon

Finger is to hand as toe is to	foot, knee, arm, shoe, nail
Car is to highway as boat is to	river, dock, ship, sail, water
Pencil is to lead as pen is to	pig, ink, write, paper, hand

A test for the range of information of a testee might allow a multiple choice response to questions of general information:

The pitcher has an important place in
Tennis Football Basketball Handball
Cribbage is played with
Rackets Mallets Dice Cards
The Holstein is a kind of
Cow Horse Sheep Goat

Testees are given an opportunity to show how well they are able to follow directions (which is really a test of attention and memory) by setting up confused written directions for them to carry out:

1. Cross out the smallest dot · · · · · ·
2. Put a comma between these two letters G H
3. Do you understand that each letter is to be made like printing and put in the parenthesis after the problem? If so, write C in the parenthesis. ( )

Arithmetical questions are set up in the following manner:

1. Cross out the number which does not belong in the following series. 2 4 8 10 16 32

   72 36 18 9 6

2. If a pair of shoes were bought for $3.50, what change would be received from a five dollar bill?

   2.50 1.50 1.00 3.50 .50

To attempt to measure planning ability, some tests resort to the use of mazes although they are not common in employment tests. Reasoning and ingenuity may be measured by combination true-false and mutilated sentence.

1. Cows milk give true-false
2. Write are with to pencils true-false
3. Live not are in houses to true-false

A simple type of memory test is one in which the testee is shown a number of pictures and names under them for a few minutes and later tries to link together the names and faces.

As to manual ability, most of the tests measure the underlying functions such as precision of finger movements, coordination of eye-hand, and manipulative ability of hands and fingers. In addition, block visualization and block assembly measure one's perceptive and structural visualization abilities. Most of the dexterity, coordination, and manipulative skills are tested by the use of pegs, boards with holes, pins and washers to be assembled, patterns to be traced, and items to inspect. (This paper will not take up the mechanical comprehension tests and tests of technical intelligence which have to do with what can be labeled mechanical ability as opposed to manual ability.)

In the field of personality and interests, some tests attempt to measure judgment with a multiple choice of response to an imagined situation.

John does not smoke. He is invited to a smoker. If he accepts the invitation it would be best for him to:
_____Try to smoke
_____Refuse politely without comment
_____Apologetically say that he has not yet learned to smoke
_____Explain the injurious effects of smoking

In this same area the true-false method of measuring observation of human behavior has been occasionally resorted to:

T F 1. In pleasure the corners of the mouth are pulled down
T F 2. We generally like those who bring us good news
T F 3. All men are created equal in mental ability

Other tests will give statements to be marked true or false which are indicative of stable or instable temperaments. Or they will list occupations, activities, or types of people for the testee to indicate his like, dislike, or indifference. Or they may submit stimuli such as standardized ink blots or special pictures and base their measurements on the testee's response.

Difficulties in Testing

The main difficulties in using tests to measure traits of any kind can be considered as falling into three categories: the enormous quantity of tests, the fact that so many of the traits are intangibles, and the difficulty of determining the most adequate and objective method of making the measurement.

It is possible to get some idea of the number of tests by merely looking over more than a full tray of publications on mental tests listed in the Library of the University of North Carolina or by reviewing one of the bibliographies in the testing field containing more than 3,000 tests of various kinds. As Laird has observed, "We do not need more tests, we need more inquiry into the existing ones."[2]

With regard to the question of intangibility, in dealing both with intelligence and personality, and perhaps interests, aspects of humans and human behavior must be coped with that are not only difficult to define but equally difficult to measure. One would assume that these aspects could be measured by obtaining a response to various imagined situations, but the very subjective nature of these factors makes it more difficult to devise situations which parallel actual ones. Furthermore, it is impossible to say that measuring in such situations will adequately sample a person's reactions so that predictability of future responses will be valid. This question lies close to the next one of how best to measure the traits.

Should measurements be made by using verbal questions and verbal response? Will a testee respond under such conditions as he would in a real situation? Should use be made of stimuli in the form of ink blots, words, or pictures and interpretations made of the testee's oral response? Can he express his reaction naturally in words so that we can rely on the words for measurement? Will he express his true reaction?

---

[2]Laird, Donald A., The Psychology of Selecting Employees, McGraw Hill Book Co., New York, 1937, p. 221.

Or take a narrower problem: the measurement of a single component of intelligence, reasoning. Is it fair to assume that the solving of a syllogic problem as the one following calls for the use of reasoning to an extent measurable and to an extent comparable to industrial situations: "George is older than Henry. Henry is younger than Bill. George is younger than Bill. True or false."? Or by asking, in writing, that the testee point out the relation of the letter O in the following series:

1. A O U U A
   U O A U A
   A O A A U

Can we assume that the other elements such as visual perception and familiarity with the alphabet do not enter this test so much as to destroy its usefulness for measuring reasoning power?

The answer lies in the fact that the mind cannot be divided into clear cut categories of one kind or another. It is impossible to devise a test which measures one mental aspect to the exclusion of all others. And when a test is said to measure memory, it doesn't mean that it measures only that quality. As Burtt states, "If a person hears a list of words and then tries to reproduce the list, his efficiency will depend not only on memory but also on the extent to which he pays attention to the original reading."[3] When a test is referred to as one of ingenuity, or memory, or learning it means that the test emphasizes that more than anything else.

---

[3]Burtt, H. E., Principles of Employment Psychology, Houghton Miflin Co., New York, 1926, p. 63.

B. The Measurement of Intelligence

Aspects of Intelligence

The problem of measuring intelligence has been approached by determining what factors go into the make-up of intelligence, then testing those factors after they have been isolated. Laird names seven factors contributing to general intelligence isolated in a study by Thurstone:

1. Number facility
2. Word fluency
3. Visualizing ability
4. Memory
5. Perceptual speed
6. Inductive thinking
7. Verbal reasoning (inductive thinking applied to words)[4]

In the Personnel Journal, Richard Schultz advances the idea that there are six primary mental abilities in which we should be interested when devising a test of intelligence:

1. Number facility
2. Space perception
3. Word meaning
4. Word fluency
5. Reasoning
6. Memory[5]

(This does not differ widely from the factors listed by Thurstone.)

To test these aspects, most intelligence tests are made up of similarities and opposites (including analogies and proverbs), information

---

[4]Laird, op. cit., p. 273

[5]Schultz, Richard S., "————", Personnel Journal, Vol. 25, September 1946, p. 3.

(including vocabulary), completion (sentence completion, disarranged sentences, and narrations completion), arithmetic, and number series.

Some complaint has been expressed against the make-up of the usual intelligence tests because they favor words, numbers, space-forms, and pictures and ignore three dimensional objects and situations containing human beings. But as Poffenberger observes, it seems to take a "higher grade of intellect to deal with symbols, abstractions and general notions than with concrete objects and particular situations."[6] Mental alertness tests attempt to measure the speed and accuracy at which one understands and reacts to these ideas, symbols, and relationships.

For the purposes of this paper it will suffice to define intelligence as simply the application of mental abilities to a situation or task. It can then be said that an intelligence test represents a sampling and grading of what people can do when faced with situations and tasks. Admittedly this generalizes an approach to the question "What does an intelligence test do?" and it should be said further that mental ability is the way an individual uses his intelligence to meet a new situation by devising new methods or by adapting the results of previous experience. Tests attempt to relate the amount of this ability possessed by one individual to the amount commonly held by other individuals. This concept agrees

---

[6]Poffenberger, A. T., Principles of Applied Psychology, Appleton Century Co., New York, 1942, p. 286.

with Thurstone's idea that the use of intelligence requires the application of the several factors which make up man's intelligence.

Even though this capability of man is among the most abstract, the various tests applicable to intelligence have been among the most commonly used in the field of measurement of humans. As noted by Poffenberger[7] this has been due to the belief that intelligence is necessary for vocational success. Further than that, the amount of intelligence of any individual is indicative of the vocational level to which he may aspire. Expressed another way, his intelligence is a measure of whether he has too much or too little mental ability for satisfactory adjustment to a particular job. For this reason measurements of intelligence are extremely important in the selection and placement of employees.

Intelligence Tests in Practical Use

In 1945 Kornhauser asked most of the nation's experts in industrial psychology "How well do the tests work out in classifying people's mental ability in business and industry?" and received the following replies:

Extremely well	7%
Rather well	60%
A slight help	33%
No help at all	0%

Two statements made by individuals in this survey are worth repeating:

---

[7]Ibid, p. 283.

(1) for industrial situations they need to be supplemented by special ability tests and (2) for clerical work they meet the need rather well.[8]

Hay reports a correlation between intelligence test scores and production records of bookkeepers of .35.[9] However, on the basis of the number of bookkeepers involved (40), this figure scarcely has significance. In the group of the 20 poorest operators were 9 with an I. Q. of less than 88; and in the 20 best operators there were only 3 with I. Q.'s lower than 88.

Several significant studies have been reported by Tiffin[10] which show the relationship of intelligence to success in clerical work. One of these showed a correlation from .34 to .57 between Otis intelligence test scores and job efficiency among bank employees. Another reported a correlation of .57 between transactions handled by cashiers and a test battery which included the Otis test, a sales checking test, a change making test, a manual dexterity test, and the Bernreuter scale. A third study brought to light the interesting fact that the employees who missed the most questions on the Otis test were the ones most criticized by company supervisors. Another showed that the greatest turnover on easy jobs was by employees

---

[8]Kornhauser, A., "Are Intelligence Tests Worth While," American Magazine, Vol. 140, July 1945, p. 40-41.

[9]Hay, Edward N., "Tests in Industry. Practical Illustration," Personnel Journal, Vol. 20, May 1941, p. 1.

[10]Tiffin, Joseph, Industrial Psychology, Prentice Hall, Inc., New York, 1946, p. 54.

making higher scores and the greatest turnover on the more difficult jobs was by employees with lower scores. This latter study points out one definite advantage in using intelligence tests: placing individuals on the proper job.

Pond and Bills have reported a study which compared the test scores to jobs held and drew the following conclusions: (1) that there is a definite relationship between intelligence and advancement in clerical work and (2) the relationship holds independently of the amount of schooling.[11]

Shellow mentions a significant correlation in connection with an intelligence test she devised covering completion, tabulation, syllabizing, spelling, abbreviation, paragraphing, judgment, analogies, proverbs, sentence corrections, and grammar rules. She found that the test scores for stenographers correlated .73 with rankings of these stenographers.[12] (Some of the items in Dr. Shellow's test are not really intelligence items, but more in the nature of trade test questions.)

Although there seems to be no question about the relationship between clerical ability and intelligence, there is some doubt as to whether it is necessary for executive success! Cleeton and Mason report that experiments have found intelligence tests have no validity when used to predict executive ability.[13]

---

[11]Pond, M. A., and Bills, M., "Intelligence and Clerical Jobs, Two Studies of Relation of Test Score to Job Held," Personnel Journal Vol. XIII, 1933, pp. 41-43.

[12]Shellow, Sadie M., "An Intelligence Test for Stenographers," Journal of Personnel Research, Vol. V, 1926, pp. 306-308.

[13]Cleeton, Glen U., and Mason, Charles W., Executive Ability. Its Discovery and Development, Antioch Press, Yellowsprings, Ohio, 1946, p. 134

In addition to these studies, Poffenberger has commented on the close relationship between turnover and mental alertness, accenting the previously mentioned advantage of using intelligence tests to determine on what job to place an individual. As he points out, there is a need for setting an upper limit to scores for a job because proper adjustment means satisfaction as well as high output. He adds further that "a person who is vocationally unsuccessful is in an occupation which fails to make sufficient demands on his intellectual capacity to keep him interested and at work."[14]

Scott emphasizes much of what has been said with his comment that mental alertness scores are more prophetic of the future wage or salary of any large group of applicants than any other single measurement.[15]

To summarize the general position of intelligence tests in selecting and placing employees the following conclusions can be drawn:

(1) Such tests are probably more important for selecting clerical employees than for industrial employees.

(2) They should be a part of a testing program which includes tests to measure other human traits.

(3) Intelligence tests have a proven correlation of a high order to clerical success.

---

[14]Poffenberger, op. cit., p. 301.

[15]Scott, W. D., and others, Personnel Management, McGraw Hill Book Co., New York, 1941, p. 155.

(4) Intelligence tests are helpful in placing employees on jobs where they will receive satisfaction in their work.

C. The Measurement of Manual Abilities

Defining Manual Abilities

Maier points out that what is generally called mechanical ability consists of a combination of general intelligence and certain aspects of muscular ability.[16] Among mechanical ability tests are those that measure this intelligence factor, such as understanding and comprehension of mechanical relationships. The other factor, motor ability, is the one which will be covered in this paper. It is intended to discuss only the tests which relate to finger dexterity, coordination, rhythm, speed, reaction, and precision.

Aspects of Manual Ability

Since ability represents the things we are able to do without further training, the measurement of ability consists of finding out what a person can do. In many organizations it is necessary to determine how much muscular coordination, bodily dexterity, or manipulative ability an individual has before assigning him to work. An applicant with good eye-hand coordination would not necessarily make a good assembler of aircraft instruments because that type of work calls for precise movements of the fingers. The girl who can coordinate the movements of both hands well would not make a good currency counter unless she also had finger dexterity and eye-hand coordination. There is no advantage in trying to judge

---

[16] Maier, Norman, Psychology in Industry, Houghton Miflin Co., New York, 1936, p. 170.

applicants by size or shape in an effort to determine their manual abilities because no relation has been found between anthropological measurements and dexterity.

The various tests of dexterity have indicated a wide range of individual differences and show little or no correlation with each other. One study in which a large number of motor tests were measured showed an average correlation of plus .15.[17] The high amount of coordination may be limited to certain parts of the body and may vary for heavy and light work, detailed and rough work, and fast and slow work. Since the skills are so highly individualistic the tests for them are relatively simple.

Manual Tests in Practical Use

Typical Tests which may be used are not discussed in detail but are mentioned briefly below:

Finger dexterity can be tested by placing pegs in holes arranged in various patterns which determine the relative importance of finger and arm movements.

Precision may be measured by requiring the testee to plunge a stylus in a hole which is uncovered rapidly, or by having him follow a line on a revolving drum or record.

Speed can be tested by seeing how rapidly the testee can tap a stylus. The tapping can be measured by an electric counting device. This is also a measure of arm coordination.

Rhythm is measured with a telegraph key on which the testee duplicates the pattern produced on a record.

---

[17]Garfiel, E., "The Measurement of Motor Ability," Archives of Psychology, Vol. 9, 1923, No. 62, p. 32.

Reaction has been tested in experimental fashion by psychologists for many years by noting and comparing the time between a signal, such as a flashing light, and the response of the testee.

Coordination has also been measured frequently. The most typical technique requires the subject to make different types of movements with each hand in order to control the action of some object. One test allows the vertical and horizontal control of a beam of light by means of two cranks, one of which is operated by each hand. The testee uses the cranks to make the beam of light follow a particular course.

In actual practice, a combination of several types of tests which would measure the different kinds of motor skills required in any task would be the most advisable procedure.

D. The Measurement of Visual Acuity and Visual Skill

Practically any job in industry requires good general vision and a great many jobs call for skill in a specific visual function. Inspectors need keen vision at close distances; truck drivers need perceptual ability in discriminating between relative distances of objects; and dye workers must have good color discrimination.

Most firms employ some sort of eye test, either administered at the office or as a part of the medical examination. But the difficulty is that most such tests measure the ability to see at standard distances. While this type of test is better than none at all and although it may weed out some employees whose vision is definitely deficient, it is not at all good in predicting job success. For example, one firm actually found that its eye test results were causing it to hire the poorer workers.[18] In this organization, a particular operation required continuous sharp vision at a distance of eight inches. And the operators with poor vision were consistently producing more work. The test used to select operators by measuring their distance acuity was causing the company to reject employees whose vision was keenest at this short distance.*

The type test used was one of the kind commonly used, such as the Snellen test which consists of several rows of block letters of

[18]Tiffin, op. cit., p. 139

*The specific relation between visual acuity at 20 feet and production was minus .60.

decreasing size and which is administered by determining separately for each eye the smallest letters that can be read. As Tiffin points out, this type of test is subject to external influences such as illumination, glare, and opportunity to memorize the chart.[19] In addition, it measures "readability" of letters, giving the more literate applicants an advantage.

Attention is currently being given to the fact that there are other characteristics besides distance acuity which have a relationship to job requirements. Two of these characteristics are depth perception and the postural characteristics of the eyes. Depth perception is a familiar term, while the latter expression is somewhat novel. The postural attitudes of the eyes, called phorias, have reference to the posture that the eyes assume under test conditions which eliminate the need for both eyes to converge on a single point. In this situation the eyes may either converge or diverge from what would be required in normal seeing at the test distance. The postures are then measured in terms of angular deviation from that standard in both lateral and vertical directions.

The reason for measuring these unfamiliar aspects of vision is to provide a means for efficiently utilizing an employee's visual assets. The Bausch and Lomb Optical Company has developed a machine called the Ortho-rater to do this measuring and contends that its use allows a company to place applicants, who would ordinarily be rejected, into jobs commensurate with visual ability. Before the test results can be accurately applied,

---

[19] Ibid., p. 128.

an organization must first make a survey of the visual requirements of its jobs in order that a "job standards profile" (similar to a job analysis) can be prepared. The developer of this instrument indicates that an intimate knowledge of the visual characteristics of employees will increase quality and quantity of production by amounts varying from 15 to 20 per cent. Tiffin reports a little more objectively the following results in connection with tests for phorias:

Average Hourly Errors of Clerks Who Passed and Failed Vision Tests[20]

	Passing Test	Failing Test
Phoria, Far	.26	.37
Phoria, Near	.22	.51

The need for this type of measurement has not been completely demonstrated to management although the Purdue Industrial Vision Institute says that the instrument has been thoroughly tried in a variety of industrial plants and is in use by some of the leading companies in America.[21] Most of the recent texts in the personnel field deal briefly if at all with visual skills and their measurement despite the fact that such tests hold promise for good results in placement of employees if additional experience continues to demonstrate their validity.

[20] Ibid., p. 139.

[21] "Industrial Vision Institute," reprint from Industrial Medicine, April, 1945, p. 342.

E. The Measurement of Personality

General

The significance of intelligence and ability has been impressed on most personnel men so thoroughly that they have been frequently guilty of overlooking the significance of personality traits. Intelligence and ability tests are not only more available but are more objective; tests of personality are less so because they measure a subjective trait. Regardless of an employee's mental ability and skill, unless he is on a job where he has no contacts with others and where his responses to situations are less demanding, he must fit into an organization from the standpoint of personality. He is thrown into contacts with others, with situations, and with things. He must make adjustments called for by contacts with employees and supervisors; he must respond normally to situational changes; and he must be suited for handling machines, materials, or tools. If his characteristics make it impossible for him to work with others, he will fail or his success will at least be improbable. In order to help measure the presence or absence of specific traits of this kind, psychologists have attempted to develop scales to test personality. These are becoming more available in standardized forms but personality is still the most neglected field of test development as far as industry is concerned.[22]

In dealing with these traits, evidence has been found to support the theory that they are consistent to some degree over periods of time.

---

[22] Cleeton and Mason, op. cit., p. 142.

The correlations reported in Poffenberger are shown below:

Repetition of Bernreuter Test After Long Intervals[23]

	Nervous State	Self Sufficiency	Dominance
1 YEAR, 30 cases	.78	.61	.53
2 YEARS, 15 cases	.67	.38	.54

The significance of the correlations in this instance is low enough to raise a question as to their value.

Difficulties in Measuring Personality

In ordinary testing, the test results are measured against some criterion in order to know what the testee has done in comparison with others. In the field of personality testing, it is this problem of developing reasonable criteria that balks the psychologist. For example, in measuring the amount of extraversion of an individual by asking him questions, which answers denote presence of that quality? In constructing these tests, reliance has been placed on the experimental technique of putting the question to a number of individuals whose characteristics of personality were already known, either from clinical observation or personal contact. Then the answers of the person being tested can be compared to the previously classified answers.

In personality testing, the unreliability is considerably greater than in tests in which one must arrive at a correct score or perform some task in a specified manner. The answers to personality test questions can

[23]Poffenberger, op. cit., p. 334.

be twisted. The testee is responding verbally, which is an unemotional process, to a series of written questions and even if a testee is sincere his answers may not reflect what his response would be under real conditions.[24] Where such tests are used before an applicant is hired, he may intentionally change his answers to what he feels are desirable ones to help him get the job.

Practical Use of Personality Tests

Among the first tests of this type to receive wide use was the Bernreuter Scale developed to measure four different personality traits. The test itself consists of a series of questions which can be scored in such a way as to measure emotional stability, extraversion-introversion, self sufficiency (absence of need for companionship, encouragement, and sympathy), and dominance-submission. As previously indicated, the difficulty in using this test is in encouraging an applicant to give frank answers about himself. Some of the questions such as "Do you consider yourself a nervous person?" or "Are you easily discouraged?" are so obvious in what they are getting at that the ordinary employee is inclined to answer them to his best advantage. The test is satisfactory enough, if its limitations are recognized, to be used for placement purposes but its reliability would be subject to doubt in instances where it was used prior to employment.

---

[24] Beaumont, op. cit., p. 75.

Most tests in the personality area have been adaptations and revisions of previous scales, each new one incorporating a means of measuring some additional element. The Humm-Wadsworth Temperament Scale was developed in this way primarily for industrial use and has been used more extensively in industry than any other personality test.[25] The scale has some 300 questions which are to be answered yes or no and by differential scoring the authors claim to be able to classify temperament according to components usually associated with mental abnormalities, the assumption being that even normal people have varying degrees of these abnormalities and that those with excessive amounts can be recognized. It would probably be easy to recognize the traits in someone with whom one comes into much contact, but the possibility of observing them accurately in an applicant or an employee in one or two interviews is so remote that it makes these tests appealing. The components that this test tries to measure are these:

N Normal - self control, self improvement, inhibition
H Hysteriod - self preservation, selfishness, crime
H Manic cycloid - elation, excitement, sociability
D Depressive cycloid - sadness, retardation, caution, worry
A Autistic schizoid - daydreams, shyness, sensitiveness
P Paranoid schizoid - fixed idea, restiveness, conceit
E Epileptoid - ecstasy, meticulousness, inspiration

The test has received rather extensive use, particularly in the aircraft industries and utility companies on the West Coast.

In one organization, a policy was adopted of rejecting applicants

---

[25]Tiffin, op. cit., p. 113

who were weak on traits N or H, with P higher than normal, with M, D, or A high or higher than normal, and with high E scores accompanying physiological evidences of epilepsy. They used the test on 185 engineering employees and found 184 met these standards. The doubtful one and one other were discharged later for maladjustments.[26] The most encouraging uses of this test have been in eliminating "unfits," those whose maladjustment is relatively obvious in the tests, and in helping present employees overcome weaknesses. If they are interested in self improvement the resultant profiles show personality deviations which can be used to counsel them. The authors claim that the test can be relied on in 60% of the cases, but as Tiffin shows, the scores on the seven components may shift when a person is changed from a "frank" situation to a "job application" situation. He showed that the scores had a tendency to be changed towards the normal trait and away from the undesirable traits.[27] Humm has attacked this criticism by evaluating the number of "no" answers given and making an analysis of the balance among the various components.[28] Since an extremely negativistic person accumulates more than the expected number of "no's" and the extremely suggestible persons tend to accumulate fewer than the expected number, an acceptable range for the "no" criterion can be established. Humm has also developed a technique for correcting scores with unacceptable "no" counts.

---

[26]Ibid., pp. 113-119.
[27]Ibid., p. 117
[28]Lawshe, C. H., Principles of Personnel Testing, McGraw Hill Book Co., New York, 1948, p. 81.

Several relatively new tests are making their way into the industrial field and encouraging studies of their value are being recorded as experience with them increases. These include the Minnesota Multiphasic Inventory, the Rorschach Test, and the Guilford-Martin Inventory.

The first of these, the multiphasic inventory, will probably prove most adaptable to industry as it is highly reliable and relatively easy to administer, taking about forty-five minutes in most cases. It has about 550 items in the form of questions which the subject answers, although good results are being obtained in using only 320 questions. By differential scoring, the following categories can be measured: hypochondriasis (H_s), depression (D), hysteria (H_y), psychopathic deviate (P_d), masculinity-feminity (M_f), paranoia (P_a), psychasthenia (P_t), schizophrenia (S_c), and hypomania (M_a). It can be administered satisfactorily from the manual of instructions and interpreted for employment and placement with an additional amount of training. It holds promise of being a very good test for industrial use since it is successful in pointing out those who are in need of some sort of psychiatric aid. The Testing Service of the University of North Carolina has used the test in helping with the selection of applicants for the State Highway Patrol. In one instance, two applicants on the borderline of psychoses were rejected. And in another case, an applicant pointed out by the test as doubtful had to be released later from the Highway Patrol because of maladjustment. As far as can be ascertained, its use by industry thus far has not been widely reported. (Incidentally, the test has been prepared in Braille in order that it can be used in

counseling the blind.)

The Rorschach test uses imagination and association of ideas as indicators of personality types. It is based on the theory that "similar personality types have similar imaginative and associative reactions to figures which are vague and indefinite."[29] It is administered by showing a standardized set of ink blots to an individual and asking him to interpret them. All the possible verbal responses have been classified as to types of personality which use these comments. Although its industrial usage has not been sufficiently evaluated, Steiner reports an interesting study of its results for 144 clerical workers who were assigned an overall personality rating ranging from excellent to poor on the basis of their Rorschach scores:

Personality Rating Based on Rorschach	No.	Satisfactory Adjustment to Work	Poor Adjustment to Work
Excellent) Above average)	109	106 (97%)	3 (3%)
Average) Poor	35	10 (29%)	25 (71%)

When these employees were followed up a year later it was found that of those whose responses were 50% or more "poor," only 34% had made a satisfactory adjustment to work.[30]

---

[29]Beaumont, op. cit., p. 72

[30]Steiner, Matilda E., "The Use of the Rorschach Method in Industry," Rorschach Research Exchange, Rorschach Institute, Vol. XI, No. 1, 1947.

The Guilford Inventories include three different scales to measure thirteen components of personality: (1) the inventory of factors STDCR measures social introversion-extraversion, thinking introversion-extraversion, depression, cycloid depression, and rhathymia; (2) the inventory of factors GAMIN measures general activity, ascendance-submission, masculinity-femininity, inferiority of feelings, and nervousness; and (3) the Personnel Inventory I is intended to measure objectivity, cooperativeness, and agreeableness.

It is interesting to look at some of the questions used in the GAMIN scale to measure the various components:[31]

General activity - Are you inclined to be quick in your actions?
Are you inclined to rush from one activity to another?

Ascendance - Do you usually speak out in a meeting to oppose someone you know is wrong?

Masculinity - Do you like love scenes in a movie or play?

Inferiority - Do you suffer keenly from feelings of inferiority?

Nervousness - Do you often become irritated over little annoyances?

The latter scale of Guilford's measures three of the sub factors of the paranoid component[32] and Martin believes it is the mose useful in identifying the potential troublemaker, citing the following cases:

---

[31]Martin, Howard G., "The Construction of the Guilford-Martin Inventory of Factors GAMIN," Journal of Applied Psychology, Vol. 29, 1945, p. 298.
[32]Lawshe, op. cit., p. 83.

Fifty-one employees in a manufacturing plant and forty-three textile mill employees were rated by supervisors as satisfactory or unsatisfactory and were tested on this scale. In each instance, the test properly placed between 70 and 75% of the employees.[33]

Other Methods of Measuring Personality

Before leaving this subject, the part played by two other means of measuring personality should be noted. The first of these, the Thematic Apperception Test mentioned in Chapter I of this paper, follows the general principles of the Rorschach technique. It relies on verbal responses given by a testee who is confronted with several standardized pictures. He is supposed to relate what has happened to the subject in the picture, what the present situation is in regard to the subject, and what the future will bring. The test administrator then compares the response to typical responses representing known components of personality. So far, the test has been used only in the testing of executive traits where personality is of prime importance. Wyatt makes the statement that this test is "one of the foremost devices in the study of personality."[34] Whether the test can be successfully adapted to industry remains to be seen.

The other technique is a method known as free association which has been widely used in clinical analysis of personality maladjustment.

---

[33]Martin, Howard G., "Locating the Troublemaker with the Guilford-Martin Personnel Inventory," Journal of Applied Psychology, Vol. 29, 1944, pp. 461-467.

[34]Wyatt, Frederick, "The Interpretation of the Thematic Apperception Test," Rorschach Research Exchange, Rorschach Institute, Vol. XI, No. 1, 1947.

In this method the subject is asked to respond to a list of one hundred or more carefully selected words. The responses are noted and compared to responses indicative of certain temperament patterns for use as a basis for informal questioning. The objection to this approach has been that the distributions of responses vary considerably with factors such as maturity, race, and special interests and do not always relate to emotional elements.[35]

Honesty Tests

A person's relative honesty is a facet of personality which has been treated lightly in testing development. It seems that individuals are reluctant to admit lapses of or deficiencies in honesty and resent any questioning in this direction as an infringent of personal rights. For this reason, such testing as has been developed has not been welcomed by industry because of the expected reaction from applicants and employees. The principal means of measurement have been through the use of devices reflecting changes in blood pressure, changes in the body's electric potential, and changes in breathing. The Keeler polygraph, popularly known as the lie detector, makes use of these items as well as the variations in the pulse rate to detect deception. Despite the lack of interest in tests to measure honesty, the use of the tests have disclosed startling results. Scott mentions one of these: Among Chicago bank personnel, 15 percent or 2500 employees have, after taking the test, admitted the theft of some money, and 62% of the tellers who were tested admitted theft of

---

[35]Cleeton and Mason, op. cit., p. 158.

small or lage amounts![36]

Summary

While the term personality is not clearly defined by the majority of psychologists, it has been used in this section to indicate tendencies of behavior. These acts of behavior have a tendency to cluster into what are called components which temperament scales attempt to measure. Such tests are becoming more valuable as new data is received on them, but their best usage is in eliminating definitely maladjusted applicants and in locating employees with personality deficiencies.

---

[36] Scott and others, op. cit., p. 188.

F. The Measurement of Interests

What is the relation of the measurement of interests to the selection and placement of an individual? Bingham remarks that "—people tend to find the keenest satisfaction in those activities which challenge their sustained attention—" and that "the man who is interested in the same things that interest his colleagues - is apt to feel at home among them."[37] Success can no longer be measured in terms of output - some place must be given to the enjoyment of work.

Before examining the ways of measuring interests, agreement should be reached on the meaning of the word "interest." Woodworth proposes that we consider it "—a drive towards activity of the capacity to which it is attached."[38] The Warren Dictionary of Psychology defines it as "a feeling which accompanies special attention to some content." Bingham calls it a "tendency to become absorbed in an experience and continue it."[39] Not only is there the positive phase indicated in most of these definitions, but there is the negative side as expressed in the aversion an individual may have which makes him turn away from an object or activity. So that, as far as this paper is concerned, interest can be considered as applying very simply to an individual's likes and dislikes and feelings of pleasantness and unpleasantness.

---

[37] Bingham, W. V., Aptitudes and Aptitude Testing, Harper and Bros., New York, 1918, p. 61.

[38] Woodworth, R. S., Dynamic Psychology, Columbia University Press, New York, 1918, p. 74.

[39] Bingham, op. cit., p. 62.

Difficulties of Measuring Interests

The ability of testees to distort the results of personality tests is equally true with regard to interest tests. If the testee is an applicant and knows the nature and demands of a job, he can influence his score in the desired direction. The most acceptable use is in cases where a person is being hired anyway and will be placed or transferred later on the basis of his interests. In such a case the testee understands that it is to his advantage to report his answers as truthfully as possible. Despite this drawback to interests tests, they show two encouraging results: they have a consistency in scores over a period of years and they have a high correlation with success. Tiffin reports a study showing a correlation of .66 between interest in physical science and success in engineering.[40] And Bingham reports high correlations of near .75 in retesting after a five year interval.[41] It should be noted that after an individual has been in an occupation for several years, his interests not only become more fixed but also more identified with that line of work. For this reason, one should be cautious in interpreting the results of interests tests given to mature young men who have followed a vocation for several years.

---

[40]Tiffin, op. cit., p. 121.
[41]Bingham, op. cit., p. 74.

Interests Tests in Practical Use

Every interests test proposed has historically been received with skepticism and criticism.[42] Most of these tests have been constructed around the idea that salient characteristics of various occupations can be found by a kind of job analysis and that an individual can like or dislike the characteristics without being too familiar with the job of which they are a component. As Strong expresses it in his test manual, "Men engaged in a particular occupation have been found to have a characteristic pattern of likes and dislikes which distinguish them from men following other professions. — a man will be more effective in his vocational career if he is engaged in work that he likes —."[43] (The question naturally arises whether the likes and dislikes of the occupational group are not acquired during their experience in that occupation.) In his manual Strong gives a well known example of the relation of the amount of insurance sold to interest ratings:

---

[42] Scott and others, op. cit., p. 203.

[43] Strong, E. K., Manual for Vocational Interest Blank for Men, Stanford University Press, 1938.

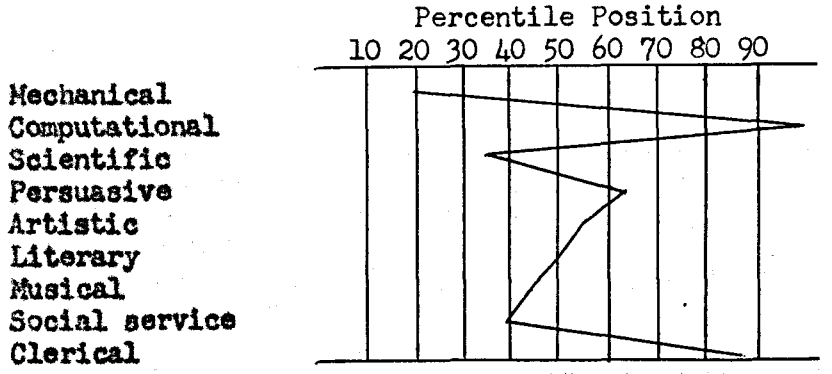
Average Amount Sold in Thousands	Number of Cases	Percentage of Each Interest Group Selling Given Amount					
		C	B-	B	B+	A	*
0-49	19	31	20	17	21	2	
50-99	37	44	20	33	26	13	
100-149	29	12	20	17	8	18	
150-199	40	6	20	28	16	26	
200-	56	0	20	5	29	41	

The two most prominent measuring devices for interests as far as industry is concerned are the Strong Vocational Interest Blank and the Kuder Preference Test. The former one of these has been in use for more than twenty years and employs the questionnaire technique allowing the testee to indicate his like, dislike, or indifference. The testee indicates this response to each of 420 separate test items including occupations, amusements, activities, school subjects, and personal characteristics. The results are then compared to those of people in various occupations. The reliability varies with the occupation for which the blank is scored, the scale for Certified Public Accountant being the least reliable. The reported reliability is frequently above .80.[44] Thurstone has made a multiple factor study of the vocational interests shown by an analysis of this blank's results which indicates that there are only four general basic interests: science, language, people, and business.[45]

[44] Scott and others, op. cit., p. 205.
[45] Tiffin, op. cit., p. 121.

*An A rating means one has the characteristics of persons successful in that line; a B rating carries the same implication with less certainty; and a C rating means one does not have such interests.

The latter test, Kuder Preference Test, has several hundred sets of questions which allow the testee to indicate what he prefers most as well as what he prefers least. A typical set of questions might ask, "Which would you rather be: secretary to a great artist, conductor of an orchestra, or a nationally known social service worker?" The test is intended to measure the relative amount of an individual's interest in nine fields: scientific, mechanical, computational, musical, artistic, literary, social service, clerical, and persuasive; and the test results are usually portrayed in profile fashion. Lawshe gives a good example of this type of profile in his group profile based on the mean performance of 27 male accountants on this test:[46]



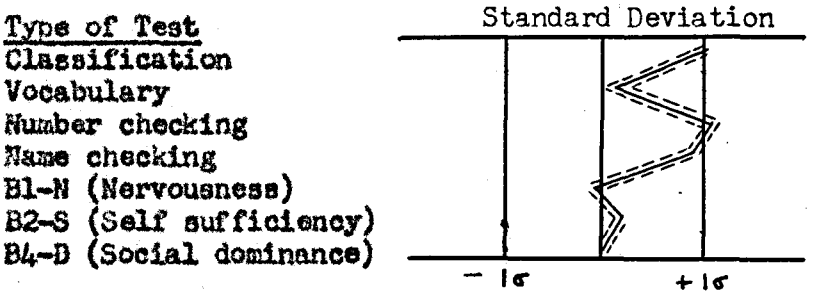
The test is extremely easy to give and takes only about 45 minutes. There has not been as much use of this test by industry as the Strong blank, but all reports indicate the test is especially helpful in placement since it allows comparison of the individual's profile with an occupational group pattern.

---

[46]Lawshe, op. cit., p. 90.

Occupational Ability Pattern Technique

The subject of interests tests is not complete without the mention of one other rather comprehensive technique of testing which seems to fall conveniently in this section of the paper. The same technique mentioned in the previous paragraph is employed to present test results in a profile representing an occupational group. However, in this method, the profile represents the results of several tests aimed at disclosing the talents in a particular vocation. The profile of the individual can then be matched against the group pattern to observe differences. For example, an occupational ability pattern for bank clerks might be prepared from mean scores of certain pertinent tests and presented in this form:



(The dotted lines represent the probable errors of the various tests used.)

While this is not exactly a pure test of interest, the relationship that exists between interests and ability and the similarity of developing occupational group patterns to the method of interests measurement are sufficient reasons for placing the discussion in this section.

Summary

An individual's likes and dislikes are related to his success and job satisfaction and by using the questionnaire method we can measure fairly well what those likes and dislikes are. By comparing them to the known interests of occupational groups a testee can be told two things: (1) if you follow this line of endeavor your interests will be similar to men who have been successful in this field, or (2) if you follow this field it is likely that you will not be well adjusted because your interests are so alien to those of people who are in this occupation. Perhaps the most advantageous use of such tests is in ruling out those occupations in which a person would not fit.

CHAPER III

TEST CONSTRUCTION, VALIDITY, AND RELIABILITY

In this part of the thesis attention will be directed at the more practical aspects of testing having to do with the development of adequate techniques. This section will deal with some of the problems of constructing a test and evaluating it in terms of validity and reliability.

By construction is meant finding criteria, assembling the test, and improving its make-up; by validity is meant the closeness of agreement between the scores and the criterion; and by reliability is meant how consistently the test measures what it is supposed to measure.

A. Construction of a Test

Criteria

The first step in the careful preparation of a test is to determine the criteria against which the test is to be measured. If the test is to select currency counters, is the criterion to be the total number of pieces counted? If the test is to select the more successful clerks for a filing department, what is to be considered as constituting success? Is it less absenteeism, fewer errors, greater earnings, or higher merit ratings?

Criteria can be divided into four groups: production data, personnel data, the judgment of others, and job samples. In the first group are such things as quantity, earnings, the time to complete a job, and quality. In personnel data, are length of service, absenteeism, rate of advancement, and training time. Under the heading of judgment would fall the ratings by others. Though this last criterion is used often in studies of test results, Lawshe notes that "most merit rating systems presently in use in industry - are not sufficiently valid for use as test criteria."[1] Job samples consist of a portion of a job which has been standardized.

In selecting a criterion we must take into consideration the distribution of the criterion data, the influence of experience, the influence of age and sex, the influence of education, and the influence of conditions of work. If the distribution of the criterion shows little distinction between the best worker and the poorest, it is not acceptable. If there is a selective process automatically going on - that is, if the best workers on bookkeeping machines are retained and the poorer ones are released - the production data would be unreliable because the range in proficiency would be too narrow. In other cases, the production of the workers may be actually limited by the speed of the machines. The extent to which experience on the job influences proficiency must be determined so that this factor will not affect the criterion. The same reasoning

---

[1]Lawshe, C. H., Principles of Personnel Testing, McGraw Hill Book Co., New York, 1948, p. 21.

applies to the other factors, as can be realized from an example given by Stead and Shartle concerning working conditions: "During the selection of the experimental group in the study of department store sales persons, it was discovered that many factors seriously affected the amount of merchandise sold. Some of these were part time demonstrating of merchandise, keeping stock in order, assisting the buyer, arranging merchandise displays, and having to sell a particular type of merchandise. Since it was impossible to determine the effect of these diverting duties on sales, it was necessary to eliminate these individuals from the experimental group."[2] The measurement of job performance should be independent of the time the measurements were made. When one person's production varies from time to time with respect to other workers, that measure is not a reliable indicator of performance. For example, on certain days each month a currency counter might have the additional task of counting the large denomination bills and her usual work would suffer accordingly.

After investigation has revealed that an adequate indicator of success is available, the next step in construction involves the selection of sets of test items. It is advisable to select criteria that will be available through the test development. Instances are frequent in which the change of methods, policy, or type of personnel interrupt the continuity of criterion data so that they are rendered useless.

---

[2]Stead, W. H., Shartle, C. L., and others, Occupational Counseling Techniques, American Book Company, 1940, p. 93.

Selecting Items

The essay type of question is considered too unreliable for test use and more usage is being made of single answer objective forms such as the true-false or multiple choice. If a true-false question is used, there is a 50-50 chance that the answer will be guessed. On a multiple choice answer with four possible responses there is only a one in four chance that the answer will be guessed.

If multiple choice questions are used, the test developer must keep in mind some of the rules suggested by Remmers and Gage:[3]

1. Prepare some hard, some easy, and some moderately difficult.
2. Avoid trick or catch items.
3. Avoid obviously wrong alternatives.
4. Make each item as short as possible.
5. Make all alternatives or responses about the same length.
6. Place correct answers in random order.
7. Avoid cue words in the root of the item.

Since many of the items will be discarded as not discriminatory or because they are poorly worded, more items should be prepared than will be needed in the final test. Lawshe suggests preparing twice as many as will be finally needed.[4]

Developing Sets of Items

In developing a suitable measuring device, there is also the problem of assembling the selected items and presenting them in a satis-

---

[3]Remmers, H. H., and Gage, N. L., Educational Measurement and Evaluation, Harper and Bros., New York, 1943.

[4]Lawshe, op. cit., p. 183.

factory form. This involves consideration of content, appearance, instructions, time limits, length, etc.

The prime consideration with regard to content is that the items show some similarity to the job. This is important both from the standpoint of better validity as well as the effect it has on the testee. Items that seem to the testee to be remote from the job will tend to antagonize him.

As to the appearance, many psychologists suggest avoiding titles on the face of tests. They say that a title such as "Clerical Aptitude Test" may have an unsatisfactory effect. This would be particularly true in those instances where a test such as this is given in connection with a mechanical job. There is not much agreement on the best way to accomplish this, although Stead and Shartle suggest not using the word "tests,"[5] while Lawshe avoids the word "examination."[6]

In arranging the items, the convenience in scoring the test should be considered. It may be that a change in the way the items are arranged will speed up and simplify the scoring of the test without disturbing its effectiveness.

The test instructions should be concise and clear and the practice exercise should be well separated from the rest of the test so that testees won't work on into the regular portion. Adequate directions

---

[5]Stead, op. cit., p. 108.
[6]Lawshe, op. cit., p. 213.

for scoring and administering the test should be drawn up so that clerical help may be used in this phase of the program.

The order of difficulty of the items should be established by administering the test to a group with no limitation on time. Item difficulty can be measured by comparing the number who responded correctly to each item. The items can then be arranged in order from the easiest to the most difficult.

The proper time limit can be established by giving the test to a sample group and preparing a frequency distribution showing the length of time required to complete the test. If the objective is to measure accuracy, the limit can be lenient. If the objective is to measure speed, the limit should be set so that few can finish all the items.

The length of the set of items will depend on how much time will be available for testing, what other tests are to be given, and how reliable the test must be.

Selecting Certain Sets of Items

It is apparent that certain of the factors surrounding an occupation will have a prominent part in indicating the kinds of tests to be used. Some of these factors are job analysis, the results of other investigators, intercorrelations with other tests, the nature of the criterion, and the group being tested.

The results of other investigations may save a lot of time by indicating which tests have been successful in testing similar jobs.

Almost any text in the field of personnel testing lists studies of various kinds which show the types of tests liable to give the best results. Stead and Shartle give results for more than 30 occupations measured by a group of 16 different tests.[7]

As for other tests, the object is to combine tests that will yield high correlations with the criteria but low correlations with each other.

If the nature of the criteria calls for speed, then the battery of tests should include a measure of speed. In listing checks on an adding machine, both speed and accuracy are essential, but the emphasis is slightly greater on speed since other arrangements are made to take care of errors. In measuring for this job some test such as the Minnesota Vocational Test for Clerical Workers should be included since the number comparison items in this test are speed items.

Item Analysis

An important part of the construction of a test is the study of the individual items, because "the effectiveness of the total test is a function of the effectiveness of the items."[8] Item analysis has two purposes: the determination of the difficulty of the item and the determination of the validity of the item. Item difficulty is measured in terms of

---

[7] Stead, op. cit., Appendix II

[8] Ghiselli, E. E., and Brown, C. W., Personnel and Industrial Psychology, McGraw Hill Book Co., New York, 1948, p. 186.

the proportion of the population tested that answers the item correctly. Item validity is measured by observing whether the workers who have high criterion scores show a better performance in response to an item than the workers with low criterion scores. Items which are correctly answered by all or not correctly answered by any have no discriminating power and should be discarded. Items on which the correct and incorrect responses do not identify the two criterion groups are not valid and should be thrown out.

Of the two aspects, item difficulty is perhaps the easier to measure. If one item is passed by 90 percent of the population and another is passed by 70 percent the first one is the easier because the chance that any one person will get it right is greater. Items selected for the final test should vary in difficulty "within the range 5 percent to 95 percent correct answers with the majority of them having a difficulty in the neighborhood of 50 percent."[9]

Guilford's comments on this question of item difficulty make it more easily understood:[10]

1. The single item that will indicate the level of ability of an individual is one for which his probability of passing is .50.
2. The most accurate test for such an individual would be made up of items of the same degree of difficulty.
3. For discriminating between the abilities of two different individuals, the best items is one that lies midway between two items that could be passed 50 percent of the time by the two.

---

[9] Ibid., p. 187.

[10] Guilford, J. P., Psychometric Methods, McGraw Hill Book Co., New York, 1936, p. 444.

This same author points out that there is a difference between discriminating value and difficulty that is pertinent to a discussion of validity and difficulty in item analysis: he observes that an item has high discriminating value if it can be passed successfully by almost all individuals who have abilities above the amount required to respond correctly and failed by most of those with abilities below that point.[11]

In measuring item validation the simplest method is to identify two groups of people known to differ on the trait and to compare the performance of the two groups on the item in question.

One variation of this method is to divide the papers on each question into two groups - one correct and one incorrect. The average score of those who answered it correctly should be appreciably higher than that of those who failed it. If the opposite result is noted or the difference is not clear cut, the item is unsatisfactory.[12]

Another way is to divide all the papers into groups classified according to scores (for example, the top fourth, the two central fourths, and the lowest fourth.) Then the number of correct answers for each item can be checked. The top fourth ought to have the largest number of correct answers to any item, and the middle group should have more than the lower. Where this relationship is not found the item is discarded. Guilford mentions a case where this method was used to trim down a test containing

---

[11] Ibid., p. 426.

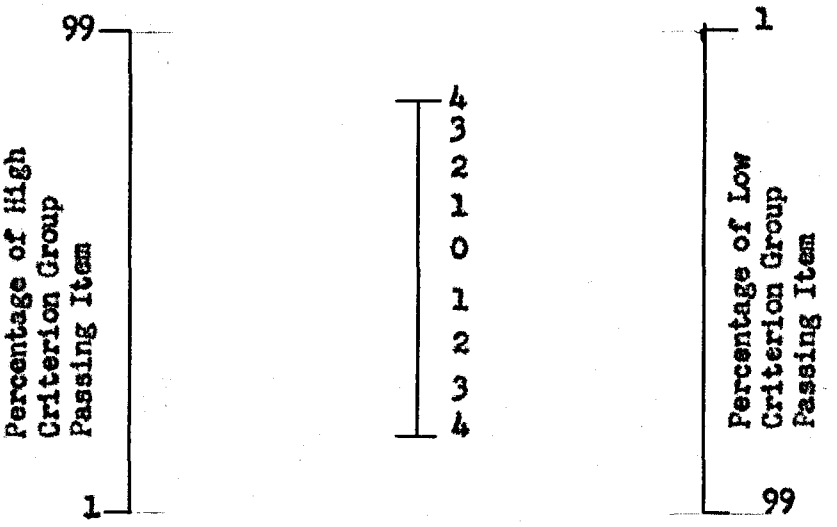
[12] Yoder, Dale, Personnel Management and Industrial Relations, Prentice Hall, Inc., New York, 1946, p. 250.

400 items with a validity coefficient of only .49 to 85 items with a validity of .71.[13]

This latter method is similar to what Lawshe calls the "criterion of internal consistency."[14] The extreme scoring groups may be taken as the criterion and the performance to the two groups may be compared. (Extreme scoring groups might be the top 20% and the lowest 20%, or the top 10% and the lowest 10%.) He has developed a nomograph to show the relationship between the two groups as a numberical value. The scale has been reproduced in part on the following page. On the left the scale shows the percentage of the high criterion group passing the item and the one on the right shows the percentage of the low criterion group passing the item. (The intervals along the two scales are not equidistant.) By connecting the points on these two scales it is possible to read off the "discrimination value" on the center scale.

---

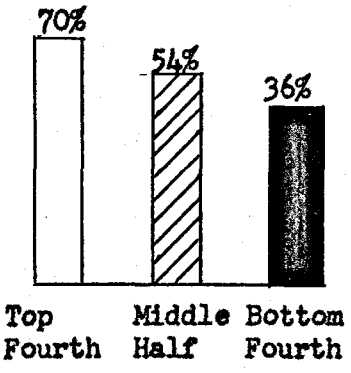
[13]Guilford, op. cit., p. 429.
[14]Lawshe, op. cit., p. 188.



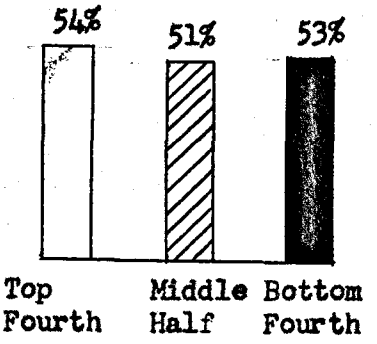
Lawshe states that a D value (discrimination value) of less than .3 or .4 means that an item would contribute very little to the validity of a test and that a D value of 1.2 is highly valid. As an example of how this nomograph can be used he shows a frequency distribution of the D values of items in an experimental test. The table is reproduced on the following page. The selection of a cut-off point at .4 enables them to discard 60 items which were felt to be of doubtful value.

D-Value	Frequency
1.1	1
1.0	1
.9	1
.8	7
.7	8
.6	13
.5	13
.4	12
.3	17
.2	9
.1	12
.0	14
-.1	4
-.2	1
-.3	2
-.4	0
-.5	0
-.6	0
-.7	0
-.8	1
	116

Another method of item validation is against external criteria. Some means is first found for identifying the groups, such as over-all grades in school. Then the performance of each group on each item is determined. Item No. 1 below would be valid and No. 2 would not be.



ITEM NO. 1



ITEM NO. 2

There are, of course, many refinements of item validation too complicated and numerous to be brought into this paper. However, one other method will indicate the direction some of these have taken. In the Vincent overlapping method, the criterion group is divided into two parts, those passing and those failing. The median score of the passing group is established, then the percentage of the failing group who exceed this median is taken as a measure of validity. A high coefficient of overlapping would indicate low validity.

$$VO = Pf > Mdn_p$$

VO = Vincent coefficient of overlapping
$Pf > Mdn_p$ = the proportion of the failing group exceeding the median of the passing group.

Battery Development

A test battery may be defined as "a group of items combined statistically into a single measure of occupational potentiality."[15] Battery is simply a term assigned to the entire group of tests which have been selected to be administered. It is not possible to combine into a battery all the measures of potentiality but only those which are most important. Whether such a battery is useful or not depends

---

[15]Stead, op. cit., p. 129.

on its validity, cost, convenience, and time required for administration.

The key problem in developing a battery is one of determining which of the trial tests are to be retained. The tester is interested in having as many tests as necessary to measure different aspects needed on the job and in keeping the battery from being unwieldly. The tester may have administered number checking, vocabulary, arithmetic, name comparison, personality, and interest tests to a sample group and will have developed correlation coefficients between the tests and the criterion. His next step will be to prepare intercorrelation coefficients to show the co-variation of the results on each test with test other test. On the basis of these data, the best tests for inclusion in the battery can be found. Tests which show a high degree of correlation with others are obviously measuring the same element in an individual and probably add little to the validity of a battery. There are several methods of selecting the best tests for the battery, ranging from casual inspection to complex statistical techniques. The Wherry Doolittle Test Selection method is one of this latter type for selecting the measures which when combined yield the maximum validity.[16] It selects the tests; it indicates when the point has been reached where the addition of another set of items to the battery adds more chance error than it contributes in efficiency; and it assigns the proper weight to each test.

---

[16]Ibid, p. 131.

B. Validity

Introduction

In considering a test to be used as an aid in the selection and placement process it is necessary to know that it actually measures what it is supposed to measure. The judgment of some expert should not be relied on to tell us that the test is valid; the test should stand or fall on the basis of scientific validation.

The means of evaluating a test in this respect is to compare it with something in the nature of an independent criterion - a condition definitely representing the characteristic to be measured. This criterion should be some element or factor about the job which is indicative of success or failure inasmuch as the question is "Does the test aid in identifying those persons who are most apt to be successful on this particular job?"

Whenever possible, criteria should be objective measures of job performance since such scores are less subject to errors of judgment and provide a more precise discrimination of differences in proficiency. When the criterion makes use of job performance, the range of proficiency of the workers should be as large as possible. If only the best or poorest workers are selected, the test operates under the severe handicap of having to make a distinction between a good worker and a better one or between a poor worker and a poorer one. The validation group should contain individuals at about the same levels of job proficiency and in the same proportions as they are found on the job.

After selecting the criterion we can compare it to the test scores by correlation analysis.

It should be noted that where several tests are being evaluated, a substantial portion of the tests will not show a satisfactory predictive power and for that reason one should start with two or three times more tests than are desired in the final battery. This will have an effect on the testing time, making it much longer for the validation study than for the time required after several of the tests are discarded.

General Methods of Validation

Since validation of a test requires some criteria with which to compare the test scores, we might select our criteria from the following: (1) production criteria, such as the number of bricks laid, the number of checks listed, or the pieces of currency counted; (2) action criteria, the measurement of the activity while it takes place, such as the speed with which 1,000 checks are sorted; or (3) subjective criteria such as rankings by a supervisor. In addition to these we might even rely on personnel data such as absenteeism, length of service, rate of advancement, training time, accidents, etc.

Lawshe suggests two basic fact finding techniques for determining whether a test will be useful: the present employee method and the follow-up method.[17] In the present employee method the first step is to analyze

---

[17]Lawshe, op. cit., p. 13.

the job to note the demands that it places on the employee. Next, the trial battery of tests is chosen with reference to the availability of employees and the relative importance of adequate placement of the particular job. Then comes "one of the most important steps in the test validation procedure,"[18] the identifying of criterion groups, one satisfactory and the other unsatisfactory. Next we administer the battery of tests and compare the test results to the criterion. If both of the groups average the same, then the battery has no value for selecting potentially successful employees on this job.

The other method consists of almost the same steps - analysis of job, selection of trial battery, and comparing test results. It is on the third and fourth steps that the follow-up method differs. The third step is to test all new hired employees and the fourth step is to classify them after a period of time on the basis of some criteria.

Ghiselli and Brown discuss the factors to be taken into consideration in these two methods.[19] Workers are not similar to applicants in training, experience, and age, and are likely to be dissimilar in terms of interest and attitude. When established workers are used, the coefficient of correlation on which the predictive power of the test is based will "in most instances be lower in value than if a group of applicants were used." But if a test does stand up under this type of validation it will have even more predictive power when used with applicants because it will have indicated its capability to distinguish between varying

---

[18]Guilford, op. cit., p. 39.
[19]Ibid., p. 172.

amounts of ability in a restricted range. These two authors believe that applicants should be the group on which final validation studies are made and that, if possible, several hundred applicants comprise the group. They also suggest that information on age, sex, training, and experience of this group should be gathered for studies on their importance on both the test and the criterion.

Techniques of Measuring Validity

The process of employing statistics to determine the reliability of observed differences and relationships so that generalizations can be made with a degree of confidence is called statistical inference.[20] This is one of the concepts used in measuring the validity of tests.

Correlation, which is a means of describing the relationship between two variables, gives us a way to see whether the test is a worth while selective device. Correlation is commonly expressed in the form of numerical values varying from -1.0 which is perfect negative correlation to +1.0 which is perfect positive correlation. If the correlation is highly significant, then the knowledge of the value of an item in one series will suggest the general position of the paired item in the other series. For example, if the test scores are known for an individual on a battery of tests which correlate +.80 with the earnings of salesgirls in a department store, the individual's probable earnings can be predicted.

Is the correlation between test score and criterion greater than would occur by chance? If the coefficient of correlation is not above

---

[20]Edwards, A. L., Statistical Analysis for Students in Psychology and Education, Rinehart and Co., Inc., New York, 1947, p. 13.

certain known chance possibilities, it shows no useful relationship. This characteristic can be appraised by reference to tables which show the fiducial or confidence limits. (In the 5% tables are shown the values of r, the coefficient of correlation, that would appear by chance 5 times out of 100. In the 1% table are shown the values that would appear 1 time out of 100. If the correlation is above the value in the 5% table it is considered significant; if it is above the value in the 1% table it is considered highly significant.)

There is a point which concerns the reliability of such estimates to be considered in predicting from correlations. The accuracy of the predictions will depend on the degree of correlation between the two series. If the correlation is high, prediction may be accurate; if correlation is low, the prediction can only be approximate. These limitations on prediction are usually indicated by the standard error of estimate. This defines the limit within which approximately two-thirds of the estimates based on a given correlation may be expected to fall. It is obtained by applying the formula $\sigma_e = \sigma_y\sqrt{1+r}$ where y is the criterion, r is the coefficient of correlation and σ is the standard deviation. For example, if the correlation between a test and the criterion is + .90 the standard error of estimate is .436 sigma units on the criterion scale. This means that the predicted score $\pm$ .436 σ would be correct about 68 times out of 100. An additional column on some tables will also show the "efficiency" of such predictions, in this instance about 56% better than chance.

The table shows that a test which correlates only .80 with a criterion will predict only 40% better than chance. As Bingham observes, this "warrants thoughtful scrutiny by any user of test data who may not have previously looked into the mathematical laws of probability."[21]

Ghiselli and Brown suggest that the lower limit of usable coefficients is in the neighborhood of .35 to .40. For tests to be included in a battery the lower limit is in the neighborhood of .20 to .25.[22] In formulating the latter statement the authors assume that the correlations between the tests to be combined is .40 or less.

A means of measuring test validity which does not involve correlation is to compute the percentage of successful workers scoring above or below various critical scores. The workers can be divided into successful and unsuccessful on the basis of job performance, and the percentage of those successful scoring at or above the score levels can be indicated. For example:

Score of 30 - 60% successful
Score of 40 - 70% successful
Score of 50 - 80% successful

In cases where the measure of criteria is not in the form of a continuous series of scores representing the various levels of proficiency, the workers can be divided into excellent, good, average, and poor and the

---

[21] Bingham, W. V., Aptitudes and Aptitude Testing, Harper and Bros., New York, 1942, p. 257.
[22] Ghiselli, op. cit., p. 185.

validity can be determined by calculating the average test score for each group.

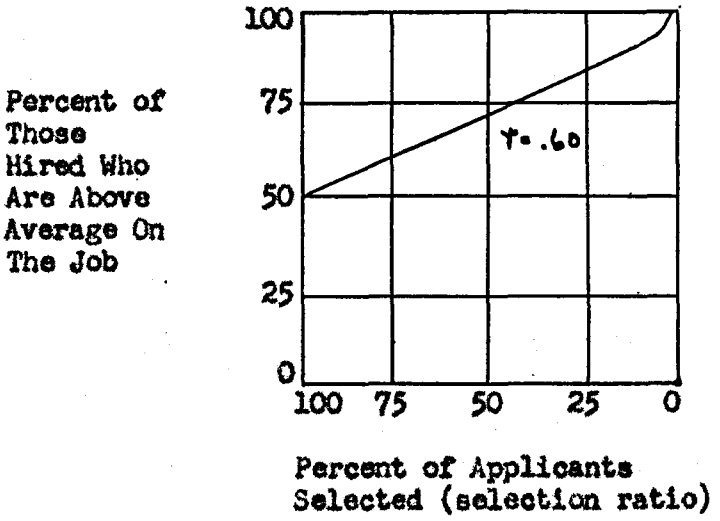
Ordinarily, however, where the criterion scores are in a continuous series, correlation methods are best.

Validity of Batteries

The validity of single tests bears a direct relationship to the correlation between test and criterion. But the validity of a test battery is a function of the correlation between each test and the criterion as well as the closeness of the co-variation between each test and every other test. The correlation between test and criterion should be as high as possible and the intercorrelation between tests as low as possible. From a validity standpoint, when adding another test to a battery does not increase predictive power enough to justify the work, the test can be ignored.

Validity and the Selection Ratio

The selection ratio is the ratio of the number of applicants to be selected to the total number of applicants available. As the chart on the following page indicates, if the selection ratio is increased, the prognostic value of the validity coefficient is decreased. And if the selection ratio decreases, the prognostic value goes up.



In the above illustration, where r is .60, if the 25% of applicants making the highest test scores are selected, 81% of them may be expected to be above average.[23] The relationship of this factor to a discussion of test validity is pertinent to anyone planning to make use of tests in selecting employees.

---

[23]Ibid., p. 183.

C. Reliability

Definition and Methods

In testing, the term reliability refers to the consistency with which a test measures what it is supposed to measure. "If individuals retain the same relative positions when measured twice by the same device then that device is reliable."[24]

There are three general methods of finding this reliability. The first way is to repeat the same test on different occasions. However, this involves the memory factor if the interval is short and the learning factor if the interval is long. A second method is to give two tests as nearly alike as possible. This avoids the above difficulties but requires the preparation of twice as much material. The third method is the split test method of dividing a test into two halves comparable in difficulty. Usually this is accomplished by considering the odd numbered questions as one half and the even numbered questions as the other half.

The disadvantage to this latter method, that the reliability of a test is related to its length, can be overcome by applying the Spearman-Brown formula for estimating the reliability of the whole form from the calculated reliability of half.

---

[24]Ibid., p. 186.

An example will show how easily this is done:

	Score on Odd	Score on Even
A	2	3
B	3	2
C	3	2
D	4	5
E	4	3
F	5	7
G	5	6
H	7	8
I	8	9
J	9	5

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{4.8}$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{5.6}$$

$$r = \frac{N \Sigma xy}{N \sigma_x \sigma_y} = .733$$

(is the standard deviation, X^2 is the sum of the squares of the difference between each odd score and the mean odd score, Y^2 is the same for the even scores, and r is the coefficient of reliability.)

.733 represents the reliability of half the test. The reliability of the whole test would be $r_t = \frac{2r}{1+r} = \frac{1.466}{1.733} = .855$.

If there is a test which has a reliability of .90 and it is necessary to know how many times as long the test will have to be in order to obtain a reliability of .95, the following variation of that formula is used:

$$A = \frac{.95(1 - .90)}{.90(1 - .95)} = 2.11 \text{ times as long}$$

(Guilford believes that the split test method when used with mental tests probably gives an estimate of reliability that is too high as it does not take into account changes in individuals from day to day and month to month.[25])

Errors of Predicting

When a score is obtained from an individual on a given test it cannot be said positively that the numerical value represents his true score. There are so many factors of unreliability and so many chance errors of administration and scoring that we must consider his score as a band or zone about the point represented by the numerical value. To measure this band the statistical device known as the standard error of estimate is employed and reference is made to the band or zone as the obtained score plus and minus this standard error. Bingham's explanation of how this operates is as follows: suppose that an individual has taken three tests, A, B, and C, with reliabilities of .96, .90, and .80 respectively. and suppose he has scored just average, exactly in the middle of the scale on all of them. His estimated true score in this case would be average also. But on test A the estimate is less likely to be in error than on tests B and C. In test A, reference to a table of the functions of r shows that the chances are 68 in 100 that his true score lies in a zone plus and minus the standard error of estimate of a true score which in this instance is .196 σ. Thus the zone runs from + .196 to - .196. On test B the standard error of estimate is $\pm$ .3 σ. And on test C the standard error of estimate is

---

[25]Guilford, op. cit., p. 411.

± .4 σ which is nearly twice the width of the zone in test A. An estimated true score is defined as the most probable score if all variable errors of measurement were eliminated. It is sometimes described as the average error the person would make if it were possible to give him a great many equivalent forms of the test under identical conditions.[26]

Most of the tests which have a value in estimating an individual's aptitudes range from .85 to .97 as to reliability and the corresponding standard errors of estimate are less than .36 but more than .17 of a sigma unit.

Factors Influencing Reliability

The reliability of a test is dependent on a large number of factors many of which are unrelated to each other. Some of the more important ones are listed below:

1. Number of items
   The greater the number of items the more reliable is the test.
2. Testing time
   The longer the test time the greater the reliability.
3. Range of difficulty
   Items that are so easy all pass or so hard all fail do not help in distinguishing individual differences. If the range is narrow, the reliability is greater provided it is not applied to a group which is too homogeneous.
4. Probability of chance
   The more likehood of chance, the lower the reliability. (In a two response test like the true-false, the reliability for a test of 100 items was found to be about .84. When the number of alternative responses was increased to 3, 5, and 7 the reliability was increased to .88, .89, and .91 respectively.)[27]

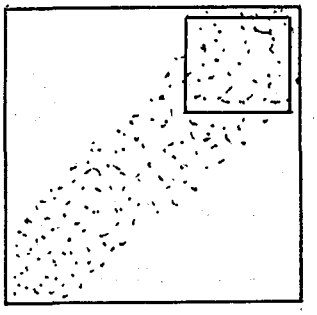
---

[26] Bingham, op. cit., p. 255.
[27] Guilford, op. cit., p. 417.

Miscellaneous factors such as catch questions, distractions, illness, worry, and cheating all reduce the reliability.

Reliability in Different Ranges of Ability

Frequently persons dealing with tests overlook the fact that the reliability is higher for a test when the range of talent is wider. This can be easily illustrated with a rough diagram. The big square represents self correlation scores for a large group of the population, such as all the students from the seventh grade through college. The small square just represents a segment of the population group such as the college seniors.



In the large population group there is a definite pattern, showing high correlation as to reliability, while in the small group the dots are very widely scattered with no clearly discernible pattern.

How Much Reliability is Desirable

There is some difference of opinion on just how high the correlation coefficient of reliability should be. Ghiselli and Brown believe that it should be .85 to .95 if the test is to be used for selection or placement.[28] Guilford thinks that for measuring individual scores it should be at least .90 and preferably above .94.[29] And Yoder says .80 correlation is desirable.[30] It is probable that all of the authors are setting about the same limits since the validity factor influences the need for high reliability. It is not possible to say that unless a test has a specified reliability, it should be discarded. A test with a validity of .70 and a reliability of .75 is better than a test which has a validity of only .50 and a reliability of .90.

Item Analysis for Reliability

As a general rule, item analysis for reliability is not carried out, for the reliability of items is assumed from the general reliability of the entire test. If it is desired to measure the reliability of the items, which in this case means whether each item is measuring what every other item is measuring, then the testees are divided into a high and low group on the basis of scores and the proportion of each group answering each item correctly is computed. "If the proportion of the high group answering it

---

28Ghiselli, op. cit., p. 186.
29Guilford, op. cit., p. 416.
30Yoder, op. cit., p. 247.

correctly is significantly greater than the proportion of the low group answering it correctly, the item is considered reliable."[31]

[31]Ghiselli, op. cit., p. 188.

CHAPTER IV

CONCLUSIONS

General

The same job in any organization is always performed in slightly different ways by each person assigned to that job because of the differences that exist between individuals. These differences may relate to varying degrees of intelligence, to dissimilarities in motor dexterities, or to unlike personal characteristics. Undoubtedly these characteristics exert their influence in making one employee better than another on such a job. One of the steps leading to adequate selection of employees is the determination of just which ones of these differences contribute to job success. The next step is to adopt selection methods which will identify the desired characteristics among applicants.

Experience has shown that such methods of selection as the use of pooled judgment, employment interviews, and application blanks are not <u>in themselves</u> highly valid determinants of the presence or absence of desirable individual characteristics. Judgment is influenced by personal prejudices; interview results fail to disclose accurately the important individual differences; and application blanks, with such items as age, education, sex, and family background, offer the narrowest clues to predicting job success. The personnel manager who chooses to rely

solely on these means of selecting employees is applying ineffective techniques. The technique of testing should be used to supplement other methods.

Despite the failure of these other methods in the field of choosing and placing employees, industries have shown a relative lack of interest in pursuing the development of more adequate techniques. Methods of reducing the costs of production have received a great deal of attention while the field of improving employment techniques has been relegated to a subordinate position.

Testing has expanded to such a broad point that anyone studying the subject must have a workable classification of tests in order to picture more easily the various fields into which it has developed. A simplified classification would include the following categories of tests: intelligence, manual abilities, visual acuity and skill, personality, and interests.

Most authorities on testing seem to agree that firms considering the installation of tests should begin their operations with these two thoughts in mind. First, that some form of job analysis should precede test installation. It is not possible to develop tests which will be adequate predictors of job success without knowing enough about the qualities needed on each job to select the appropriate tests. Second, it is better to use a combination of tests since all jobs require more than one ability and since no one test can measure all abilities.

Types of Tests

Tests of intelligence are constructed to measure the different mental elements of the over-all component of intelligence. They rely on question items such as opposites, similars, analogies, pictures of structures, coding, and number problems. The questions attempt to measure number facility, word fluency, a person's ability to visualize forms and shapes, and his memory span. In addition, they attempt to measure an individual's speed and his ability to reason. Tests of this type seem to be highly reliable in testing for clerical success. More particularly, they are helpful in determining the level of work for which an individual is best suited. Probably their best use is in eliminating those applicants whose high intelligence would make them dissatisfied with the work and those whose low intelligence would not qualify them for doing the work.

Manual ability denotes those things we can do without further training and includes such motor abilities as finger dexterity, coordination, rhythm, speed, reaction, and precision. It should be noted that the various motor abilities bear little relation to each other. An individual who has extreme skill in tapping at fast speeds may have poor finger dexterity. These tests usually consist of peg boards, assembly devices, tapping devices, etc., and one of their advantages is that they are simple and very easy to administer. Validity of these tests is highest when the tests are closely related to a careful analysis of the skills required on a particular task. For example, operators of an IBM proof machine are

required to pick up individual checks and drop them into a slot after operating a one-hand keyboard to list the dollar amount. The important phase of this operation is not the precision with which the operator can drop the check into the slot but rather her finger dexterity in separating one check from a pile of checks and her eye-hand coordination in transporting the check from its initial position to the slot opening and releasing the check at that point. In designing a test of abilities in this type of work a thorough analysis would indicate the relative importance of finger dexterity, eye-hand coordination, rhythm, speed, etc. The application of tests so selected gives test results which correlate highly with successful employees.

Since many jobs call for a specific visual skill, the eye tests being given by many firms today may be causing the rejection of applicants whose vision would make them outstanding on a job. Hosiery menders must have keen vision at approximately 8 inches. Adding machine operators engaged in listing checks must have keen vision at about 15 inches. The ordinary eye test which measures visual acuity at 20 feet does not have an adequate relationship to these specific job demands. The visual requirements of jobs should be studied in an effort to utilize most efficiently an employee's visual assets. A new technique for testing such assets has been developed by the Purdue Industrial Vision Institute and can be administered with the aid of a machine called the ortho-rater. It will provide employers with a better tool for selecting and placing employees on the basis of visual skills.

In addition to the physical and mental attributes mentioned above, an employee's responses to situational demands are another determinant of job success. In some jobs there is a definite requirement for an individual who reacts in a specific pattern to persons, things, and situations. To predict with any degree of accuracy how an individual will react requires a knowledge of the personality traits of that individual and it is the purpose of personality testing to disclose the presence or absence of the various traits. For this reason such tests are becoming increasingly important - primarily for placement purposes. Of major importance in any consideration of personality tests is the fact that unreliability is probably greater than in any other type of testing. The answers to the various questions are subject to distortion by the person tested, particularly when the tests are used in the hiring stage. Furthermore, since most of these tests require interpretation by a psychologist, their use by many personnel managers may be problematical. However, since most cities have college testing bureas or mental clinics which are usually cooperative in administering or interpreting tests of this nature, their use should increase.

Interests tests achieved their greatest success in the field of vocational conseling and have subsequently been applied to selection and placement. Since a person's enjoyment of work has a distinct bearing on his efficiency at work, it is apparent that his interests are related to his success. Interest tests are constructed around the idea that an individual can measure his likes and dislikes and that by comparing them to the likes and dislikes of occupational groups can see the similarity

of his interests to various vocations. Care should be taken in the use of such tests where the person to be tested has been in a particular occupation for any length of time because in these cases his interests have a tendency to become identified with that line of work. From an industrial standpoint, the tests offer more advantages in the placement of employees than in their selection.

Test Construction, Validity, and Reliability

Test construction involves the selection of criteria against which the results of the test can be measured, the selection of items to be used in the test, the development of sets of these items, the analysis of each item, and the combining of the sets of items into a battery. Such things as production data, personnel data, job samples, and employee progress reports are useful as criteria for evaluating tests. The choice of a test or the original designing of a test can be made much easier by studying the results of earlier investigators. After the test has been chosen or designed and tried out, it can be analyzed item by item by applying some of the statistical techniques of item analysis. This will be helpful in improving the tests and making adjustments in it with regard to difficulty and length.

The validity of a test, one of its more important characteristics, can be determined by comparing test results to some independent measure which reflects the qualities needed on the job. To phrase this another way, validity consists of ascertaining whether the persons who

do well on the job also do well on the tests. It can be measured by simpler methods such as the averages of criterion groups or by the application of statistical techniques. A coefficient of correlation of + .35 to + .40 between the test and the criterion appears to be a useful minimum standard.

Another important characteristic of tests is their reliability. This characteristic, the consistency with which a test measures what it is supposed to, is subject to more external factors than validity. It can be influenced, among other things, by the number of items in a test, the amount of time it takes to complete the test, and the range of difficulty of the items. For most tests reliability should be between + .85 and + .95 although the reliability can be slightly lower if the validity is high.

Present and Future Problems

Although testing has been extensively developed during the past 50 years, there are several problems to be faced in applying tests to the actual selection and placement of employees. Such problems can be grouped in the following manner:

A. Limitations of tests.

Personnel managers in their efforts to introduce the subject of testing to management should point out thoroughly that tests cannot perform miracles. One of the reasons why there was a delay in the development of testing techniques by industry was that numerous firms rushed to

use this new technique during the 1930's expecting the tests to work wonders. The subsequent disappointment in the results as measured by their expectations led to the discarding of test programs in a wholesale manner. Emphasis should be placed on this aspect of testing when tests are presented to management as a worth while project. Personnel managers would do well to stress that the best personnel on any job means less supervisory problems as well as more efficiency and to point out that "personnel testing can contribute to these objectives."[1] Tests should not be represented as a cure-all for management problems, and they will not always yield excellent results. A testing program can be best measured in terms of whether it selects fewer poor employees and more good employees than previously used employment practices. Most tests have either met this standard or can be tried in a particular situation to see whether the standard can be met.

B. Selecting the proper tests.

The selection of the proper tests to be used by any particular concern with regard to any particular job is a minor problem. By making the analysis of the job from the standpoint of its skill requirements and by studying the work of other investigators it is easy to draw up a list of suitable tests with which to experiment. If there is any difficulty in this case, it lies in the fact that there are too many tests to choose from and the inexperienced employer must rely largely on what he can read

---

[1]Lawshe, C. H., Principles of Personnel Testing, McGraw Hill Book Co., New York, 1948, p. 9.

about the various tests. Bingham, in the appendix to Aptitudes and Aptitude Testing,[2] presents a thorough summary of many of the outstanding tests as well as statistical information about them showing their usefulness. And other texts discuss the practical usefulness of many tests in many situations.

A subordinate part of this problem is the applicability of testing to supervisory or executive positions. Ordinarily, in organizations these positions are the most important both from a salary point of view as well as from an organizational standpoint. Testing for positions on such levels has not achieved as much success as it has for clerical and manual positions. It is quite possible that in the next few years, with the attention that this problem is now receiving, adequate tests for these positions will be developed. The present complication in this phase of the development of testing lies in the determination of exactly what characteristics produce good executive material.

C. Use of tests by smaller firms.

The use of tests by the smaller firms, those with less than 1,000 employees, will probably never be completely solved. Most testing requires a highly trained administrator and the size of these firms precludes the possibility of their having such a person. The amount of training necessary is open to some question. Many writers believe that a psychologist with clerical experience should be put in charge of all testing programs. Other writers have pointed out that any reasonably intelli-

---

[2]Bingham, W. V., Aptitudes and Aptitude Testing, Harper and Bros., New York, 1918.

gent personnel man with some specialized training in testing techniques and evaluation can administer most of the tests. For this reason many universities are offering specialized courses and training either through their regular curriculum or through institutes and seminars to encourage personnel men to use testing and to teach them the information they need. It should be possible over a period of years to supply industries having less than 1,000 employees with men who have had or can obtain this training and can also perform other duties for the company. Also of help in meeting the problem of the small concern are the testing bureaus of schools, the aid of local psychologists, and the testing work done by the state employment services.

D. The use of personality testing in industries.

This field of testing is being subjected to a great amount of research and study. In most instances, except for a few of the rather insignificant tests, this type of test has been administered only by trained psychologists. Inasmuch as this is a more difficult field of testing both from the standpoint of administering the tests and of interpreting test results, its usage has been confined to concerns where psychologists are available or where the concern has easy reference to outside advice by trained men. It is doubtful if the interpretation of such tests will ever be simplified to the point where the ordinary company without a man highly trained in psychology will be able to make satisfactory use of them. It may be that some concerns will find it possible to select a member of their present personnel organization and send him to a university for this type of training. Many universities have inaugurated

courses of study to teach such techniques as the Rorshach method to industrial representatives and perhaps other personality testing techniques will receive the same treatment in the future. However, for the ordinary firm there is still another problem, and that is the determination of what personality characteristics are needed on various jobs. Here again the results of other investigators may prove helpful but generally the small concern with its lack of clear demarcation among the various types of jobs will always have difficulty in determining what part temperament plays in each job.

Summation

It is understood that the final test of ability is an employee's actual efficiency on the job. There are, however, certain individual qualities, which influence this efficiency and which can be measured before an employee is on the job. Inasmuch as human judgment is subject to many errors, the use of tests in aiding the selection and placement of employees offers advantages. In many companies today there are employees who are doing unsatisfactory work in the positions to which they have been assigned. This situation results in inferior work and high turnover. Any improvement which can be developed to select employees as well as to place them will be financially worth while.

While tests have repeatedly shown high correlation with productivity, wages, turnover, and adjustment to work, anyone using or planning to use them should recognize their limitations and weaknesses. They should

realize that such tests do not constitute a perfect instrument for relating individual traits to occupational requirements. A great danger lies in test usage by persons who expect too much of them. Users of tests should realize that when statements are made to the effect that they will improve employment techniques, it does not mean necessarily that tests will select the best possible employees. In many cases a test will only rule out applicants whose chances of success are poor. Even this, however, is an improvement on many employment practices.

Current and prospective users of tests can be guided by the following principles:

1. Job requirements should be carefully analyzed and studied before making use of any tests.

2. On the basis of this analysis, tests should be chosen which measure those skills actually called for by the job.

3. After administering the tests to experimental groups, the results should be compared with some carefully selected criteria.

4. The user should keep in mind that testing is simply a scientific supplement to other employment and placement procedures and should not expect startling results in the quality of employees selected or placed.

BIBLIOGRAPHY

Beaumont, Henry, The Psychology of Personnel, Longmans, Green, and Co., New York, 1945.

Bingham, W. V., Aptitudes and Aptitude Testing, Harper and Bros., New York, 1918.

Burtt, H. E., Principles of Employment Psychology, Houghton Miflin Co., New York, 1926.

Cleeton, Glen U., and Mason, Charles W., Executive Ability - In Discovery and Development, the Antioch Press, Yellow Springs, Ohio, 1946.

Cook, David W., "Psychology Challenges Industry," Personnel Series No. 107, American Management Association, 1947.

Drake, C. A., and Oleen, H. D., "The Technique of Testing," Factory Management and Maintenance, March 1938.

Edwards, A. L., Statistical Analysis for Students in Psychology and Education, Rhinehart and Co., Inc., New York, 1947.

Garfiel, E., "The Measurement of Motor Ability," Archives of Psychology, Vol. 9, 1923.

Ghiselli, E. E., and Brown, C. W., Personnel and Industrial Psychology, McGraw Hill Book Co., New York, 1948.

Guilford, J. P., Psychometric Methods, McGraw Hill Book Co., New York, 1936.

Hay, Edward N., "Tests in Industry. Practical Illustration," Personnel Journal, Vol. 20, May 1941.

Hunt, Thelma, Measurement in Psychology, Prentice Hall, Inc., 1936.

"Industrial Vision Institute," reprint from Industrial Medicine, April, 1945.

Jucius, Michael J., Personnel Management, Richard D. Irwin, Inc., Chicago, 1947.

Kornhauser, A., "Are Intelligence Tests Worth While," American Magazine, Vol. 140, July 1945.

Laird, Donald A., The Psychology of Selecting Employees, McGraw Hill Book Co., New York, 1937.

Lawshe, C. H., Principles of Personnel Testing, McGraw Hill Book Co., New York, 1948.

Leake, M. Martin and Smith, Thyra, The Scientific Selection and Training of Workers in Industry, Isaac Pitman and Sons, Ltd., London, 1932.

Lishan, John M., "The Use of Tests in American Industry: A Survey," Personnel, January 1948.

Maier, Norman, Psychology in Industry, Houghton Miflin Co., New York, 1936.

Management Review, American Management Association, New York, September, 1948.

Martin, Howard G., "The Construction of the Guilford-Martin Inventory of Factors GAMIN," Journal of Applied Psychology, Vol. 29, 1945.

Martin, Howard G., "Locating the Troublemaker with the Guilford-Martin Personnel Inventory," Journal of Applied Psychology, Vol. 28, 1944.

National Industrial Conference Board, Inc., New York, "Experience With Employment Tests," Studies in Personnel Policy, No. 32, 1941.

Newsweek, Vol. XXXII, No. 3, July 19, 1948.

Pigors, Paul, and Myers, Charles A., Personnel Administration, A Point of View and A Method, McGraw Hill Book Co., Inc., New York, 1947.

Poffenberger, A. T., Principles of Applied Psychology, Appleton Century Co., New York, 1942.

Pond, M. A., and Bills, M., "Intelligence and Clerical Jobs, Two Studies of Relation of Test Score to Job Held," Personnel Journal, Vol. XIII, 1933.

Remmers, H. H., and Gage, N. L., Educational Measurement and Evaluation, Harper and Bros., New York, 1943.

Schultz, Richard S., "_____", Personnel Journal, Vol. 25, September 1946.

Scott, W. D., Clothier, R. C., Mathewson, S. B., Spriegel, W. R., Personnel Management, McGraw Hill Cook Co., New York, 1941.

Shellow, Sadie M., "An Intelligence Test for Stenographers," Journal of Personnel Research, Vol. V, 1926.

Stead, W. H., Shartle, C. L., and others, Occupational Counseling Techniques, American Book Company, 1940.

Steiner, Matilda E., "The Use of the Rorschach Method in Industry," Rorschach Research Exchange, Rorschach Institute, Vol. XI, No. 1, 1947.

Strong, E. K., Manual for Vocational Interest Blank for Men, Stanford University Press, 1938.

Telford, Fred and Moss, F. A., "Suggested Tests for Patrolmen," Public Personnel Studies, 1924, Vol. 2.

Tiffin, Joseph, Industrial Psychology, Prentice-Hall, Inc., 1946.

Woodworth, R. S., Dynamic Psychology, Columbia University Press, New York, 1918.

Wyatt, Frederick, "The Interpretation of the Thematic Apperception Test," Rorschach Research Exchange, Rorschach Institute, Vol. XI, No. 1, 1947.

Yoder, Dale, Personnel Management and Industrial Relations, Prentice-Hall, Inc., 1946.

VITA

The author was born in Greensboro, N. C. in 1918 and attended public schools in North Carolina, graduating from Chapel Hill High School in 1935. He attended the University of North Carolina and graduated with a B. S. degree in Economics in 1939.

He was employed by the Federal Reserve Bank of Richmond from 1939 until he received a commission in the United States Naval Reserves in 1942. He left the Navy as a Lieutenant, senior grade, in 1946 and returned to the Federal Reserve Bank. At the bank he held a position in the accounting department until he became a member of the job analysis committee, which drafted job descriptions for all of the jobs in the bank. In July 1947 he was transferred to the Charlotte Branch of the Federal Reserve Bank as Personnel Manager, a position which he holds at the present time.