Errors in placement are legendary in military circles. Stories are told of highly trained technicians assigned to cleaning latrines, of mathematicians forced to be typists, or of dull and untrained people given responsible jobs. Errors in placement are dramatic in the military services because there a man improperly placed cannot be fired for incompetence, nor can he quit.

Civilian organizations face the same basic problems even though they are not so well dramatized. In civilian life, too, individuals are frequently placed on jobs for which they have little or no talent or in which the talents they have are wasted. People rejected by an organization with one set of hiring standards are hired by another—and they succeed. People who are hired with high hopes frequently fail. Such errors result in waste of manpower, expense to business, and frustration to the individual.

Personnel testing has one specific objective: to contribute to the increasingly effective use of manpower within an organization. At all manpower levels, the effectiveness of utilization must be gauged by one or both of two kinds of measures: (1) how well people do in the performance of their jobs, and (2) how much personal satisfaction each individual finds in the performance of his job (Katzell, 1957). These two measures are quite general. They suggest many more specific questions to which, through testing, answers can be predicted for individual applicants: How fast will he work? How much skill will he develop during the training programs? How often will his work be marred by errors?
How consistent will his production be? What kinds of attitudes will he develop toward his work? How well will he be motivated to produce? How long will he stay on the job? How regular will he be in attendance? By permitting prediction of answers to such questions as these, tests can contribute significantly to the improved selection and placement of applicants.

It should be made clear at the outset that personnel testing does not provide completely accurate prediction of performance on the job. In the first place, variations from one worker to another are due to many causes. These may include attitude toward supervisors, working conditions, or training, as well as the possession of required aptitudes and motives at the time of hire. There are many personnel activities besides selection and placement that try to improve the use of manpower: industrial engineering; wage and salary administration; or attitude surveys, suggestion systems, and other techniques for taking the industrial relations pulse.

Beyond these necessary limitations, accuracy is further limited by the fact that prediction must be based upon imperfect measurement. Measures of satisfaction, of production, of aptitudes, and of motives—all such measures are subject to considerable error. In short, it is unrealistic to expect any selection program to be perfect. Nevertheless, it is both realistic and essential that efforts continually be made to improve employment office prediction.

Such efforts can be based upon psychometrics, defined as the application of mathematics and statistics to psychological data (English & English, 1958). The psychometrist is the specialist in scientific psychology who works on problems of measurement; most of his work has been done with psychological tests. The test specialist is a research worker: he uses available tests and develops other measures as he needs them. He uses, and in some situations develops, mathematical principles and procedures for making predictions of future behavior from measures available at the time of employment.

The job of the test specialist is to provide a scientific basis for selection and placement. This he does through psychometric analyses, by developing, and determining the validity of, generalizations about relationships between characteristics of job applicants and their subsequent behavior at work. These generalizations, stated mathematically, enable prediction within known limits of error. Prediction of future behavior, whether mathematical or merely implicit and subjective, forms the basis on which an employment interviewer or supervisor makes a decision about an individual applicant: to reject him, to hire him, to reject him for one job but consider him further for another, or to place his name on a waiting list.

A Need for Caution

Psychological testing is finding ever-increasing use in American business. This is fine. Competent use of tests and of test techniques applied to other parts of the employment process can help industry take long strides toward the more effective utilization of manpower.

Psychological testing is also finding ever-increasing abuse in American business. Testing programs are installed merely because they are stylish. Tests are chosen because of catchy names or clever promotion, without considering what they might be measuring, if anything. Tests are given, and applicants are selected, without any effort being made to discover the facts about what the tests will do.

Many times a person who genuinely understands mental measurement, and who also understands the nature of job performance and job satisfaction, can pick out a really useful test or test battery on his first try. Even the experts, however, can guess wrong—and they often do. The real experts, therefore, are very much concerned with finding out for sure whether a test will actually help make good predictions of on-the-job behavior. This is called validation, or testing the test (Tiffin & McCormick, 1958).

The “mental measurements differs” are not concerned with validation. They pick tests on theory or on looks or because of someone’s testimonial and are satisfied as soon as the choice is made. Since most tests are fairly good in some respects, and since almost anything is a little bit related to almost anything else, the tests chosen in this way sometimes really do provide a better work force. Most of the time they have no relationship either to performance or to satisfaction, they do no particular good, but neither do they do much real harm. With appalling frequency, however, the incompetent testing program actually results in the selection of applicants who are least satisfactory.

A Point of View

Individuals are of primary importance to a free society, where each man has a fundamental right to a job for which he is qualified—as long as he fulfills its demands. Some men in trades or professions may work independently, but most will find employment by joining some organization.

In a free society, organizations are formed when people voluntarily join together to reach common goals. They earn this privilege by producing goods or services of value to at least a segment of society. A legitimate organization deserves, therefore, to prosper and grow according to its contribution to society (Eells & Walton, 1961).

As it grows, it adds new members for certain functions and responsibilities. The interests of society, as well as of the organization, are served best when the new members are fully qualified for their jobs. The interests of society are compromised when a fully qualified person is denied a place given to one less qualified. If many serious errors of selection are made, moreover, the organization may fail—with resulting human and economic waste.

The decision to apply for work is made by the individual, but the decision to hire or not is made by the organization. This text, therefore, clearly is written from the viewpoint of the employing organization. This is not, however, at all inconsistent with a regard for the integrity of the individual in a free society. The point of view of this text is that it is both wasteful and immoral to deny qualified people desirable and available employment for invalid reasons, including some whims known as “company policies.” Tests and test technology used in employment procedures can markedly reduce the frequency and degree of wasteful selection errors. The employing organization has a responsibility to itself and to the society that supports it to be sure that it uses these tools competently and wisely.
The Measurement of Individual Differences

Plato proposed a series of tests for the selection of the guardians of his ideal republic. Today's office manager gives practice dictation to an applicant for a stenographic job. From Plato to the office manager, personnel selection has had its basis in the fact that people are different.

The fact of individual differences is inescapable. We see it in such obvious traits as height, weight, or hair color. We are well aware of it in less obvious traits such as personality, intelligence, or achievement. As we meet people we recognize these differences: some are dark, others light; some are heavy, others thin; some are tall, others short; some are pleasant to be with, others disagreeable; some are bright, others dull; some do superior work, others do poor work.

Such descriptions are, of course, misleading. They imply that each trait can be described adequately in terms of only two categories: tall-short, pleasant-disagreeable, superior-inferior, and so on. Such a classification may be convenient and, for some purposes, even adequate. But, certainly, more discriminating classification of people is possible. One might hesitate, for example, to classify some people as either tall or short. A third category might be better, that is, medium. Moreover, within any category there are still differences that could be described. Some tall people, for example, are taller than others.

All of this implies measurement. Measurement which stops with only two or three classes is, however, extremely gross; it will not provide for very fine discriminations. More explicit measurement, by providing many small categories and classes, makes possible finer discriminations. Sometimes the choice between available applicants is so obvious that fine discriminations are not needed—differences are great, and a classification into two or three categories is fine enough. More often, the choice is not so obvious and finer discriminations are needed.

Test scores are typically scattered throughout a large enough range to make finer discriminations possible. Examples of two test distributions are shown in Figure 1-1. Scores on a verbal skill test, for a group of 123 female applicants for office work, ranged from a low of 26 to a high score of 47 on a 50-point test; their scores ranged from 4 to 15 points on an arithmetic test where 15 points was the maximum score. One can question whether the arithmetic test was discriminating fully (although it was doubtless satisfactory for practical purposes); there is no way of knowing which of the eight who made the top score was best in arithmetic.

Test specialists think of personal attributes as the dimensions along which people differ. Psychometric descriptions of less obvious differences between people require (1) identification and definition of the dimension to be measured and (2) an appropriate scale of measurement. If a dimension is well understood, there are usually alternative ways of measuring that are equally satisfactory. Height is a dimension typically measured in inches, although it can be measured in centimeters or "hands." Weight is another physical dimension, typically measured in pounds avoirdupois. Hair color is not typically measured at all, but it could be measured in terms of the wavelength of light reflected from it.

Height, weight, hair color, and other physical dimensions are relatively easy to measure. In the first place, they can be measured independently. Although height and weight may be correlated as natural phenomena, they are independent dimensions in that each can be measured by separate scales, uninfluenced by the other. A scale of measurement that describes variations in one dimension along a straightforward numerical sequence without being influenced by any other characteristic is said to be unidimensional.

Unidimensional measurement of psychological characteristics is a psychometric goal that has not yet been satisfactorily attained. The scatter plot shown in Table 1-1 shows almost no correlation between verbal skill and arithmetic skill; these are considered by most authorities to be two independent dimensions—as independent as height and hair color, perhaps. But they cannot be measured independently; instructions for the arithmetic test are verbal, and most subjects tend to verbalize the problems.

Personality, ability, and job performance—the characteristics important to the personnel testing specialist—are hard to measure because the dimensions themselves are not clearly understood. Research being done to try to define more precisely the basic independent dimensions will be described in Chapters 3 and 4. The search can best be characterized as moving by successive approximations; much research remains to be done, but useful work can be...
accomplished with the present knowledge. Definitions with varying degrees of independence and clarity have been offered for many psychological dimensions, and for each there are available scales of measurement, such as intelligence tests, behavior ratings, or production counts. Some of these are quite good. The competent test specialist is never fully satisfied, continually seeking improved measurement as a basis for improved prediction.

Selection

The personnel selection problem can be defined as choosing, from a number of available applicants, a smaller number to be hired for a given job. Basic to this process is the assumption that variation along some trait dimension (i.e., some measurable characteristic of the applicant) is related to variation along a job performance dimension. When the nature of that relationship is determined (or more typically, guessed at), selection can be based on descriptions of the applicants available. The description can be done through measurement, but even when it is not, the same pattern is used. This description of the process of selection is not limited to careful scientific research; it is implicit in even the most unscientific casual procedure.

For example, applicants will differ in the strength of handclasp. A sales manager may believe that there is a positive, direct relationship between strength of handclasp and effectiveness as a salesman. Using a simple three-category rating as a measure (of which he may be unaware), ranging from "weak" through "OK" to "firm," he may select only those applicants whose handclasp he has rated "firm." If he does something like this, his selection is based upon implied measurement in accordance with a hypothesis he may never recognize as one.

The test specialist prefers to work with explicit measures, such as psychological tests, of the traits he believes to be related to performance. For example, it is usually accepted that intelligence is related to performance in almost any kind of training activity. The test specialist, however, is not willing to accept this without using a specific measure of intelligence, such as the Adaptability Test, and checking within the specific situation the relationship of intelligence to performance. The Adaptability Test was used as a screening device for a premanagement training course in the Seattle area by the Boeing Airplane Company, with results shown in Figure 1-2. It will be seen that clear evidence exists that the more intelligent people, as intelligence is defined by a specific test, actually do tend to perform better in that training course. This is evidence, not mere supposition; the difference is crucial to personnel testing.

Fig. 1-2. The percentage of premanagement trainees rated superior in training-course performance in each of four score ranges on the Adaptability Test.

of the applicants available. The description can be done through measurement, but even when it is not, the same pattern is used. This description of the process of selection is not limited to careful scientific research; it is implicit in even the most unscientific casual procedure.

For example, applicants will differ in the strength of handclasp. A sales manager may believe that there is a positive, direct relationship between strength of handclasp and effectiveness as a salesman. Using a simple three-category rating as a measure (of which he may be unaware), ranging from "weak" through "OK" to "firm," he may select only those applicants whose handclasp he has rated "firm." If he does something like this, his selection is based upon implied measurement in accordance with a hypothesis he may never recognize as one.

The test specialist prefers to work with explicit measures, such as psychological tests, of the traits he believes to be related to performance. For example, it is usually accepted that intelligence is related to performance in almost any kind of training activity. The test specialist, however, is not willing to accept this without using a specific measure of intelligence, such as the Adaptability Test, and checking within the specific situation the relationship of intelligence to performance. The Adaptability Test was used as a screening device for a premanagement training course in the Seattle area by the Boeing Airplane Company, with results shown in Figure 1-2. It will be seen that clear evidence exists that the more intelligent people, as intelligence is defined by a specific test, actually do tend to perform better in that training course. This is evidence, not mere supposition; the difference is crucial to personnel testing.

of the applicants available. The description can be done through measurement, but even when it is not, the same pattern is used. This description of the process of selection is not limited to careful scientific research; it is implicit in even the most unscientific casual procedure.

For example, applicants will differ in the strength of handclasp. A sales manager may believe that there is a positive, direct relationship between strength of handclasp and effectiveness as a salesman. Using a simple three-category rating as a measure (of which he may be unaware), ranging from "weak" through "OK" to "firm," he may select only those applicants whose handclasp he has rated "firm." If he does something like this, his selection is based upon implied measurement in accordance with a hypothesis he may never recognize as one.

The test specialist prefers to work with explicit measures, such as psychological tests, of the traits he believes to be related to performance. For example, it is usually accepted that intelligence is related to performance in almost any kind of training activity. The test specialist, however, is not willing to accept this without using a specific measure of intelligence, such as the Adaptability Test, and checking within the specific situation the relationship of intelligence to performance. The Adaptability Test was used as a screening device for a premanagement training course in the Seattle area by the Boeing Airplane Company, with results shown in Figure 1-2. It will be seen that clear evidence exists that the more intelligent people, as intelligence is defined by a specific test, actually do tend to perform better in that training course. This is evidence, not mere supposition; the difference is crucial to personnel testing.

The number bracketed after the name of any published test refers to the entry number for that test in Buros's Tests in Print (1961). This useful volume is a complete index of psychological tests. The Adaptability Test is described in entry 719 by indicating the population for which it is intended (job applicants), its dates of publication (1942; revised 1954), authors and publisher (Tiffin & Lawshe, Science Research Associates, Inc.), and a listing of the volumes of the Mental Measurements Yearbooks (Buros, 1958, 1941, 1949, 1953, 1959), where bibliographies and critical reviews may be found.

Private communication from J. R. Niven; used by permission.
With such evidence, specific rules can be laid down for using the test in the actual selection process. The general procedure recommended for use at Boeing puts testing at the end of the employment procedure. Thus, those who are tested are considered about equally qualified on other counts. Those with the highest scores are the ones actually chosen for training.

**Placement**

So far we have described *interpersonal* differences, that is, those along a particular dimension by which one person can be said to be different from someone else. Selection is based upon these interpersonal differences. Personnel testing should also be concerned with what might be called *intrapersonal* differences: differences within the individual in the extent to which he possesses or exhibits certain characteristics.

Using a standard scale of measurement for each of several dimensions, we might find that a given person is about average in height, weight, or hair color, and even in most personality or intellectual traits, yet weaker than most in verbal abilities and stronger than most when working with numbers or spatial problems. The terms "weak" and "strong" derive their meaning, of course, from comparison with other people. But they also have particular significance when used to show relationship to other traits within the same person. For example, the 50th percentile in a distribution of measures indicates an average amount of the trait being measured, in terms of interpersonal differences. The 50th percentile might indicate the strong point, however, of an individual who tends on most other dimensions to score at about the 15th percentile, or a serious weakness for the individual who typically scores on ability tests at the 90th percentile or better. Legend is full of mathematicians who could not spell, inventive geniuses who flunked out of school, and beautiful but dumb blondes. Each person has some strong points and some weaknesses.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Score Expressed in Percentile Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal skill</td>
<td>72</td>
</tr>
<tr>
<td>Arithmetic skill</td>
<td>41</td>
</tr>
<tr>
<td>Clerical skill</td>
<td>68</td>
</tr>
<tr>
<td>Reasoning skill</td>
<td>72</td>
</tr>
<tr>
<td>Typing skill</td>
<td>49</td>
</tr>
</tbody>
</table>

*Fig. 1-3* Test profiles for three female applicants scoring alike on a measure of verbal skill.

**Introduction to Personnel Testing**

This is shown graphically in Figure 1-3, showing the scores on a battery of tests of 3 of the female applicants scoring 41 on the verbal skill test in Table 1-1. Of the 10 applicants with that score, the ones scoring the lowest, the highest, and in the middle of the arithmetic distribution were chosen for the illustration; their scores, in percentile ranks, are shown for five tests. The verbal skill test is heavily weighted on spelling ability, so it is probably not unidimensional. It is best interpreted as measuring both intellectual ability and skill needed in office work. The arithmetic tests and reasoning tests are more intellectual, while the clerical and typing tests are more specifically related to office work.

Applicant C does better on intellectual abilities (arithmetic and reasoning tests), but not so well on office tasks (clerical and typing tests). Applicant B shows a similar pattern, although the ability level is lower on every test except typing. Applicant A shows a quite different pattern; she is weaker on the intellectual tests but stronger on those testing routine office skills. If all three girls are hired, it might be well to place B and C on jobs calling for independent thinking and problem solving, and to place A on a more routine, closely supervised job.

Such intrapersonal differences can be important in selection, and they form the basis of the *placement* process, or classification. General textbooks in industrial psychology often use the expression "selection and placement" as if it should be hyphenated. This frequent pairing of the two words suggests to the uninformed that they are, if not synonymous, at least inseparable processes. Unfortunately, this is not true. While modern psychology has made much real progress in improving personnel selection, much remains to be done in solving the problem of personnel placement.

Selection, in its simplest form, is concerned with choosing one applicant among many for a given job. Placement, on the other hand, is concerned with choosing from a number of jobs the one job best suited to the strengths and weaknesses of a particular applicant. In placement, the decision to hire has already been made; the applicant is going to have a place on the payroll. Where that place will be is the question yet to be answered. It must be answered on the basis of what we are calling here intrapersonal differences. On what jobs will the individual's strengths be put to best use? On what jobs will his weaknesses not interfere? In short, on what job will this individual make his best contribution to the employing organization and to himself? It is regrettable that many companies do not include these questions in their employment procedures. All too often an applicant is considered only for a single job or job family and is hired or rejected without regard for other places in the organization where he might perhaps make a greater contribution. Such a one-track approach in employment policy is hardly the road to maximum utilization of manpower.

4 These data have been taken from a situation in which privately constructed tests were used. These interpretations should not be considered as applying to all such tests.

5 When the arithmetic of the classification problem is more fully considered in Chap. 6, it will be apparent that the problem is really more complex; optimum placement must also, of course, consider the needs of the organization.
The Employment Process

Testing is only one small part of the total employment process, although there are important interrelationships between testing and other aspects. A typical pattern of employment activity is diagrammed in Figure 1.4 (Uhrbrock, 1938). Although details of the procedure may vary from one organization to another, the pattern shown here is typical. Perhaps the most important feature of the process is that the decision rests essentially upon rejection; if an applicant is not rejected at some point in the succession of hurdles, then he is hired. While this may fall short of an idealistic statement of the selection process, it is at least a realistic one.

Recruiting

Recruitment gets applicants into the employment office. Some recruiting policies include (1) promoting or transferring employees from within, (2) making a special effort to hire the handicapped or members of minority groups, and (3) advertising or listing job opportunities with specialized employment agencies. Less obvious but equally important recruitment is based on company reputation and other less tangible forces within the community. Stromberg (1948) found, for example, that one result of a testing program is an improvement in the caliber of applicants coming to the employment office. He suggested that the use of tests establishes, as part of the company reputation within the community, the idea that the company seeks and is getting the best people in the labor force.

Recruitment also includes the simple device of keeping the door to the employment office open so that applicants may walk in off the street. The characteristics of the immediate neighborhood do much to determine the kinds of applicants this "recruiting" brings in. Changes in the community (e.g., racial or national composition, socioeconomic status) will be reflected by changes in the applicants coming in for employment.

Recruitment is important to personnel testing since its procedures, explicit or implicit, define the nature of the applicant population upon which research is done. Changes in recruitment policy and activity result in changes in the applicant population. A test that is valid on a sample of the kind of applicant population recruited at one time may prove to be less valid, or perhaps still more valid, as the applicant population changes.

Interviewing

Interviews and interviewing procedures are notoriously fallible. Even so, the interview is the focal point of the employment process. No part of the process, including testing, is foolproof. The interview is needed as a place where subjectivity may deliberately enter the process before decisions are reached.

Interviewing provides for a preliminary screening of the applicants. The obviously unfit applicant should not be encouraged to waste his time, nor should the company want to expend its time and resources in false encouragement. The preliminary interview may result in the immediate rejection of the few obviously unqualified applicants, and for the others it can serve as an orientation tool by which direction is given to the rest of the employment process. If tests follow a screening interview, then the interview (like recruiting) defines the population to be studied by the test specialist.

The influence of the preliminary interview can be pronounced. Depending on the conduct of the interview, an applicant can become tense and frightened or he can become relaxed and confident. He may become more motivated to get the job, or he may approach the tests with an "I don't really know if I want this job or not" attitude. Sound personnel policy dictates that the interviewer try to increase motivation and to eliminate unnecessary stress and fear. Good research technique demands that the influence of the interview be kept as constant as possible for all applicants.

The function of a final interview is to integrate all information gathered about
the applicant from interviews, tests, application blanks, references, etc. Such information is seldom completely consistent; one kind of information may be unfavorable, but this can be counterbalanced by something better about the applicant. The final interview is the point at which there is properly some degree of subjective juggling and sorting of perhaps conflicting impressions so that a final decision to hire or to reject can be made. Interviewer judgments can be quantified and studied as if they were test scores.

Applications and References

In nearly all organizations some sort of application form filled out by the applicant is used to help in assessing the applicant's background and qualifications. The interviewer examines the completed form and makes certain judgments, which are seldom very precise. They are more likely to be simple generalities, such as "this is a good guy" or "this one doesn't look so hot." At a higher level of sophistication the judgment becomes a subjective effort to predict future performance. For example, it may be predicted that "an applicant who is more than thirty years old and has never married will be a neurotic employee with an excessive number of absences." A prediction like this makes enough sense that it can be verified; therefore, application blanks can be quantified and used as if they were tests. Various kinds of hunches that employment interviewers play while looking over an application blank can be, at this level of sophistication, put to an empirical test. If such hunches are right, they provide further information that can help improve selection; if they are shown to be wrong, then the interviewer is warned against them and can stop being misled by irrelevant information.

An individual's background can also be assessed by checking his references. A telephone conversation will often get information far more reliable and accurate than will be obtained in writing, but even written references can be useful if the writer is asked to answer specific questions and is assured that his replies will be confidential.

The typical letter of recommendation, written upon request of the applicant, is often useless. It is likely to be filled with glittering generalities from which no reliable inferences may be drawn. When such a letter is negative, of course, it can be useful simply because this is so rare; the letter that "dams with faint praise" may also provide information to make an employer cautious.

The local credit bureau is a reference source that is frequently overlooked. Credit bureaus usually have rather full information on any applicant who has been known to the community for any length of time. Frequently an applicant who appears promising in the employment office seems less desirable when it is learned that he habitually neglects financial responsibilities.

These references, of course, do not help to choose between otherwise equally good applicants as much as they help in finding out who the most undesirable applicants are and getting them rejected. Rejection is not always based upon predicted job performance; it may also be based upon policy. An applicant, for example, who is considered immoral or one who is physically dirty, may rank high in terms of predicted performance ability, yet be rejected on the basis of a company policy against having immoral or unclean people in the organization.

Testing

Since most of this book is about testing, the comments below are given only to show how tests fit into the total picture of employment activity.

Tests are simply part of the process. They can be a big help when the final personnel decision is reached, but they do not make the decision. An employment manager who relies blindly on tests, who will not under any circumstances temper test results with his own judgment, has abdicated his managerial responsibility. This is not to suggest that tests have no value; in fact, where test scores and other quantitative data strongly suggest a particular decision, it will rarely be wise to go against that decision. Nevertheless, rare events do occur, and the final responsibility for deciding whether a particular case represents the usual thing or the rare event rests with the employment manager, not with the test constructor.

The point in the process at which to test varies from one company to another; there seems to be little objective evidence to indicate that tests are either more or less effective when they are administered early in the process or else. There is, however, one aspect of timing which is most important: If a trial test battery—the collection of tests that have not yet been validated and are not being used for actual selection—is administered at a particular point in the process, then the test battery finally adopted for actual selection should be used at that same point. The reason refers again to the sampling of the applicant population. The applicant population, at the beginning of the process, is a relatively unselected one. At each successive stage, however, there are vague and perhaps poorly defined changes in the nature and homogeneity of that population which may make significant changes in the pattern of test results.

This becomes especially important if the trial test battery has been administered very early and the adopted test battery postponed until the final stages. If each successive step in the process has been valid and related to the same traits involved in the trial test battery, then correlations between scores on the adopted test battery will be markedly lower (because of restriction of range) than those obtained during the validation of the trial test battery.

Personnel Decisions

A decision to hire, not to hire, or to hold on a waiting list is the end product of the employment process. Cronbach and Gleser (1957) show clearly that this is the basic problem in employment and the problem on which testing programs and future research must focus. They point out that the principle stated by Hull (1928) that "the ultimate purpose of testing is to estimate or forecast aptitudes from test scores" has been the foundation of nearly all work on test theory. Cronbach and Gleser prefer to abandon this point of view. They accept the usefulness of accurate estimation, or prediction, but they insist that the ultimate purpose of testing, and of the employment process generally, is simply to arrive at a decision to hire or not to hire.

Their point is well taken; there is certainly a real need for more research on applying decision theories to personnel testing. Nevertheless, this view needs to be tempered somewhat. Although there is seldom any interest in literal pre-
dictions of future performance, the personnel decision to hire or not hire is based upon at least an implicit prediction of the behavior likely to be shown on the job.

Developing a Testing Program

A technically superior testing program can fail. Testing must fit into the particular administrative climate of the organization in which it is applied. Development of a testing program must therefore follow many steps. Some of these are technical, others are administrative. Basically, the steps can be listed as follows:

1. Establish policies and get support from management, supervision, and labor union.
2. Put the program in charge of a qualified person.
3. Identify needs.
4. Select a trial test battery.
5. Validate the tests.
6. Incorporate the procedure into the employment process.
7. Make provision for periodic review.*

Personnel Policies and Testing Programs

Quite often policy simply grows, like Topsy, without much planning or direction or even much real awareness of what policy is. A testing program (or any other) is not likely to be very effective in such situations. If policy makers in an organization consider a program trivial and unworthy of their serious thought, it will be. An adequate policy is essential, both in terms of providing for support and encouragement of a program and in terms of defining its limits.

Testing will not be successful unless top management officials—those with power over budgets, facilities, and personnel assignments—support the program in a well-defined way. Written policy statements are desirable, including definite decisions concerning at least:

1. Methods of gathering data, particularly from present employees
2. Definition of advisory responsibilities of those in charge of testing
3. Establishment of lines of authority over testing programs
4. Procedures for assuring maintenance of ethical standards and professional integrity of personnel handling the program
5. Reporting of individual test results


*This list follows, with some modification, the views of Lawshe (1948). In his presentation, he made a strong plea for competence throughout; he pointed out that it is not enough to be “merely interested in testing.”

INTRODUCTION TO PERSONNEL TESTING

Top management support is essential, but it is not enough. The test specialist must have the support of supervisory personnel down the line. The employment process usually allows the foreman or supervisor to make the final decision. If he does not accept the testing program, then tests simply do not enter into the decision at all. There is no merit in pushing the program through by the sheer weight of higher authority. If selection is done so that it reduces the authority and prestige of the supervisor—if it is done by “going over his head”—he will feel threatened and resentful and not very cooperative in doing the spadework necessary to the program.

In getting support there are two problems to face. One of these is overcoming the resistance of the skeptics. Equally serious, and often worse, is toning down the enthusiasm of the oversold. In either case, the test specialist will approach the happy middle ground much more quickly if he is able to establish a friendly, personal relationship. Many are the legends of industrial psychologists who spent their first few months with their organizations “just getting acquainted.” Unfortunately, success in establishing sound working relationships frequently seems to involve an element of luck. Gaining the support of a confirmed skeptic often hinges on the accuracy of test-based predictions in a single, dramatic case.

Union support is also needed, particularly if present employees are used in validation. Union support can usually be gained where the company has already established a general reputation for fair play and where there is an existing history of cooperation and mutual respect. If there has been, instead, a history of discord, proposals for employee testing will not be very well received; testing may be considered just one more example of management’s efforts to exploit its employees.

Competent Leadership

No organization should attempt a testing program without having it done under the supervision of someone who is specifically trained in the use of psychological tests. Several undesirable results can be expected when inadequately or improperly trained people handle testing programs; the final result, in the long run, is that such programs are dropped. The end of a program usually follows discovery that, at best, it has not done much good, or, at worst, it has resulted in the selection of the very people who should have been rejected.

One illustration may be enough. In a large bakery chain, many drivers were quitting before their productivity had paid the expense of hiring and training them. A so-called “consultant” was called in, who was presumed to be, from his own advertising, familiar with psychological tests. He rode with the drivers, analyzed the job from a “psychological” point of view, and—on the basis of “insight”—identified certain personality characteristics which he felt bakery drivers should have. Up to this point, there can be no serious quarrel with his procedure. Having identified traits on which selection might be based, however, he immediately jumped to a recommendation that the company hire only those persons scoring highest on some designated measures of these traits. The company then put the program into effect under the leadership of a totally untrained person, who went for many years without bothering to evaluate the effectiveness of the tests.
In due course, a qualified psychologist was hired on a full-time basis. He immediately questioned the effectiveness of the tests for reducing turnover—the original problem—and began a careful analysis of the data accumulated over several years. He found a significant relationship between test scores and tenure on the job, but the relationship was negative. In short, this company's selection program had been giving first preference, on the basis of unvalidated tests, to the very people most likely to become dissatisfied with the rigors of the job and quit! Many thousands of dollars were lost because the man in charge did not know how to evaluate tests or even that it was necessary to do so.

Need Analysis

The addition of testing to the selection process is based on the belief that something needs to be improved within the organization and that the testing, through improved selection, can effect that improvement. Obviously, if tests are introduced on this basis, the success of the testing program must be evaluated in terms of the degree to which the needed improvements actually occur. Need analysis is the essentially diagnostic task of finding out what is wrong or what needs to be improved.

Need analysis begins with a consideration of managerial problems, identifying the most serious problem areas within the organization. In part, this can be done by examining objective data; for example, records can identify the departments or jobs where productivity is relatively low or turnover relatively high. Such deviations identify the places where something needs to be done.

That "something," whatever it may be, is the criterion by which the effectiveness of a test or other selection procedure may be evaluated. If the test is good, then changes within the organization, such as increased relative productivity or decreased turnover, are associated with test performance—the criterion can be predicted from test scores. If such prediction is not possible, then the test is of no value in solving that particular problem (although it might be quite useful for something else).

There is another aspect of need analysis, beyond criterion development. Identifying a problem to be solved, or at least reduced, does not automatically suggest that improved selection is needed. Personnel problems may have several different sources. Consider, for example, the broad categories of possible sources of production problems, as shown in Figure 1-5. Aptitude for the job consists of applicant traits. Skill is developed through training—the development of aptitude. Desire may be associated with the whole range of industrial relations practices: supervision, communication, wage and salary systems, and so on. Many production problems (and many personnel problems, too) are rooted in nonpersonal environmental conditions, such as temperature, quality of tools or machinery, and other aspects of job engineering.

A need analysis cannot be considered complete until an effort has been made to determine what needs to be done to solve the problems that have been identified. Is the best approach going to be an effort to reengineer the job and its immediate physical environment, or to change people already on the job, or to improve the program of selection and placement? Clearly, no testing program can hope to solve a problem that springs primarily from inadequate equipment or supervision.

There will always be certain situations where tests can be used to better advantage than in others. Unfortunately, there is no available rule of thumb and no published psychometric device by which need analysis can be carried out objectively. A program of need analysis consists of asking questions of all persons concerned, trying to find out where they feel they could best use help in carrying out their managerial functions. Through interviewing, through examination of records, and through what, unfortunately, boils down to a question of "intuition," "expert judgment," or "insight," the test specialist reaches basic decisions on where and how he can do the most good and the criteria by which he can evaluate his work.

Trial Test Battery

Generally, two tests are better than one. A battery of tests is usually more valid than a single predictor. Moreover, a priori guesses on prospective test validities frequently do not turn out very well. The conclusion is that there ought to be several possible predictors chosen for a preliminary tryout.

The first step in designating a trial test battery is to suggest personal attributes of applicants which may be associated with the criteria identified in need analysis. Once specific traits of aptitude, interests, or temperament have been identified, ways must be found for measuring them. Sometimes this is difficult. It will frequently be necessary to devise a new test, set up a "work-sample" test, score interview or application forms, or use other "homemade" predictors. However, there are an extremely large number of commercially available tests. Some of these are excellent, others are poor; careful judgment on a basis more sophisticated than clever titles is needed. Information about tests is available in the manuals included with tests, written by the test authors, or in publishers' catalogs. Such information may lack the objectivity that might be desired, and any student of psychological testing should become familiar early in his training with Buros's Mental Measurements Yearbooks (1938, 1941, 1949, 1953, 1956). These yearbooks contain descriptions and critical reviews.
of tests by persons not associated with the authors or publishers. Other sources of detailed information include the Handbook of Employee Selection (Doresus & Jones, 1950) and the series of validation studies that appear regularly in outline form in the “Validity Information Exchange” feature of the journals Personnel Psychology and Educational and Psychological Measurement. Beyond these references, the experience of the person in charge of testing and his familiarity with the wide and growing professional literature on testing determines his ability to pick out good trial batteries.

Validation

Validation, the central and essential process of evaluating personnel testing, will receive consideration throughout this book. Two major approaches to validation should be recognized at the outset. One of these, the present-employee method, works with people who have been with the organization for some time. The trial test battery is administered to these people, whose job behavior can already be designated “good” or “poor” according to the established criterion. By this method, the test is valid to the extent that it correlates with criterion measures (or discriminates between “good” and “poor”) available at the time the test is given. This is known as concurrent validity.

The second approach, the follow-up method, takes more time. Tests are administered to applicants and filed away unscored until the applicants who have been hired have established a record. The amount of time to wait varies with different jobs and depends largely upon the amount of time required to gain proficiency. When sufficient criterion data have been collected, the tests are scored and the correlation with the criterion (or discrimination power) computed. This is known as predictive validity and should be stated with clear reference to the length of the interval between the time of hire and the time of evaluation.

The present-employee method is clearly a violation of scientific principles. Present employees constitute a decidedly different population from that composed of job applicants. With personality or interest inventories, personal history data, and other “fakable” measures, this difference is so serious that present-employee validations are virtually worthless. With ability tests, the difference may still be important, but the concurrent validities may occasionally be useful as fast “first guesses” about predictive validities. The assumption is usually made (and, unfortunately, usually unverified) that any difference between them will result in a conservative error; i.e., that natural selection among employees will restrict the range of talent so that concurrent validities will be less than predictive validities. The uncertainty means that no great faith should be placed on concurrent validities until more adequate follow-up studies are completed.

Follow-up research is necessary for any validation. An initial validation study should always be considered “exploratory” research—exploring both variables and methods of prediction. In the language of experimental psychology, the exploratory study is to generate “hypotheses” about the nature of the relationships found and the implications for the way the variables should actually be used in selection. The “hypothesis-testing” research comes next.

In personnel selection research, hypothesis testing consists, not of evaluating the test per se, but of validating a specific way of using the test as an instrument of prediction. This is the “follow-up” research; it should be a longitudinal research study (i.e., a study of predictive rather than of concurrent validity).

Use of Valid Tests

When there is evidence of validity, the tests can be put to work as selection tools in either of two general ways. One method calls for the designation of specific cutting scores: scores below which (or perhaps above which) applicants will not be hired. Organizations that use cutting scores typically use tests rather early in the employment process; in actual practice, the tests often become the sole real basis for personnel decisions in such situations.

This approach may be modified in organizations where rigid adherence to a single cutting score is undesirable or impractical. Fluctuations of personnel requirements and fluctuations in the labor market mean that a cutting score which might be appropriate on one day may be too lenient or too strict on another. Organizations which cannot use a single cutting score may therefore establish certain grades or score ranges, each with its own implications for hiring. Applicants in higher grades are preferred over those scoring in lower categories; those actually selected are those scoring best within a given period of time.

The other approach relies more on the judgment of the person who makes the final decision. It is best illustrated by the use of expectancy tables (of which Figure 1-2 is a simple example) indicating the probability of success associated with various score ranges. The test, under this procedure, is probably a less important characteristic of the selection procedure, being decisive only if “all other things are equal.” Since other things are seldom equal, the probability of success as indicated by test scores may be tempered by the subjective judgment of the interviewer about the relative importance of other bits of information he may be able to gather about the individual applicant.

Periodic Review

It has been emphasized that, because of shifts in the nature of an applicant population, tests once validated simply do not maintain the same validities year after year. It is imperative that there be some systematic provision for regular review to see how test validities are holding up. This means, of course, that a reasonable and effective system of personnel accounting should be maintained so that criterion data are kept up-to-date at all times. Such records should also include systematic information on test scores, transfers, promotions, rate changes, etc. With adequate records, regular assessments of test usefulness can be made easily, and “emergency” reevaluations made with a minimum of delay at practically any time. Only under circumstances of extreme change in applicants or jobs will appreciable delays be necessary in making new validity statements.

Science and Selection

“A fairly common fallacy among businessmen is the belief that the administration of aptitude, proficiency, or personality tests, by itself, constitutes ‘scientific
The act of selection cannot properly be considered scientific. This act is a decision about a unique individual at a unique point in time; uniqueness in this sense is simply not the stuff of science. To think of the act of decision as a professional act is helpful. It can be performed more effectively if it rests on a scientific foundation: a set of generalizations demonstrated to be valid. The professional act, then, will be to determine the usefulness of the valid generalizations, applied to the unique circumstances of an individual case.

The point might be argued that leaving out decision making would be more scientific, i.e., establishing clear cutting scores based on test validations and tolerating no deviation. But this is perhaps the most unscientific use of tests. It denies observations that every test specialist is constantly making: that there are exceptions to the generalizations represented in the statement of validity. The exceptions are his errors of prediction: the people who failed when he predicted success; his predicted failures who were successful. To ignore empirical evidence is most unscientific, and one important bit of empirical evidence accumulated through decades of personnel testing is that many errors in prediction are made. Very little research has been devoted to understanding the reasons for these errors.

Let it be clearly understood at this point that psychological testing has resulted in tremendous savings through reducing errors far below the frequency of error without tests. These savings have been obtained through a traditional approach to personnel testing. In brief outline, this approach consists of (1) the administration of tests to applicants, (2) the subsequent correlation of test scores with measures of performance, and (3) the introduction of tests thus shown to be valid predictors into the selection process. This is technology, however, not science. It is a routine use of tests and procedures that may have been developed through scientific procedures, but it falls short of a scientific ideal.

A Scientific Focus

The purpose of science is to seek invariance, that is, to develop valid generalizations or define the limits within which a generalization is true.

There can be no claim that personnel testing currently rests upon the scientific foundation of valid generalizations. The most important practical rule, stressed throughout this text, of effective testing procedure is that every test or other predictor should be freshly validated for every new situation to which it is applied. One simply cannot assume that a test which is valid in one organization will also be valid in another where the same type of behavior is to be predicted for people doing apparently similar work. A validation study produces a generalization of only limited validity, and even then only in the organization and under the circumstances within which it was derived. Even within the organization, as circumstances change through time, the validity may also change.

INTRODUCTION TO PERSONNEL TESTING

The focus of science in making generalizations is the phenomenon to be explained or predicted. In an experiment, this is typically called the dependent variable—the variable to be explained or predicted in terms of its dependence on something else. In personnel selection, this is known as the criterion—the variable of job performance or of job satisfaction that is to be predicted in terms of something else that can be assessed earlier. The very use of the term "criterion" is enough to suggest that personnel testing has not focused as much attention on the dependent variables as on the predictors. The criterion, typically, is not considered behavior to be explained so much as it is considered behavior that reflects the excellence of the test!

In short, selection will have a firmer scientific basis when more attention is given to those variables now described as "criteria." In Chapter 4, specific attention will be given to this problem, with suggestions for two specific directions for further work: (1) analysis of performance into independent components, and (2) greater emphasis upon the forms or styles of behavior involved in work than upon the economic consequences of that behavior (i.e., greater emphasis on how the salesman sells than upon how much he sells). It seems possible that there is more than one kind of behavior that will result in desired performance. Scientific generalizations may be easier to develop if these behaviors are identified first and the means of predicting them are identified second.

Formulation of Hypotheses

As the scientist studies the behavior he wishes to predict, he formulates certain hypotheses about the factors that may influence that behavior. Some of the hypotheses, within the work situation, relate to traits or past experiences that could be assessed before people begin to work.

The hypothesis is an educated guess about the relationship existing between two or more variables. One of these may, by analogy to experimental design, be considered a dependent variable (the criterion in personnel testing), and the others are the independent variables (tests or other predictors). In experiment personnel testing, the test specialist hypothesizes that certain independent variables will produce systematic variations in the dependent variables. In personnel testing, the test specialist hypothesizes that certain traits, as measured by his tests, are to a significant degree responsible for, or at least associated with, variations in job performance, as measured by his criterion. In other words, he makes guesses about the attributes (and their measures) required for proficiency on a job. For example, the decision to use intelligence tests in a trial test battery for supervisory selection implies the hypothesis that intelligence is related to supervisory effectiveness.

The usefulness of research depends frequently upon the cleverness of the initial hypothesis. "Intelligence" in a hypothesis may not be a very clever suggestion; there are too many kinds of intelligence. The amount of insight into the kinds of traits that might be related to performance may depend upon the amount of information about the classification of the dimensions of individual differences. Chapter 3 considers the problem of defining the dimensions, and the chapters of Part 2 describe various measuring instruments.
Controlled Observation

The hallmark of science is its insistence upon scientific control in observing data. The concept is best related to experimental procedures where control is used to keep the variables not a part of the hypothesis from influencing the results of the experiment.

The concept of control as the elimination of contaminating influences on research results has been carefully included in traditional personnel testing procedures in two ways: The first is the use of thorough standardization of procedure in the administration of psychological tests. As described in Chapter 7, tests are constructed to avoid ambiguities, so that the stimulus content will be unvarying. Rigid time limits may be established for some tests; instructions may be given with specific wording invariably used. The second is in the "purification" of criterion measures. Careful research workers take pains to assure that criterion data are carefully and accurately recorded according to systematic procedures. (In practice, research workers are not always "careful" about the collection of criterion data, using such records as may be conveniently available.) In many studies, criterion data will be refined so that specific sources of contamination are statistically removed or at least minimized.

There is, however, another respect in which experimental control is not usually considered in connection with personnel testing. That is the sense in which the conditions of control serve to define the limitations of the resulting generalizations. The experimental test of a hypothesis is seen as confirming it or not within the conditions of the experiment as these are defined by the nature of the controls employed.

The necessity for such a concept is not unknown in testing. Wallace and Twichell (1949) pointed out that many factors combine to determine performance or turnover or other criteria, and that these factors exercise a limiting influence on the validity of predictors. The competent test specialist, they suggested, will recognize that, if he is to fulfill his major aims, he must investigate the other factors which may affect performance level or turnover and determine their relationships to the selection procedures and the groups from which the selection is made. In short, he will recognize that he probably never is concerned with a test's validity but with a number of validities which may vary for differing sub-groups of the population and under various administrative procedures. (Wallace & Twichell, 1949, p. 279; italics added.)

To demonstrate their point, they determined validities of the Aptitude Index [2042] for two groups of salesmen: those who were hired on a commission basis only and those who were financed by advances or paid a salary during their training. Validities differed for these groups. Moreover, it was further found that validities varied also according to the applicant's statement of his monthly income requirement. In short, the validity of the Aptitude Index needs to be defined in terms of specific administrative procedures in hiring and of the applicant's financial need level.

The application of this concept of control in research has led to interest in procedures variously known as subgroup analysis or the use of "moderator variables"; this will be discussed in Chapter 6 and in Part 3. In the same context, the broader problems of research design are also considered, including the development of models for prediction that might fit the realities of job behavior better than the traditional model of linear correlation.

Replication

Scientific research requires results that can be subjected to further analysis by having someone else replicate the study, i.e., repeat it to test the same hypothesis. Only when the hypothesis can be verified by independent research studies can a generalization be considered essentially valid.

Throughout the text, particularly in Part 3, where practical problems of selection and placement are considered, there will be an emphasis upon repetition of research. One form of replication consists of following through over a period of time to see if changes in the applicant population or in the organization have produced changes that may invalidate a predictor that earlier had done a good job. Generalizations that have been demonstrated repeatedly in situations which might be different are generalizations that provide a more dependable foundation for personnel decisions.

Summary and Implications

Individuals differ. They differ in definable traits, and they differ in the ways in which they perform tasks and in the satisfactions they derive from such performance. The differences that exist from one person to another—the interpersonal individual differences—provide a foundation for employee selection. Differences in the relative strengths of specific traits within individuals, intradimensional differences, provide a basis for placement of employees as well.

Both selection and placement require, for effectiveness in the application of the fact of individual differences, some reasonable degree of refinement in measurement. Psychometric research, over a period of years, has resulted in an abundance of psychological tests, measuring traits of intellect, skill, and personality, and in a well-established set of techniques by which other variables, such as job performance and job satisfaction, can be measured. Although tests as such are only a part of the total employment process, the principles and techniques of testing can be applied to that process with a resulting reduction of errors of selection and placement.

Most personnel testing as it is typically done is more a technology than a scientific foundation for personnel decisions. A firmer scientific base can be established, however, through more careful attention to the problem of the prediction of behavior. Such attention should result in more fruitful hypotheses, tested under conditions more nearly approximating those of controlled experimentation.

A competent testing program, including the application of psychometric principles as well as the use of specific psychological tests, requires competent leadership within a well-defined policy. Although such a program may develop a solid scientific foundation, the use of that foundation remains essentially a matter of professional judgment; it is not wise to expect valid judgments from
those who cannot understand the scientific foundations that exist or who cannot recognize the difference between sheer technology of routine test validation and the science of developing the principles with general validity.

The use of the impressive mathematics of even the most elegant research design in test validation is based on judgment. Success in the prediction of job behavior is not, therefore, limited entirely by the intrinsic relationship between aptitude factors at the time of hire and later job performance. It is limited also by the competence of the people entrusted with the responsibilities for making these fundamental judgments. Adequacy of judgment is developed from knowledge and experience and a willingness to verify the conclusions drawn from basic experience. Any person interested in personnel testing must become thoroughly familiar with tests and testing techniques, with the characteristics of adequate measurement of criteria, and with the characteristics of different kinds of work. Any person actively responsible for testing must refuse to become smug about his experience or his ability to make sound judgments.

An inescapable fact about any set of measurements is that they vary. Measurements differ for many reasons. The causes of variation may be inherent in the nature of the measuring instrument, the vagaries of chance, or the characteristics of the objects or persons being measured—the latter including characteristics one did not intend to measure!

Such variation raises the question of accuracy—a term often, unfortunately, misused in mental measurement. If variations are due solely to individual differences in the trait being measured, perhaps the measurement can be considered wholly accurate. If all the variation comes from causes having nothing to do with the trait being measured, clearly the measurements cannot be considered accurate. However, one should not be too quick to consider them inaccurate, either. They might be merely irrelevant, measuring something else quite well.

**Concepts for Evaluating Measurements**

The frequent question "Is that test accurate?" is almost impossible to answer sensibly; the term "accuracy" is not really an appropriate term for mental measurements. More important are such concepts as reliability and validity. Both relate to the question of error, but in very specific ways.