Psychological Testing in Personnel Selection

Rosemary Amelia Venne

ISBN: 0-88886-152-4
Printed and bound in Canada.

Industrial Relations Centre
Queen's University
Kingston, Ontario
Canada K7L 3N6

Canadian Cataloguing in Publication Data

Venne, Rosemary Amelia, 1956-
Psychological testing in personnel selection

(School of Industrial Relations research essay series; no. 8)
Bibliography: p.
ISBN 0-88886-152-4

1. Employment tests. 2. Psychology, Applied.
I. Queen's University (Kingston, Ont.). Industrial Relations Centre. II. Title. III. Series.

HF5549.5.V46 1987 658.3'1125 C87-093256-X

TABLE OF CONTENTS

FOREWORD

The Industrial Relations Centre is pleased to include this study, Psychological Testing in Personnel Selection, in its publication series School of Industrial Relations Research Essay Series. The series is intended to give wider circulation to selected student research essays, chosen for both their academic merit and their interest to industrial relations practitioners and policy makers.

A substantial research essay is a major requirement of the Master's Program in Industrial Relations at Queen's. The essay may be an evaluation of a policy oriented issue; a limited empirical project; or a critical analysis of theory, policy, or the related literature in a particular area of industrial relations.

The author of the essay, Rosemary Venne, graduated from the School of Industrial Relations in November 1986. She is now teaching in the School of Business Administration, Acadia University in Wolfville, Nova Scotia.

I would like to express my appreciation to the author for granting permission to publish this excellent study.

D.D. Carter, Director
Industrial Relations Centre
and School of Industrial Relations
Queen's University

January, 1987

ABSTRACT

This research paper reviews the subject of psychological testing in personnel selection. The history of employment testing is traced from its beginnings in World War I to current day testing practices. Tests are described in a five category classification: intelligence, aptitude, performance, interest and personality tests. Next the various psychometric properties of tests are discussed: standardization of a test, objectivity, the different kinds of norms and reliability, and the different types of validity. The latter two topics are dealt with in some detail. The recent findings of validity generalization and its implications are considered. A section on decision theory and utility follows a discussion of classical test validation procedures. The advantages of tests are discussed in terms of test characteristics, recent research on productivity increases with valid testing programs, and alternative predictors. The following section covers the limitations of tests in three areas: the misuse of tests; test bias; and ethical concerns regarding privacy. In the concluding comments, it is noted that there are no better predictors than tests, but that tests are only one part of the personnel process. Also, the recent research findings reviewed here have yet to influence the real world of employment testing.

INTRODUCTION

Psychological testing has been part of the personnel selection process for over half of this century; at times hailed as a breakthrough and occasionally at the center of controversy. A psychological test can be defined as a formal measuring device that has been developed through careful research. Individual differences among job applicants provide the basic rationale for selection (Cascio 1982). Psychological tests are one technique in assessing these individual differences in order to provide a successful match between people and jobs.

The terms testing and selection need to be put into perspective. Selection is one part of the personnel function. Selection, which involves the choosing of an applicant for a job, implies acceptance or rejection of that applicant. Before the selection process, a firm will consider what skills are required for the particular position (job analysis), will build up a pool of candidates (recruitment), and then will choose applicants using a number of selection techniques. Just as selection is only one part of the personnel function, so is psychological testing only one part of the selection process. Besides testing, other selection techniques include: interviews, reference checks, and biographical inventories.

Underlying the use of tests in the selection process are several scientific and economic assumptions that provide the basis for using tests (Ability Tests 1982). One scientific justification for using tests in the selection process, relates to the claimed superiority of objective measurement techniques over more subjective methods (e.g., interviews). With respect to economic justification, there is the promise of efficient selection and increased productivity. A large amount of information can be obtained from tests in a relatively short period of time. Increased productivity is expected to result from the optimum match between workers and jobs. Also, Haney (1981) notes that several factors account for the prevalence of testing in our society: the special place of technology in western culture and the fact that we live in a mass society as well as an information society.

Psychological tests are most commonly used in the selection process for entry level jobs but are also used for promotion and placement decisions. Here the emphasis will be on psychological testing in the selection process. This review paper will attempt to cover many aspects of psychological testing in the personnel process, including a short history of personnel testing, how tests are classified, the psychometric properties of tests, how tests are used in the selection process, as well as the advantages and limitations of psychological testing. This review paper will be kept as non-technical as possible.

HISTORY OF EMPLOYMENT/PERSONNEL TESTING

In the early part of this century, Binet developed the first "intelligence test". This was an individual testing instrument used to assess the intelligence of French school children. These individual tests were time consuming to administer and also required skill in their administration. The first widespread use of tests for selection and classification of personnel occurred during the first World War in the United States (Stone and Ruch 1979). Group testing instruments were used to test and assign the large number of recruits to appropriate military training. These group intelligence tests proved invaluable in the classification of the million and a half recruits (with respect to their general intellectual level) during the war (Anastasi 1982).

The two tests, Army Alpha and Army Beta, were paper and pencil tests devised by a group of psychologists headed by Yerkes. These tests were based on an unpublished group intelligence test developed by Otis. Army Alpha, the more widespread of the two tests, was designed for those with reading and writing ability and included questions on arithmetic, definitions, analogies and general information (Murray 1983). Army Beta, a non-language scale not requiring reading or writing, was designed for illiterates or immigrants unfamiliar with English. Included in Army Beta were tests of mazes, block counting, pattern completion and geometrical construction (Murray 1983).

Thus, the war and its requirements provided the impetus for the development of large scale group testing (tests which were quick and easy to administer). Enthusiasm for these group testing instruments, which had proved so helpful during the war, spilled over into the civilian sector as these tests were released for civilian use shortly after the war.

The 1920s saw the tremendous growth of these group testing instruments into many areas, such as the education field. The simplified administration of these tests allowed the application of large scale testing in the schools. Anastasi (1982) notes that "the testing boom" of the 1920s, based on the sometimes "indiscriminate" use of tests, drew sharp public criticism. After all, these group tests were still "technically crude" instruments whose enthusiastic application outran their technical improvements (Anastasi, 1982).

Laird (1937), in the third edition of his book (on the psychology of employee selection), hints at some of the confusion of this period when he notes that several decades ago, psychological tests were unknown whereas "now there are several hundred tests that have been put to every conceivable use". Though Laird (1937) acknowledged that there was some confusion (i.e., he tried to "put some order into this chaos of tests"), he conceived of the growth of tests as more of a cautious scientific growth than as a mushroom growth.

Moore (1939) also makes note of the criticism and scorn that the bandwagon use of psychological tests drew after World War I. In the enthusiasm after the war, a number of businesses rushed to use these tests with their promised shortcuts to the selection process. However, a number of these tests were not found to be much more helpful than the old methods and some disappointed firms dropped the tests. Moore (1937) points out that for the decade of the 1930s, psychologists tried to "heal the wounds" which occurred during the adolescent growth spurt of the 1920s and to "establish the mental test among the technical, scientific tools of value to industry". The use of employment tests was part of a larger movement to "rationalize" industry (Ability Tests 1982). That is, an attempt was made to use objective measures in a scientific" or controlled selection procedure.

In the 1930s, businesses slowly began to pick up the use of tests, but it remained for the needs of the Second World War to demonstrate how useful tests could be and for more technical advances in tests to occur.

Though Moore (1939) and Laird (1937) do express some caution regarding the use of tests, they tend to be overly optimistic and somewhat naïve in their predictions of how "great" tests can be for the firm. Comparisons of the books of Moore (1939) and Laird (1937) with books written a decade later (post World War II) reveal quantum leaps in test knowledge and sophistication. The writings of Ghiselli and

Brown (1949) and Cronbach (1949), represent the advances and developments in psychological testing that occurred mainly during World War II. A glance at the indices alone discloses the scientific advances of the post World War II books. There was also an increase in the sheer number of tests. Ghiselli and Brown (1955) note that thousands of tests have been devised, though not all were in use.

During World War II, psychological testing was on a much larger scale than that which occurred during World War I, with approximately 13 million servicemen tested for military classification (Stone and Ruch 1979). Testing was also more complex during World War II. Anastasi (1982) reports that the scope and variety of tests used by the military underwent a phenomenal increase during the war. For example, test research using factor analysis, led to the construction of special multiple aptitude batteries for military specialists, such as pilots (Anastasi 1982).

Ghiselli and Brown (1949) also remark on the contributions to the development of psychological tests with respect to classification and placement of workers that occurred during World War II: advancement in the development and use of scientific personnel methods is greatest under abnormal conditions. They note that at such times (i.e., World War II) "attention is focused on methodology in order to achieve rapid solutions to the pressing problems of the day". Advances in statistical techniques played a part in test development (e.g., the use of factor analysis in test theory and test development).

Ghiselli and Brown (1949) predicted that the outlook for psychological tests would be their greater application following the successful use of tests by government and industry during World War II. After the war, the large scale use of psychological tests did spread to industry and continued to grow until the mid 1960s (Stone and Ruch 1979).

Along with the growth of testing, another steady development was the centralization of tests (Cronbach 1984). Testing and evaluation was expensive and the merger of commercial publishers of tests into larger units allowed for economies of scale. Cronbach (1984) points out the benefits of centralization: it allowed a steady improvement in the technical aspects of testing with the growth of qualified staff; and it permitted the amassing of information regarding a specific test (e.g., norms). He also mentions that the centralization process is an "enemy of diversity" with only minor improvements in test design occurring as a result.

Surveys over the years point to the increasing use of personnel testing in industry. In a 1939 survey of 2700 firms, 14% had some type of testing program and by 1948, 22% of the 413 firms in another survey reported using tests of some type (Lishan 1948). The use of tests was even higher during World War II. In a 1940 survey, 66% of the 231 firms polled reported that they were using some type of tests (Lishan 1948). A survey conducted by the American Society for Personnel Administration (ASPA) and the Bureau of National Affairs (BNA), on a cross section of American organizations, found that approximately 90% were using tests in 1963. However by 1971, another nation-wide survey of American industries conducted by ASPA and BNA, revealed a decrease in the use of personnel tests. In contrast to the 1963 survey, only 55% of the organizations report using some type of employment tests in 1971 (ASPA-BNA Survey 1971).

The main reason for the decline in the use of tests was the change in United States legislation occurring in the mid 1960s. A major factor was the Civil Rights Act of 1964, particularly the Equal Employment

Opportunity Act - Title VII, which "forbids any employment practices (including personnel testing) which discriminates against members of minority groups and which have no legitimate business purposes" (ASPA-BNA Survey 1971 p.1). With this act, Dunnette and Borman (1979) point out that "selection practices became a matter for public and legal concern" (p.481).

The Equal Employment Opportunity Commission (EEOC) is responsible for implementing and enforcing the provisions of Title VII of the Civil Rights Act. The EEOC, which has jurisdiction over all occupational categories, has set a number of testing guidelines. For instance, the required level of statistical significance for correlation purposes is set at 95% (ASPA-BNA Survey 1971). U.S. federal prohibitions against discrimination in employment have had significant effects on the use of tests, producing a complex set of restrictions on an employer's hiring decisions (Ability Tests 1982).

Jain (1974) remarks on the two major thrusts of the EEOC testing guidelines. First, there is the requirement that there be a "demonstrable relationship" between hiring procedures and critical job success criteria. The second point relates to validation procedures which must not "differentially" reject applicants solely due to their minority group membership. It is up to the employer to prove non-minority discrimination on the latter point. Jain (1974) also notes that a considerable amount of EEOC dealings have been related to the use of employment tests. The basic intent of the EEOC testing guidelines has been upheld by the U.S. Supreme Court rulings (e.g., Griggs v Duke Power Company). Thus, the legislative changes of the mid 1960s have had a substantial impact on the use of employment tests (i.e., Title VII and EEOC testing guidelines).

The use of employment tests in Canada does not seem to be as widespread as in the U.S. Jain (1974) notes that fair employment practice Acts in all Canadian jurisdictions prohibit discrimination in hiring and conditions of employment. Fair employment practices provisions in Canada are contained in the human rights acts of the provinces. Dessler and Duffy (1984) point out that with respect to testing these laws require proof of validity (must show a relationship between test and job success) and proof that any tests used do not unfairly discriminate against any subgroup (minority or non-minority). Also these laws pertain not only to tests but to all screening devices used in the selection process.

The problem of discrimination with respect to employment tests, which the EEOC test guidelines sought to correct, will be dealt with in a later section on the limitations of tests.

Stone and Ruch (1979) attribute the subsequent decline in the use of employment tests to the fact that many organizations were previously using tests that they had not locally validated. Following the legislative changes, many employers discontinued their use of these unvalidated tests due to "pressures or fear of pressure from regulatory agencies" (Stone and Ruch 1979 p.4-136). The 1971 ASPA-BNA survey also mentions that several organizations were dropping unvalidated tests (since some organizations found the required validation process too costly and time consuming or they lacked a feasible way to evaluate the tests).

Thus, the consequences for employment testing have been a decline in test use and a general tightening up of test procedures (i.e., more emphasis on validation). Though the quality of employment testing is thought by some to have improved (with the imposition of the EEOC testing guidelines), the very rigid

nature of these standards has created a situation where "adequate or useful tests are being abandoned or struck down along with the bad" (Ability Test 1982 p.31).

This brief history has traced the use of psychological tests from the early part of this century until now. The present personnel testing situation will be more fully developed in the current test use section. Before discussing the characteristics of psychological tests and the development of testing procedures in industry, a classification of tests needs to be outlined.

CLASSIFICATION OF PSYCHOLOGICAL TESTS

As Cronbach (1984) points out, there are numerous ways to classify tests. Tests can be classified along a number of their dimensions: group vs individually administered, paper and pencil vs performing a specified task; and power (progressively harder test items) vs speed (a timed test).

Most classification schemes group tests by what the tests are designed to measure. Anastasi (1982) in her book on Psychological Testing, arranges tests into three general categories: tests of intellectual levels, tests of separate abilities and personality tests. A five way scheme seems to be the most typical classification (Stone and Ruch 1979, Strauss and Sayles 1972, Flippo 1971). Stone and Ruch (1979) classify tests as: cognitive aptitude tests, psychomotor abilities, job knowledge and work sample, tests of vocational interest, and tests of personality and temperament. Similarly, Strauss and Sayles (1972) sort tests in another five way scheme: performance tests, intelligence tests, aptitude tests, interest tests and personality tests. Here the classification will be based on the latter scheme. Psychomotor tests will be discussed in the aptitude test section.

The five main types of tests to be discussed are 1) general intelligence tests, 2) aptitude tests, 3) performance tests, 4) vocational interest tests, and 5) personality tests.

1) General Intelligence Tests

First, it is necessary to define the term "intelligence test". Anastasi (1982) notes that the term "customarily refers to heterogeneous tests yielding a single global score, such as an I.Q." (p.15). Though intelligence is not a single unitary ability, an intelligence test yields a single score representing a number of abilities.

The term intelligence is a broad one with a number of meanings. A number of early researchers (i.e., Thurstone, Guilford) were interested in identifying the different abilities (using factor analysis) contained under the construct of intelligence. In a multifactor theory, Thurstone postulated about a dozen group factors which he labelled as "primary mental abilities" (e.g., verbal comprehension) (Anastasi 1982). Cronbach (1984) points out that factor analysis, the main technique for studying intellectual performance from 1925 to 1965, is no longer the dominant method in this research area. Anastasi (1982) notes that there has recently (in the late 70s and early 80s) been a rekindling of interest in researching the construct of intelligence.

There are a number of intelligence tests in use, ranging from the time consuming, individually administered type, to the brief, group-administered paper and pencil tests. An example of the former is the

Stanford-Binet Intelligence Scale, which presents a wide variety of tasks and requires skill in its administration and scoring. Examples of the latter include the Wesman Personnel Classification Test and the Wonderlic Personnel Test. These tests are easily administered, brief, and objectively scored group instruments.

2) Aptitude Tests

In contrast to intelligence tests, the term "aptitude test" (or ability test) refers to tests measuring relatively homogeneous and clearly defined segments of ability (Anastasi 1982 p.15). The distinction between the two terms is one of degree and specification (with intelligence viewed as a general trait and aptitude viewed as a specific trait). Guion (1965) refers to special aptitude tests as "specialized" measures of intellectual abilities (e.g., the intellectual factor of perceptual speed is involved in clerical aptitude tests). Stone and Ruch (1979) define cognitive ability or aptitude tests as "those that measure an individual's capacity or latent ability to learn as well as to perform a job that has been learned" (p.4-138).

Early psychologists gradually realized the limited nature of the so called intelligence tests, which mainly reflected verbal skills. They saw the need for more specific and precise ability or aptitude tests. Whereas the main use of intelligence tests was in the education area, the aptitude tests (and later the multiple aptitude batteries) were used mainly for selection, classification and counseling purposes in military and industrial settings. Factor analysis, used in determining the different components of intelligence, was also a great aid in the development of the multiple aptitude batteries, which proved to be so helpful in the selection of military specialists (Anastasi 1982).

There are numerous aptitude tests in a wide variety of skill categories. Also, there are multiple aptitude batteries, which provide a profile of test scores (for a number of abilities). These are used mainly in the education area for counseling and by the military. Anastasi (1982) lists a number of areas in which special aptitudes can be tested: artistic aptitudes, musical aptitudes, mechanical aptitudes, psychomotor skills and clerical aptitudes. The latter three aptitudes are the most common ones tested in the employment selection area.

Tests of mechanical aptitude involve measuring a number of skills, such as psychomotor, perceptual, and spatial skills, as well as mechanical reasoning. An example of a common mechanical aptitude test is the Bennett Mechanical Comprehension Test. This test is a paper and pencil instrument investigating mechanical reasoning. This test functioned well as a predictor of successful pilots during World War II. A number of psychomotor skills, such as manual dexterity, can also be measured by perceptual aptitude tests.

Stone and Ruch (1979) treat the psychomotor abilities area as a separate test section. They note the increasing attention given to the measurement of human strength, coordination and dexterity, due in part to the miniaturization in the electronics industry. In factor analytic research of motor tests, Fleishman (1975) has identified a number of factors in psychomotor functions. These factors include finger dexterity, manual dexterity, wrist-finger speed, and arm-hand steadiness. These psychomotor ability tests generally tend to be apparatus tests rather than the group administered paper and pencil type. Anastasi (1982) notes the custom-made nature of many psychomotor tests, which are designed to simulate the requirements of a specific job. Psychomotor

tests are typically used in the selection of military and industrial personnel. One example of a common psychomotor test is the Crawford Small Parts Dexterity Test, which measures manual and finger dexterity.

The last type of aptitude test to be discussed is perhaps the most common one in industrial settings, the clerical aptitude test. One of the main abilities that the clerical aptitude tests measure is perceptual speed. Stone and Ruch (1979) define perceptual speed as the "ability to perceive quickly and accurately familiar similarities and differences in detail" (p. 4-138). A common example of a clerical aptitude test is the Minnesota Clerical Test, which has two subtests, one in number comparison and another in name comparison. This test emphasizes speed and accuracy while other clerical aptitude tests also stress business information and language usage.

3) Performance Tests

Performance tests, also called job-knowledge or work-sample tests, are usually given to experienced workers. Stone and Ruch (1979) define job-knowledge tests as "tests designed to measure how much the applicant or candidate for promotion already knows about the kinds of work involved for which he/she is being considered" (p.4-140). In the personnel area, performance tests have two functions which Stone and Ruch (1979) describe. First, these tests serve as a check on the applicant's job knowledge. Performance or work-sample tests help to distinguish a skilled worker from potential "trade bluffers" and less skilled workers. The second function is to assess the work knowledge of present workers who are in line for promotion or transfer.

The specific worth of any performance or work-knowledge test depends on the care taken with the job analysis. Before the work-knowledge test is designed, a thorough job analysis needs to be carried out for the chosen job. Stone and Ruch (1979) define a "good work-sample" test as one that has been developed on the basis of careful job analysis and which consists of tasks that are truly representative of the nature of the work for the specific job being tested.

Examples of work-knowledge or performance tests range from a simple timed typing test to a trade test for a machinist (which is usually orally administered), to a written or individually administered job-knowledge test developed for a specific job in a specific firm (known as an in-house test).

Achievement and trade tests are often included in this category. Though this type might also be discussed under the second category (i.e., aptitude tests) its purpose more properly places it under the performance or work-knowledge category, since achievement tests are concerned with what the applicant has accomplished and already knows.

4) Vocational Interest Tests

Cronbach (1984) makes a useful distinction between tests of maximum performance (ability tests) that would best be described by the tests mentioned in categories one, two, and sometimes three, and between tests of typical behaviour that would best be described by the fourth and fifth categories of tests. The latter tests would reflect one's typical performance or behaviour, and are best exemplified by vocational interest tests and personality tests.

Stone and Ruch (1979) define vocational interest tests or interest inventories, as tests that compare a person's interest patterns with the interest patterns of people successfully employed in a specific job. The rationale behind these tests is that if a person shows the same interest patterns as those individuals successful in a given occupation, the chances are high that the person will be satisfied in that occupation (Schultz and Schultz 1986).

As. Stone and Ruch (1979) point out, the occupation areas in which a person shows the most interest are expected to be the same areas where that person is most likely to find job satisfaction (assuming the ability to do the job is present). The expectation is that if an individual has the interest, that person also has the ability for a certain occupation. Since this expectation may or may not be true, certain aptitude or ability tests are often given in conjunction with interest inventories, thus providing a more comprehensive prediction of job performance. Stone and Ruch (1979) point out that interest inventories show better prediction of the criterion of job stability than the criterion of job success. They relate this finding to the interest/ability dichotomy, noting that "interest determines the direction of effort and ability the level of achievement" (p.4-141).

While these interest inventories have found more use in the educational and career counselling areas than in personnel selection, the usefulness of these tests in the latter area should not be overlooked. A consideration of a potential worker's interests is of benefit to the individual as well as to the organization, especially if aptitude testing is carried out along with the vocational interest testing.

Examples of two widely used interest inventories are the Strong Campbell Interest Inventory (SCII) and the Kuder Occupational Interest Survey. The former has a long history and has been extensively researched, yielding good reliability and validity measures (Anastasi 1982). The SCII is composed of a large number of items that deal with a respondent's liking or dislike for a variety of activities, subjects and occupations. This test determines how closely an individual's interests resemble those of people successfully employed in six broad occupation areas.

The Kuder Occupational Interest Survey is another important example of an interest inventory. It has a number of similarities to the SCII. One important difference is its much broader coverage of occupations (127 specific occupational groups, Anastasi 1982). Both of these valid and reliable instruments are group administered and computer scored.

Both Stone and Ruch (1979) and Schultz and Schultz (1986) sound a note of caution regarding the use of the self-report inventory (both interest and personality inventories) in the selection process. The "faking" of responses in these self-report inventories is mentioned as a possible problem when these tests are used in the selection process and the individual test-taker has some motivation to show his/her best side.

5) Personality Tests

Another distinction made with tests of typical behaviour is between those tests that are self-report and those that require standardized observation (Cronbach 1984). Vocational interest tests would fall in the former category, being self-report in nature. Personality tests can be self-report or may require standardized observation. Cascio (1986) makes a similar distinction, dividing personality tests into either objective personality inventories or projective measures, which require standardized observation.

Personality tests are perhaps the most controversial and criticized tests used in the selection process. For many laypeople the term "psychological test" means personality test. Werther, Davis, Schwind, Das and Miner (1985) actually define "psychological test" as tests that measure personality or temperament. In most personnel texts, the term "psychological test" is a general term referring to different types of tests, personality tests being only one of these tests. However, Cascio (1982) does make a distinction between the terms test and inventory. He notes that inventories, which reflect what a person feels, can be falsified (here categories four and five). Tests, which measure what a person knows or can do, cannot be falsified (here categories one, two and three). He astutely points out that public suspicion of testing results from a confusion of the two terms.

Personality tests are designed to measure such characteristics as an individual's emotional states, self-confidence, interpersonal relations, motivation, interests and attitudes (Anastasi 1982). Concern over an individual's personality or how a prospective employee will fit into an organization is an important consideration in the selection decision and appears to be a legitimate worry. As Stone and Ruch (1979) and Beach (1980) point out, causes of job failure stem more often from personality and related job adjustment problems rather than from ability problems. Since personality is considered to be important for successful employee performance, what the personnel manager needs is an objective, reliable and valid instrument that would help predict employee performance. Yet the current test instruments, objective personality tests and projective tests, do not fit the personnel manager's requirement.

The first type, the objective self-report inventory, is the most common personality test. These paper and pencil tests are group administered and objectively scored. As with interest or vocational inventories, these objective tests require respondents to indicate how well each item describes themselves or how much they agree with each item (Schultz and Schultz 1986). These lists of items usually describe certain situations, feelings or activities. Examples of these personality inventories are the Minnesota Multiphasic Personality Inventory (MMPI) and the Manifest Anxiety Scale. The MMPI is a forerunner among these tests and served as the basis for the development of other personality instruments.

One widely used test which is based on the MMPI is the California Psychological Inventory (CPI). The CPI is a good example of a self-report style personality test that was developed for use with normal or nonclinical populations. The CPI consists of a large number of true-false items that yield 16 principal scores (e.g., dominance, self-control, and self-acceptance). Anastasi (1982) names the CPI as one of the best personality inventories currently available. It is technically well developed and has been extensively researched.

The second type of personality test, called projective measures, is individually administered by a qualified professional. With projective tests, a person is presented with a set of vague unstructured stimuli. The meaning the individual "projects" onto the stimuli is expected to reveal latent or unconscious aspects of his/her personality (Anastasi 1982). The test results, which require interpretation by a qualified professional, are subjective and unstandardized. Two well-known projective tests are the Rorschach and the Thematic Apperception Test (TAT).

The Rorschach, commonly known as the inkblot test, involves showing 10 cards of "inkblot" patterns to the test-taker who then describes what he/she sees in these cards. Through the person's responses to these cards, a personality pattern is discerned by a qualified professional. The Thematic Apperception Test

(TAT) is another well known projective test. It consists of a number of ambiguous pictures that show two or more people in different situations. The test-taker makes up a story about what is happening in the pictures. These stories are then analyzed, again by a qualified professional. Both of these tests, the TAT and the Rorschach, involve subjective scoring and are used mainly in the clinical setting.

As mentioned earlier, personality tests are the most controversial tests used in selection and have the most shortcomings. As Stone and Ruch (1979) point out: "Personality tests have long lagged behind aptitude, ability and interest tests in terms of demonstrated usefulness for selection and placement in industry" (p.4-142). The most serious problem is that these tests have frequently been characterized by low reliability and validity (Stone and Ruch 1979). The inventory type personality tests are subject to the same possible "faking" response that is a problem with other self-report inventories. Also respondents may have a tendency to give socially acceptable answers. Anastasi (1982) discusses these two "responses sets" of faking and social desirability at length and notes that these can be a problem with self-report inventories, especially in personality measures. Also, the projective tests are time consuming to administer, rely on subjective interpretation and require the services of a qualified professional.

Stone and Ruch (1979) sum up the current status of personality tests by observing that much more research is needed before these tests are as useful in the industrial setting as they are now in the clinical setting. Many of these tests continue to be useful in counselling and clinical settings, areas for which most of them were developed. Srinivas (1984) also makes the point that tests with a high validity in a clinical setting may not have the same validity in an industrial setting.

CHARACTERISTICS OF PSYCHOLOGICAL TESTS

Having discussed the history of testing and a classification of tests, I will now begin a discussion of the relevant psychometric properties of tests. An attempt will be made to keep this section fairly non-technical. Following this, validity generalization, test validation, and decision theory will also be discussed.

A well-developed psychological test has a number of psychometric properties or characteristics. A proper test should have adequate reliability, validity, objectivity, be standardized and be based on sound norms. The interrelatedness of many of these characteristics will also be discussed.

Standardization

A psychological test can be described as a standardized measure. Anastasi (1982) notes that standardization implies uniformity of procedure in administering the test. Consistency in the conditions and procedures for administering the test attempt to ensure that every test-taker is given the "same" test. The purpose of standardization, according to Cronbach (1984) is to "obtain a measurement comparable to measurements made at other times" (p.56). Thus, with standardization, it is possible to compare the performance of a number of test-takers on the same test, since they have taken the test under near identical circumstances.

Each test needs to have a standardized procedure that is followed to the letter during each and every test administration. Every detail of the testing situation including instructions, time limits, materials used and the test environment, needs to be kept consistent.

There is evidence (Cronbach 1984; Schultz and Schultz 1986) that subtle changes in the test procedure (e.g., change in test room size) can result in a change in individual test performance. Since test results can be altered by changes and carelessness in administration, it is imperative that standardized test conditions are maintained. Of course standardization is the ideal that the test developers attempt to build into a test and that properly trained test administrators strive for. As Schultz and Schultz (1986) point out, a sound test can be rendered useless with careless administration.

Objectivity

The concept of objectivity is related to that of standardization. While standardization refers to uniformity in the test procedure and administration, objectivity refers to consistency in test interpretation and scoring. Thus, an objective test is free from subjective judgment or bias. In order for a test to be considered objective, any person scoring the test should obtain the same results as another person scoring the same test, since the scorer has no subjective input (e.g., bias) into the interpretation or scoring of test results.

Of course, true objectivity is the ideal and tests vary in their degree of this characteristic. Schultz and Schultz (1986) point out that the predominant use of objective tests (over subjective tests) in industry is desirable as it allows for fair assessment of job applicants and equitable comparisons among them. In fact, most tests in the five categories described have a high degree of objectivity, the obvious exception being the subjective personality tests. The scoring process with the Rorschach, for example, is not free from subjective bias and two scorers may obtain very different results.

Norms

The establishment of test norms allows meaningful interpretation of raw test scores. As psychological tests have no inherent standard of pass or fail, an individual's test performance is evaluated by comparing it with scores obtained by others on the same test (Anastasi 1982). Schultz and Schultz (1986) define norms as the "distribution of scores of a large group of people similar in nature to the people being tested" (p.127). The scores of this group, called the standardization group, serve to establish the frame of reference or point of comparison that allows us to compare a person's test score to the test performance of a similar group of people.

The term norm implies average performance rather than any desired level of performance. A person's raw test score is meaningless until evaluated in terms of the standardization group norms. For example, if a job applicant receives a raw score of 78 on a mechanical aptitude test, little is known about the applicant's relative mechanical ability. The score of 78 can be interpreted only when the norms are consulted. If the mean of the test norms is 80 and the standard deviation is 10, the score of 78 can be evaluated as a "typical" performance indicating that the applicant possesses an average mechanical ability.

Without disputing the significance of norms, Cronbach (1984) notes that there are circumstances in employment selection when norms are not of great importance: when there is concern with absolute

performance or when there is an attempt to identify individual differences within a group (e.g., when a manager has to hire the top ten of many candidates).

There are different types of norms with varying degrees of specificity. The term norm generally refers to the broad, published test norms derived from the standardization group. The term local norm refers to norms that an organization has specifically developed for its selection purposes. In between these two types of norms are subgroup norms.

It is often desirable to break down norms into subgroups, that is, to standardize a test on a more narrowly defined population. Subgroup norms are advised when there are score differences between the groups. For example, there could be age and sex subgroups norms for a test. Anastasi (1982) notes that how the test is used determines whether general or more specific norms are most relevant. An organization that uses a particular test in its selection process may find that subgroup norms are more relevant than the general norms for their purposes. For example, with the 1980 Bennett Mechanical Comprehension Test, a tester can compare a candidate's score with reference groups in a dozen industrial settings (Cronbach 1984). For some of these industrial settings, norm tables even separate cases by sex.

An even more specific type of norm is the local norm. Local norms can be defined as the specific norms collected on applicants for a given job by an organization. Stone and Ruch (1979) point out that "the most useful norms will be those developed by a specific employer in a particular community for a specific job" (p.4-136). The local norms that an organization uses may be more relevant for their testing purposes than the broad norms offered by the test publisher. From the ASPA-BNA 1983 survey of American employee selection procedures, it is evident that local or company norms are developed and used more often than the published norms. More than half of the firms use their own company norms as a standard for assessing candidates (ASPA-BNA Survey 1983).

Reliability

A personnel manager wants a test that he/she can rely on and have confidence in. Reliability can be simply defined as consistency of test scores. Stone and Ruch (1979) refer to reliability as the "degree to which people earn the same relative score each time they are measured" (p.4-135).

The reliability of a test is objectively determined before it is released. The test manual should report, among other points, the type of reliability investigated, the method used to determine it, and the number and nature of persons on whom the reliability was checked (e.g., a group similar to the normative group) (Anastasi 1982). The size and representativeness of the sample group is important. As Cascio (1982) points out, the larger the sample and the more the sample resembles your comparison group, the more meaningful the reliability coefficient is.

Such information regarding a test's reliability helps a personnel manager choose a test and determine if such a test will be more, less or equally reliable for the particular group being tested. The personnel manager should ask if his/her group in question is comparable to the standardization group in terms of norms and reliability.

The measures of test reliability will be briefly considered. Different types of tests require different ways of determining reliability. Anastasi (1982) points out that in a broad sense reliability indicates the extent to which individual differences in the test scores are attributable to actual differences in the characteristic being measured (e.g., mechanical aptitude) and the extent to which they are due to chance errors. As no test is perfectly reliable, a number of irrelevant chance errors affect a test's reliability. Different types of reliability have different kinds of errors associated with them.

The first type is called test-retest reliability. This is perhaps the most obvious type that comes to mind when one thinks of reliability. Test-retest simply involves administering the same test to the same group on two different occasions. The two sets of scores are checked for their correlation (expressed as the reliability coefficient). The test manual will describe the group tested and will report the time interval between the tests. Also, Cascio (1982) notes that the test manual should report any relevant intervening experiences of the test-takers (e.g., job or educational experience) since these may affect the retest scores. Srinivas (1984) reports that a two to four week interval is common. According to Anastasi (1982) the test interval should not be immediate nor should it exceed six months. Not unexpectedly, the test-retest correlation decreases as the time interval lengthens.

Cascio (1982) refers to the reliability coefficient derived from test-retest as the coefficient of stability. The question is: how stable are test scores from one administration to another? The error associated with test-retest reliability is referred tows "time-sampling". The error here corresponds to any random fluctuations from one test session to another (Anastasi, 1982). The higher the reliability coefficient (the closer it is to 1.0) the more confidence we place in the test and the less susceptible the test scores are to random changes in the persons tested and in the testing environment.

One problem associated with test-retest reliability is the effect of practice or memory recall that a test-taker has from one session to the next. For example, with only a two week interval a person may recall his/her previous responses. Both Anastasi (1982) and Cascio (1982) note that this type of reliability is not suited to most psychological tests. Only tests that are little affected by repetition are suited to reliability checks with test-retest. Therefore, Anastasi (1982) recommends test-retest be used with psychomotor and sensory tests.

The second type of reliability is equivalent or alternate form reliability. One form of a test is administered, and following a delay period, an alternate form of the test is administered to the same group. The reliability coefficient here reflects both temporal stability and consistency of response to different test forms. Cascio (1982) labels this reliability coefficient as the coefficients of stability and equivalence. Anastasi (1982) lists the two sources of error variance as time sampling (as in test-retest due to the time delay) and content sampling (due to problems in sampling test items from the same domain for both tests). It is possible (due to past experience factors) that a test-taker will find one form of a test easier than another form.

Since alternate form reliability has two sources of error, it is considered a more conservative estimate of reliability than test-retest. It is also considered somewhat more suitable than test-retest for a number of psychological tests. Since the tests are not identical as in test-retest reliability, the problem of item recall is eliminated. However the problem of a possible practice effect may not be eliminated here. Alternate

test forms are unavailable for many tests due to the expense and difficulty in constructing a parallel test form.

The third type is called split-half reliability. It involves splitting the items of one test in half and comparing the scores of the two halves. Since this method involves only one test administration, there is no problem with item recall and practice effects. Anastasi (1982) notes that this type of reliability provides a measure of consistency with regard to content sampling, and is often referred to as a coefficient of internal consistency. The one source of error here is consistency in content sampling.

There are a number of ways to split tests into two halves that are equivalent in difficulty and content. An obvious method is an odd-even split of items. A test split of the first and second halves has the potential problems of warm-up associated with the first half and possible fatigue associated with the second half of the test. The odd-even split balances out these possible problems and their effect on reliability. A random selection of items is also effective.

The two half scores are correlated using a formula that corrects for the attenuated test length (i.e., since we are comparing two halves of one test and not one whole test). This correction is necessary since reliability is positively related to test length (i.e., the longer a test, the more the content domain is sampled). Anastasi (1982) points out that many test manuals report reliability using the Spearman-Brown formula which doubles the halved test length. According to Cascio (1982) this method of determining reliability generally yields the highest reliability coefficient.

A fourth type of reliability, Kuder-Richardson reliability, also measures internal consistency using a single test administration. In contrast to split-half reliability, this last type is based on the consistency of responses to all test items (Anastasi 1982). Kuder-Richardson reliability is a measure of homogeneity or inter-item consistency and as such has two error sources: content sampling and heterogeniety of the behaviour domain sampled. The more homogeneous the behaviour domain is, the higher the inter-item consistency will be. For example, a test of clerical aptitude that measures only aspects of perceptual speed may have a higher inter-item consistency than a test which measures both perceptual speed and business information. The Kuder-Richardson coefficient would be easiest to interpret with a relatively homogeneous test.

Anastasi (1982) notes that reliability coefficients usually exceed .80 and Beach (1980) states that reliability should exceed .85. When the reported reliability coefficient reaches this level, we can have a certain degree of confidence in the test. How we use the scores affects our need for the level of reliability. Aiken (1979) points out that when comparing one individual's score against another (vs comparing groups), a higher level of reliability (greater than .85) is required.

The concept of reliability has been questioned. Brown (1976) points out that the concept of reliability is not the end all of tests but rather a "step on a way to a goal"-the goal being a well constructed test. Cronbach (1984) notes that the error of measurement (which relates to the interpretation of a single score) is a more accurate measure than reliability (which relates to the group's score) and sometimes more useful.

The error of measurement, unlike reliability, allows us to predict the range of fluctuation likely in a single score as a result of chance factors (Anastasi 1982). Put another way, the error or measurement allows one to think of test scores as a range rather than a single point (Cascio 1982). The reliability coefficient is affected by the range of individual differences in the sample as well as the size of the sample, whereas the error of measurement does not have these limitations.

The error of measurement and the reliability coefficient can be thought of as alternate ways of expressing test reliability, with the former being independent of the variability of the group on which it is computed (Anastasi 1982). Also, Anastasi (1982) gives a useful guide for the two measures: "If we want to compare the reliability of different tests, the reliability coefficient is the better measure, to interpret individual scores the error of measurement is more appropriate" (p.127). A number of test publishers now have report forms which allow evaluation of scores in terms of the error of measurement.

Validity

Before discussing the last characteristic of tests, validity, I will briefly discuss the inter-relatedness of these test properties. Stone and Ruch (1979) point out that it is important for personnel managers to understand these test characteristics and the relationships among them.

If one were to nominate one of these test characteristics as the most important, that one would probably be validity. The other characteristics (objectivity, standardization, norms, and reliability) can be seen as a means to an end, the end being a valid measurement instrument. To be reliable tests need to be standardized and objective (high levels of consistency in test administration and scoring, respectively), while norms provide a frame of reference allowing us to interpret a raw score in terms of a defined group (the standardization group). Also, a valid instrument must have reliability, but reliability in and of itself does not ensure validity: that is, reliability is said to be a necessary but not sufficient condition for validity. Statistically speaking, a test's reliability coefficient sets the upper limit on a test's validity, since a test's correlation with an outside criterion cannot be higher than it's own reliability (Cronbach 1984). Therefore, all of the previously discussed test characteristics are necessary before a test is deemed to be a valid measurement instrument.

Validity, can be simply defined as the "degree to which the test actually measures what it purports to measure" (Anastasi 1982 p.27). Cascio (1982) finds this definition too simple as it implies that validity is established once and for all by a single study. He stresses the importance of investigating the many inter-relationships between a test and other variables. The term 'validation' more properly captures the idea of learning as much as possible about the inferences that can be made from a psychological test (Cascio 1982).

The concept of validity is difficult to establish and as with reliability, there are a number of ways to assess it. Dunnette and Borman (1979) contend that the oversimplification of validity into types leads users to place a great emphasis on choosing a type of validity rather than asking "why" they are using the test. First, the user must specify the purpose of the test: what inferences are to be made from the test? This is also the caution that Anastasi (1982) expresses: that validity needs to be established for whatever use the test is being considered.

The term validity is itself misleading. As Guion (1976) points out, the term validity is not singular though we speak of one test's validity, and validity is not a property of the test itself, but more a property of how the test is used (e.g., making predictions from the test's scores). Guion (1976) also cautions that dividing validity into types is a simplification and suggests that proper validation may require the comprehensive investigation of all of the four types of validity.

Here validity will be discussed under two main categories or approaches: 1) criterion-oriented validity, and 2) rational validity. In the first approach, the one most used in employment testing, there is concern with establishing a correlation (validity coefficient) between the test score and some measure of job performance (Schultz and Schultz 1986). Rational validity, the second approach, involves the nature or content of the test, notwithstanding its correlation with an external criterion (Schultz and Schultz 1986). Campbell (1976) also divides the validity taxonomy into two categories: validity for practical decision-making (criterion-related) vs validity for scientific understanding (rational validity).

Criterion-related validity is at the very heart of employment testing. Criterion-related validities are required whenever individual difference measures are used to predict behaviour and are most often used in the selection process (Cascio 1982). According to Guion (1976) research on selection has typically used the correlation coefficient to compare variations in applicant traits (e.g., mechanical aptitude) to variations in subsequent job performance of those hired (e.g., some performance measure of a machinist). Assessing the relationship between predictor (test) and criterion (job performance) has a number of problems associated with it, problems that were also encountered in the use of the correlation coefficient in reliability measures.

As with the reliability coefficient, the validity coefficient can also be affected by sample size. The larger the sample, the more likely a particular study is to find a significant relationship between predictor and criterion scores. Also, the sample must be representative of the persons for whom the test is recommended (Cascio 1982). That is, the validity group should be composed of a sample with the proper age range, education and vocational situation for comparison purposes.

The criterion-validity approach includes two types of validity: predictive and concurrent. With predictive validity, the criterion (some measure of job performance) is assessed after a specified delay. With concurrent validity, the test and criterion measures are taken at the same time. Anastasi (1982) notes that the logical distinction between predictive and concurrent validity is not just time, but the "objective of testing: diagnosis of current states vs prediction of future outcomes" (p.137). The former has a present orientation while the latter is future oriented. For example, we are asking the question: Can Chris perform the machinist job well (concurrent) vs will Chris be able to perform the machinist job well (predictive)?

In the selection situation, predictive validity involves giving the test to all job applicants and then hiring them regardless of their test performance; at a later date, a measure of job performance is obtained and the test scores and the criterion scores are correlated to determine how well the test functioned as a predictor of job performance (Schultz and Schultz 1986). There are a number of factors to consider regarding the time interval between assessment of predictor and criterion measures. Cascio (1982) asks the question: "When has the worker been on the job long enough for us to appraise his or her performance properly?" (p.151). Therefore the amount of training given and job difficulty are two factors to consider. A ball park estimate of six months work experience is given.

Concurrent validity has essentially the same paradigm as predictive validity except that both test and criterion measures are assessed at roughly the same time using current employees. That is, the test measures of current employees are compared with their current job success.

Schultz and Schultz (1986) note that concurrent validity is used more often than predictive validity in industry. Anastasi (1982) points out that concurrent validity is often used as a substitute for predictive validity. Also, evidence from concurrent validity studies has functioned as preliminary data for future predictive validity studies, providing indirect evidence on the predictor-criterion relationship (Cronbach 1984). The use of predictive validity, with its longer time frame is more expensive and sometimes not a feasible practice since if all candidates are hired, some of them may turn out to be poor performers and thus cause a financial drain on the organization.

As the concurrent validity procedures assess only workers currently on the job, Schultz and Schultz (1986) question the ability of the test to discriminate between good and poor performers (since some of the poorer workers are probably no longer on the job, having quit or been demoted or fired). Also with concurrent validity, the effect of job experience on validity is ignored (Cascio 1982). When concurrent validity functions as a substitute for predictive validity, the supposed superiority of the latter over the former has been questioned. One group of researchers (Barrett, Phillips and Alexander 1981) has found that both methods yield similar results with cognitive ability tests. Thus for these type of tests, concurrent validity appears to function well as a substitute for predictive validity.

Both predictive and concurrent validity involve criterion measurement. There are a number of problems with criterion measurement. It is important to keep in mind that "any predictor will be no better than the criterion used to predict it" (Cascio 1982 p.152). This point cannot be emphasized enough.

Precaution must be taken against any possible "criterion contamination" in a validity study. Criterion measures must be gathered independently of the test scores. For example, a supervisor who is rating the job performance of a group of workers must not have any knowledge of their respective test scores (as this knowledge may influence his or her job performance ratings of the workers and result in criterion contamination). To guard against this contamination, test scores must be kept confidential.

Criterion measures must be carefully chosen and assessed. Cronbach (1984) advises that it is more realistic to consider using multiple criteria in place of the single "ultimate criterion", although this may be more time consuming and expensive. Researchers endeavour to obtain reliability in criterion measures, especially in rating measures (measures that are not totally objective in nature). Cronbach (1984) notes that ratings are commonly used as criterion measures as they are cheaper and easier to collect than are measures of work output or observed skill. Though ratings, as a criterion measure, may be subject to a number of judgmental errors, Anastasi (1982) points out that ratings are an important source of criterion data, when obtained under carefully controlled conditions. She considers the use of ratings as the very core of the criterion measure. Other measures used as criteria are: sales volume, hourly output, probability of turnover, length of service, absence/tardiness record, and completion and success in a subsequent training program.

If criterion-related validity is seen as the heart of employment testing, and if the predictor is said to be no better than the criterion used to predict it, then traditional concerns of predictive efficiency and

measurement accuracy should focus on the criterion measure as well as the predictor (which now receives most of the attention). Criteria should be reliable and should be evaluated by the same psychometric standards that are applied to predictor variables (Wiggins 1973). The performance appraisal criterion, which is commonly used as a measure of job success, can be thought of as a psychological test. However, serious questions arise as to the subjective nature of performance appraisals, which involve rater judgment. Dunnette and Borman (1979) review research on the "criterion problem" and note that the area is receiving increasing attention. Recent research on rater training does suggest that errors can be reduced when efforts are taken to help raters do a better job. Firms might consider training their staff members who are involved in conducting performance appraisals or ratings in connection with test validation.

When reporting a test's validity in the manual, the actual coefficient is only one piece of information that will be provided. Cascio (1982) lists the conventional definition of small (.10), medium (.30) and large (.50) validity coefficient values, representing the relationship between the predictor and criterion. For example, a validity coefficient of .40 is considered to represent a strong predictor-criterion relationship. Schultz and Schultz (1986) contend that validity coefficients of .30 to .40 are acceptable for use in the selection process. Anastasi (1982) notes that even validity coefficients as low as .20 may be effective in the selection process, and that the size of the coefficient needs to be examined in light of how the test is to be used.

With predictive validity, test manuals should also provide information on the length of the time interval between predictor and criterion measurement, as well as reporting on any training program during this time interval. The number, composition and other relevant characteristics of the validation group should be specified. Since validity is reported in terms of the correlation coefficient, the relative heterogeneity of the sample used needs to be reported, as a wider range of scores may lead to a higher validity coefficient (Anastasi 1982). Also, Anastasi (1982) contends that when reporting data for criterion-related validities, test manuals should describe the specific details of the criterion measure used as well as the job duties involved (since job titles do not properly describe a job).

Another problem associated with criterion-related validity results from pre-selection. Pre-selection refers to the fact that the workers in the validity study represent the best workers of the group who applied for the job (especially with concurrent validity). Anastasi (1982) notes that with pre-selection, the variability of the group's predictor-criterion scores will be curtailed at the lower end of the distribution (leading to a lowered validity coefficient).

Also the form of the relationship between predictor and criterion should be noted. Validity is usually reported using the type of correlation coefficient (e.g., Pearson Product Moment) that assumes the relationship is linear and uniform throughout the range (Anastasi 1982). One group of researchers (Schmidt, Hunter, McKenzie and Muldrow 1979), in their work on the relationship between predictor and criterion have found that these assumptions are usually met.

Rational validity is the second main category. Two types of validity are discussed under this approach: content and construct. Though these types of validity involve the inherent nature of the test itself rather that its relationship with an outside criterion, these types of validity also have relevance to employment testing.

The first type, content validity, involves the examination of the content of a test to ensure that it includes a representative sample of the domain area to be measured (Schultz and Schultz 1986). Content validation is mainly involved in evaluating educational achievement tests, an area where content and proper sampling is of great concern. For a number of tests used in the selection process, (especially in job knowledge or work sample tests), the content validation method is also used. This type of test was previously discussed under category three of test classification: performance tests. The trade test is a common example of a work knowledge test. Content validation is used here to answer the question: does a certain job performance test contain a fair sample of the job performance domain it purports to represent?

There has been a growing interest in content validity in the area of personnel selection, championed by Guion (1976). Anastasi (1982) notes that content validation is applicable to job knowledge or job performance tests and that it depends on a thorough and specific job analysis (carried out to verify the resemblance between the job activities and the test). The job analyst should use various sources of information (e.g., reports of job experts, supervisors and actual job incumbents) to help obtain a clear picture of all that the job entails. Job samples or simulations represent another attempt to reproduce actual job functions (Anastasi 1982). A typing or a filing test are examples of job sample tests. A test simulating flight conditions for pilots is an example of a job simulation test. Assessment centers, which are often used for managerial selection, typically use a number of job simulation tests as part of their test batteries.

One specific method of developing a job performance test is the job element method, often used for blue collar industrial jobs. Anastasi (1982) defines job elements as "those specific job behaviours that differentiate most clearly between marginal and superior workers" (p.436).

Guion (1978) proposes that the term "content oriented test development" more aptly describes content validity, since it involves non-correlational inferences about test construction. Content validity is not quantified like predictive validity. It involves a judgment process. Cascio (1982) points out that content validity is of growing importance in employment testing as certain tests require content validation, and predictive validity is not always practical.

The second type of rational validity, construct validity, is concerned with determining the underlying construct or trait that the test is attempting to measure. Examples of a construct are intelligence and mechanical comprehension. Construct validity can be thought of as the most theoretical type of validity. Yet Anastasi (1982) contends that construct validity is a comprehensive concept that includes all other types of validity since construct validation requires the gradual collection of information from many sources. That is, all information collected from all types of validation studies, whether predictive or content validity, add to the accumulated information for a particular construct. Construct validity is often considered outside of the realm of employment tests, as a non-applied, indirect type of validity. Yet criterion-related validity can add to the knowledge of construct validity and in turn benefit from the findings of construct validity. Cronbach (1984) points out that every test is impure to some degree (i.e., it does not measure exactly what it purports to) and identifying these impurities in a test is one part of the process of explanation or construct validation. For example, a better understanding of the construct underlying mechanical aptitude tests may lead to practical consequences such as improvements in the test material, resulting in a weeding out of certain impurities in the test.

There are a number of research techniques for investigating construct validity put forth by Cronbach (1984), Anastasi (1982) and Cascio (1982): analyzing internal consistency of the test; content validation by expert judges; stability of the scores over time; correlation with a number of practical criteria (criterion-related validity); studies of group differences on test scores; factor analysis of a group of tests showing the inter-relationships of the data; and convergent (showing that a test correlates with variables it should show relation to) and discriminant validity (showing that a test does not correlate with variables it should not be related to).

Cronbach (1984) describes construct validation as a complex, fluid, never-ending process of piecing together many parts of data and evidence. Construct validation can be likened to a slow refining process with new data helping to weed out the impurities in a test.

Before ending this section on types of validity, the subject of face validity should be briefly considered. Face validity is not relevant in the technical sense, but refers to what the test superficially appears to measure (Anastasi 1982). Cascio (1982) points out that since tests are used to help make personnel decisions, face validity is very important as it may affect the test-taker's motivation and reaction to the test. Face validity really concerns an organization's public relations and rapport with potential employees. If a test appears irrelevant to the test-taker, poor motivation and lack of cooperation may result. Also the test-taker may have a negative opinion of the organization where he or she took the test. An important question regarding a test is: Does the test appear to be related to the job area that the test-taker is applying for? Anastasi (1982) contends that to be effective, a test needs to have face validity and she recommends improving face validity where possible by reformulating test items in terms relevant to the particular test. For example, a mathematical reasoning test for pilots can be reworded to appear relevant for this particular group.

VALIDITY GENERALIZATION

With respect to criterion-related validity, local or firm by firm validation has been the recommended ideal. The assumption behind local validation is that a test's validity is specific to the organization where it was validated (and also specific to the job on which it was validated). Cascio (1982) calls this assumption of "situation specific validity" one of the "orthodox doctrines of personnel psychology" (p.159). This doctrine has been challenged by a group of researchers who hold that "validity generalization" is the more proper assumption: that validity derived for one job in one firm can be generalized to cover the same or similar job in another firm (Schultz and Schultz 1986).

The assumption of situation specific validity and the requirement of local validation presents a number of problems. Firm by firm validation (as well as job by job validation) is expensive, time-consuming and in many cases impractical to carry out (e.g., when job sample size is small). Lack of validity generalization means that each time a test is used, its validity needs to be checked afresh. For example, an organization using a mechanical comprehension test for its machinists cannot rely on the results of another comparable firm's validity study (of their own machinists) but needs to carry out their own validation check. More seriously, Cascio (1982) contends that only with the assumption of validity generalization (from firm to firm and from job to job) will it be possible for personnel psychology to advance past its present stage of a technology to that of a science. Also, Guion (1976) earlier noted that the lack of validity generalization presents a serious constraint in the usefulness of standardized tests in personnel selection.

The empirical foundation for this belief of situation specific validity was the finding of considerable variability in observed validity coefficients from study to study, even though jobs appeared to be identical (Ghiselli 1966). One group of researchers (Schmidt and Hunter 1977) asked the question: Was the variance in the firm by firm validity coefficients due to the fact that no two jobs or no two firms were alike, or due to statistical artifacts? The latter was a likely factor. Anastasi (1982) points out that validity studies in personnel selection generally have too small a sample size (typically 50 or less) to yield a stable estimate of validity coefficients.

To test the hypothesis that statistical artifacts account for the apparent situation specific nature of validity, Schmidt and Hunter (1977) developed a sophisticated method of "meta-analysis" (using an application of Bayesian statistics) that involved a large scale re-analysis of previous validity studies. This method cumulates results across studies for a given job-test combination and corrects the variance for a number of statistical artifacts. Some of these sources of artifacts are common errors encountered in correlational studies. Examples of these artifacts are: small sample size or sampling error; differences between studies in criterion reliability; differences between studies in test reliability; and differences between studies in range restriction, namely pre-selection (Schmidt and Hunter 1981). These artifacts are listed in ascending order of importance. In one study (Schmidt, Hunter, Pearlman and Shane 1979), these artifacts listed here accounted for well over half of the variance in the distributions of validity coefficients (for the job-test combination of clerical and first line supervisor jobs and tests). Callendar and Osborn (1980) have corroborated these findings on validity generalization using a number of different procedures. It should be noted here that these robust validity generalization findings are restricted stricted to cognitive ability tests, tests which are commonly in personnel selection.

Schmidt and Hunter (1981) take the debate one step further in their attempt to refute the assumption of job specific validity. Job specific validity refers to the lack of validity generalization from one job to another (jobs in the same job family). Schmidt, Hunter and Pearlman (1981) conducted a study on job specific validity and found the validity of seven cognitive abilities to be consistent across five clerical job families. These findings suggest that there is a common core of cognitive abilities in these job families. Schmidt and Hunter (1981) contend that separate validity studies are not required for each job. Instead, tests can be validated at the job family level. From their cumulative research findings, they conclude that for cognitive tests, validities can be generalized across firms and across jobs.

These findings have a number of practical implications. Since this model of test validation allows researchers to assess the degree of generalizability of prior validity results to the present job situation, a local validation study may not be required (Anastasi 1982). If a local validation study is conducted, its results can be interpreted in concert with prior results. Cascio (1982) points out that as validity generalization evidence accumulates for different occupations, only a job analysis (rather than a local validation study) will need to be carried out to ensure that a job is in the same job family (for which a degree of validity generalization has been established). So, one major implication is the expectation of savings because less local validation studies will be required. Also, as Cascio (1982) earlier pointed out, the assumption of validity generalization could lead to advances in the personnel psychology area.

These statistical findings have not yet filtered down to practical use and to the legal system that attempts to regulate employment testing. After all, situation specific validity was long held to be the case and is reflected in legal decisions and in test guidelines. Tenopyr (1981) astutely reminds us that the much

stronger socio-political force can sometimes overwhelm the scientific force. Thus, at present there is a discrepancy between current research findings and legal policy regarding validity generalization. At present, validity generalization has very tight restrictions surrounding its use (American Psychological Association 1980). Local validation is usually recommended.

Schmidt and Hunter (1981) acknowledge that how tests are used is a matter of social policy. However, they do express cautious optimism regarding the impact of the validity generalization findings. While Cascio (1982) also applauds the recent validity generalization findings as a breakthrough, Cronbach (1984) sounds a note of caution. Cronbach (1984) believes that validities are much less generalizable than the Schmidt and Hunter group's findings propose. He suggests that a more conservative approach might be appropriate. Burke (1984) cautions that in the area of test validation we should not overcompensate for the past excesses (e.g., insistence that every test need be locally validated) by "committing similar excesses in the opposite direction" (p.113). He points out that there are other considerations (e.g., test fairness and practical utility) besides validity generalization. Schultz and Schultz (1986) are enthusiastic about the validity generalization research, noting that a resurgence in the interest of cognitive ability tests is a result of these findings. They too are aware that such an approach is very different from the one currently practiced and required by the American equal employment legislation.

Tenopyr (1981) points out that however appealing and practical the concept of broad validity generalization may be to the personnel practitioner, it will be some time before these results affect the real world. However, she does mention that some middle ground in this area might be recommended, one that acknowledges validity generalization but still calls for local validation studies on major groups of jobs. Local validation certainly will not be eliminated and nor should it be, though job to job validity generalization might be allowed in one job group for the same firm. Meanwhile, research on the concept of validity generalization goes on (Burke 1984).

TEST VALIDATION AND PREDICTION MODELS

Anastasi (1982) speaks of test users being concerned with validity at two possible stages. The first stage concerns choosing a test for use in a selection procedure. Here, a personnel manager relies on validity data reported in the test manual or other test information sources (e.g., another source is the Mental Measurements Yearbook edited by Buros, 1978). The second stage, involves an organization checking a test's validity against its own local criteria (local validation). Anastasi (1982) lists four steps in the local validation of industrial tests. These four steps involve job analysis (what are the skills and knowledge required for the job?); selecting an appropriate test; correlating the test with some criterion of job performance; and, putting together a strategy for personnel decisions, that is, deciding how to use the test scores.

The ideal full scale longitudinal validation (predictive validity) is unrealistic for the large majority of industrial situations and organizations. A number of problems prohibit full scale validation: unavailability of large employee samples performing the same or similar jobs, criterion data problems such as unreliability, and restriction of range through pre-selection (Anastasi, 1982). Some of these problems have been mentioned already with regard to predictive validity. Due to the problems associated with local validation, some organizations rely on the test manual's reported validity values, while others attempt content validation if it is feasible.

Cronbach (1984) describes a number of steps in a full scale selection procedure. The first step concerns job analysis, an attempt to identify the skills, knowledge, abilities and any special characteristics required for success in a specific job. First, the broad job category is determined, then a more systematic analysis follows. An analysis may rely on direct observation, discussion with workers and supervisors. The next step involves choosing one or more tests to measure the characteristics thought to be necessary to perform the job (characteristics derived from the job analysis). After choosing an appropriate test, the next step involves administering the test. After administering the test, a measure of job performance, the criterion, is gathered. Following this is an analysis of how the test scores are related to job success: the validity check. The selection plan is being verified here: does the test predictively distinguish good performers from poor performers? The last step involves translating the test score (predictor) into actual decisions, according to a selection plan. Often more than one test is required to measure a number of abilities required for a job. These tests used together for prediction are referred to as a test battery (e.g., a test of mechanical aptitude, spatial aptitude and manual dexterity might form a test battery). The scores from a test battery are typically combined into one of two types of linear statistical prediction models: a multiple regression equation or multiple cutoff scores.

With the first procedure, multiple regression equation the scores from the tests are combined into a weighted sum, which yields a predicted criterion score for each test-taker. The tests that are most highly correlated with the criterion will have the highest weight. The tests in the equation are often related to each other as well as with the criterion. Since the best possible predictor is desired, it is important that two tests are not highly correlated with one another as this represents unnecessary duplication. For example, if two different tests, one of mechanical reasoning and one of mechanical comprehension, cover much the same area, then only one such test is needed.

An important point regarding multiple regression equations, is the compensatory nature of the predictor scores. When the composite score obtained is more important than each individual test score, a certain lack of skill in one area (e.g., numerical aptitude) is allowed to be compensated by a higher level of skill in another area (e.g., verbal aptitude). The value of each test in the battery is allowed to vary in that a high score on one test can compensate for a low score on another test, as long as an acceptable total or composite score is obtained.

With the second statistical prediction model, multiple cutoff scores, there is a requirement that a certain minimum cutoff score is set for each test. Any applicant who falls below any one minimum test score is rejected. So only those applicants who reach or exceed all of the cutoff scores are considered. This relatively simple procedure is used when certain essential levels of a skill are required. For example, to function as a pilot there are certain required levels of a skill (e.g., visual acuity, spatial aptitude) that high levels of another skill cannot compensate for. This minimum score requirement forms one essential distinction between the multiple regression equation and multiple cutoff score procedures.

When certain assumptions (e.g., a linear relationship between tests and criterion) are true, Anastasi (1982) points out that a higher proportion of correct decisions will occur with a regression equation than with multiple cutoff scores. Multiple regression equations also allow for a ranking of applicants according to their predicted criterion score, while multiple cutoff scores allow no further evaluation. Cascio (1982) also regards the multiple regression equation as possibly the most efficient predictor. He notes that the multiple cutoff approach is practical and easy to understand, although the model can become complex

with a large number of tests with different cutoff scores. Both of these prediction models represent a single stage decision strategy. Both Anastasi (1982) and Cascio (1982) recommend that the optimal strategy in a number of situations may be a combination of both models, using these procedures in a sequential approach. The multiple cutoff model would be used first to select applicants who score above the minimum cutoff on certain tests. Then a multiple regression equation is used to predict the criterion scores on the remaining predictors.

DECISION THEORY AND UTILITY

Thus far we have been discussing the traditional or classical validity approach to personnel selection. Here the primary emphasis is on measurement accuracy and predictive efficiency (Cascio 1982). A highly valid test is an important goal in the classical validity approach. This classical model has a number of shortcomings associated with it. One major criticism is that the classical validity approach ignores "certain external parameters of the situation that largely determine the overall worth of a selection instrument" (Cascio 1982 216). That is, classical validity ignores other ways of evaluating the utility or value of a test, ways which take into account the types of decisions that results from using test scores.

A more recent and perhaps more realistic approach is that of decision theory, which emphasizes that the "outcomes" of prediction are of primary importance. Decision theory takes a much broader outlook in that it encompasses a number of external parameters of the test situation as well as the various outcomes of prediction. Here validity is viewed as a useful index which does not tell the whole story of predictive success (Wiggins 1973). Thus, validity should not be used as the sole basis in choosing a test, since it gives only one part of the picture.

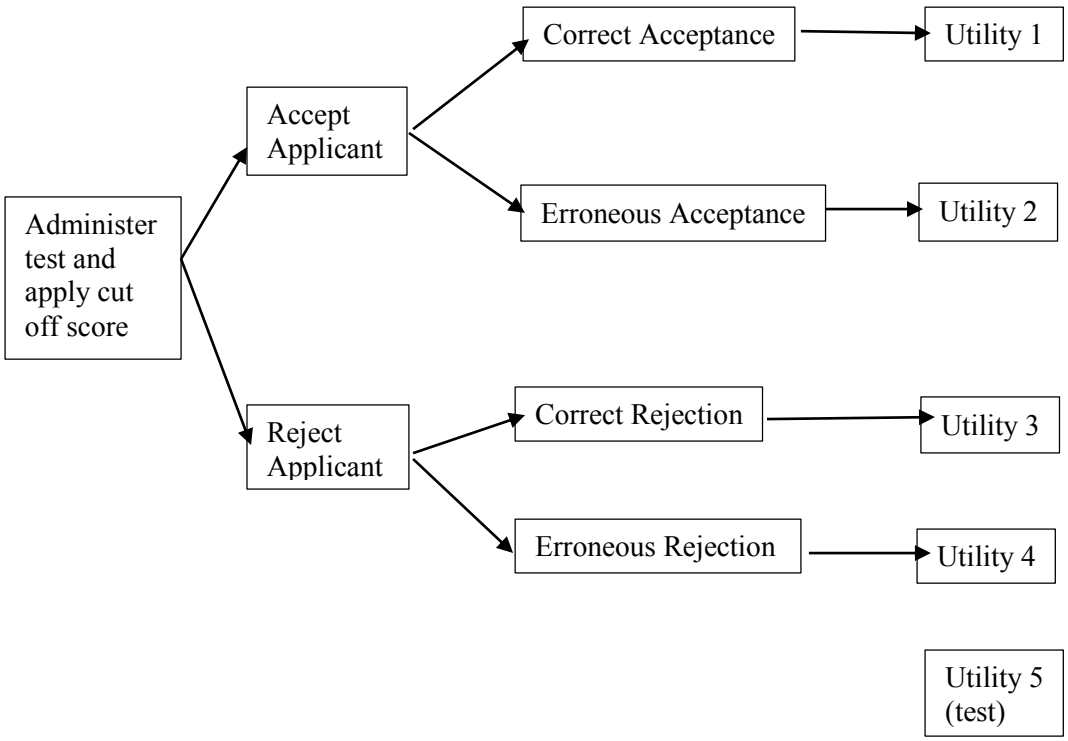
Basically, decision theory is an attempt to put the personnel decision making process into mathematical form. Anastasi (1982) notes that the decision theory process can be mathematically quite complex, and that the practical application to personnel testing has been proceeding slowly. Though it is not in widespread use, decision theory has served to focus attention on the complexity of factors that determine a test's contribution in a specific situation (Anastasi 1982). Cascio (1982) notes that the main advantage of decision theory is that it addresses the complexities of the situation and forces the decision maker to consider the kinds of judgments to be made. As this subject area is somewhat technical, only a few of the important concepts concerning decision theory will be discussed: utility, selection ratio and base rate.

Utility, can be defined as the values or judged favourableness of each decision outcome from a test (Anastasi 1982). Each decision outcome will have a particular utility associated with it. For example, with a one stage cutoff prediction model, there are four possible outcomes (see figure 1). Those applicants who score above the cutoff score are accepted and those who score below it are rejected. Some of the acceptances and rejections are correct and some are erroneous. Each of these four outcomes has a utility associated with it. A fifth utility is added to represent the cost of using the test. Classical validity attempts to minimize the second and fourth outcomes (erroneous acceptances and erroneous rejections) and treats both of these as equally costly, when in fact the former is often more costly than the latter (Cascio 1982). For example, the erroneous acceptances of a poorly qualified applicant for a certain factory job may result in damage to the machinery and may require extra supervision, both of which are costly consequences for an organization. Erroneous rejections, on the other hand, are regarded by some firms as relatively unimportant. Thus utility emphasizes the costs and benefits of selection strategies by focusing "attention

on the types of selection errors, their relative incidence, and associated costs" (Dunnette and Borman 1979 480).

Dunnette and Borman (1979) point out that there has been a failure to apply the precise but complex Cronbach and Gleser 1965 decision theory equations. One problem has been the development of a utility scale to assign values to the various outcomes. Cronbach and Gleser (1965) themselves call the assignment of values to outcomes the "Achilles' heel" of decision theory. When dollar values have been assigned to the outcomes of personnel decisions, there are often problems in costing some items in monetary terms. For example, if there is a high rate of erroneous rejections, the result may be a certain loss of public good will. How is this loss to be costed?

Figure 1, Outcomes and Utilities of a Personnel Decision Strategy



Recently there has been an increased interest in utility and decision theory following advances in the assignment of dollar values to outcomes. This knowledge allows for the general application of these utility/decision theory equations. One group of researchers (Schmidt, Hunter, McKenzie and Muldrow 1979) has derived a way to estimate the spread or standard deviation of employee job performance in dollars. Anastasi (1982) contends that the techniques developed in this research can be applied to the measurement of other outcomes of personnel decisions. Hunter and Schmidt (1983) have been successful in this endeavour. Dunnette and Borman (1979) believe that research on utility may become more frequent in the future due to these improvements in costing and due to the doubt thrown on the hypothesis of situation specific validity (the latter hypothesis would limit the application of utility to different settings).

Two important parameters that affect the utility of a test are the selection ratio and the base rate. The first one, the selection ratio, is defined as the proportion of applicants tested who are accepted (Cronbach 1984). A firm typically requires a certain number of workers and usually there are more applicants than positions open. For example, if there are 100 job applicants and an organization requires 30 workers, the selection ratio is .30. The higher the selection ratio (the more it approaches one - accepting all who apply), the less selective the firm is. Conversely, as the selection ratio approaches zero (accepting very few workers from the applicant pool) the more selective the firm is. It becomes more effective to use a test when the selection ratio is low. In practical terms this means if the firm is selecting a small number of applicants (a low selection ratio), even a test with low validity can be useful. However the higher the selection ratio (e.g., a firm selecting the majority of those who apply), the less cost effective and the less useful a test becomes (no matter how valid the test is).

The second parameter is the base rate, defined as: "the proportion of successful employees prior to the introduction of the test for selection" (Anastasi 1982 165). Thus, the base rate is related to the range of performance in an unselected work group (Ability Tests 1982). If in the past an organization has found that without using a selection test, 50% of its applicants are judged successful, their base rate is said to be 50%. Using a valid test will improve predictive accuracy most when base rates are in this middle range (Anastasi 1982). With a base rate of 90% most candidates will be successful, so a test will not be of much use.

Both of these parameters (selection ratio and base rate) have consequences for the utility of using a test. One important point from decision theory is that the validity of a test does not equal its utility. The overall cost and benefit of a test also depends on the selection ratio, the base rate, and the derived utility of each outcome, as well as the validity.

CURRENT USE OF TESTS IN PERSONNEL SELECTION

It is difficult to get a handle on the current amount of testing in personnel selection. Test use is guided by employer needs and availability of tests, as well as well by legal test guidelines. One source (Tenopyr 1981) believes that the actual amount of testing in personnel is over estimated. That there has been a decline in test use (since the advent of equal employment opportunity legislation) seems to be an un-disputed fact. However, Schultz and Schultz (1986) recently noted a resurgence of interest in cognitive ability tests following the finding of validity generalization for these tests. A recent Newsweek article

(1986) on testing also refers to the boom in all types of employment testing after their decline in recent decades.

The results of a 1983 ASPA-BNA survey of employee selection procedures sheds some light on current test use. Comparing this 1983 survey with the 1971 ASPA-BNA survey on personnel testing (mentioned in the history section) reveals some trends in employment testing. However it is difficult to compare some aspects of these two surveys due to differences in emphasis and differences in the reporting of results. In the 1983 survey, test usage results are categorized by type of test: 75% of the firms surveyed use skill performance tests/work samples; 22% use job knowledge tests; 20% use mental ability tests; and, 9% and 6% use personality tests and physical ability tests, respectively (ASPA-BNA Survey 1983). Also a small percentage (6%) report using the services of an assessment center for their testing needs. In the 1971 survey, 55% of those firms surveyed report using some kind of test (here test use is not classified by test type). In the 1983 survey there is a large discrepancy between the reported use of skill performance tests/work samples (75%) and all other types of tests (e.g., use varies from 22% to 6%). It is difficult to ascertain if overall testing did actually increase or decrease from survey to survey because of survey differences in reporting test use and in the number of firms surveyed (the later survey is the more thorough of the two).

In a 1975 survey mentioned by Tenopyr (1981), approximately 60% of the larger firms report using tests while only 39% of the smaller firms do. She also notes that the majority of these firms report cutting back on their test use during the past five years. Both of these trends (higher levels of test use in larger firms and cutting back on test use) are also apparent in the ASPA-BNA surveys. In both the 1971 and 1983 surveys, employers report changes in their testing practices (e.g., a small percentage of the firms in the 1983 survey and an undetermined number of the firms in 1971 mention discontinuing all or part of their testing programs). Reasons for selection procedure change "focus on problems of validity and defensibility" (i.e., problems with validation or fear of legal action, ASPA-BNA Survey 1983 p.10). Only a very small percentage of the 1983 firms (mainly larger firms and non-business establishments) report having any legal challenges to their selection practices (here the survey does not mention if all of these challenged procedures were test related). Larger firms also reported using more testing than smaller firms.

Both the 1971 and the 1983 surveys identify office/clerical applicants as the most frequently tested group. Tests are more commonly used in the non-manufacturing sector (e.g., banks, public utilities, communication and retail sales) than in the manufacturing sector (Ability Tests 1982).

With respect ⁻0 validation, only 16% of firms in 1983 report having validated any of their selection procedures with United States federal guidelines, compared to 53% of firms who validated their tests in 1971. Large firms are more likely to do so than smaller firms. However more than a third of the firms in 1983 have used an outside consultant to validate their current selection procedures (here the survey does not specify if these procedures are all test related). When validation was carried out, the preferred criterion measure was the formal performance appraisal or supervisor's ratings of performance. With respect to norms, more than half of the firms in 1983 use company or local norms.

A very small percentage of the 1983 firms provided cost data on their testing programs. Annual operating costs for testing ranged from $25 to $5,000,000. Test development and validation expenditures started at

$1000 and went as high as $100,000. Cost per applicant went from a low $3 for a health care organization, to a high of $135 for a medium sized manufacturer. The 1971 ASPA-BNA survey also inquired about costs of testing programs. Taking inflation into account, the costs are similar. Interestingly, a majority of the 1971 firms reported not knowing what their testing programs cost them.

Thus far, all of these data apply to the United States. Though Canadian test use may not have been as widespread as American use, Dewey (1981) notes that there has not been as much of a corresponding decline in Canadian test use as there has been in the United States. It is also difficult to estimate current test use in Canada. A mini survey of the Toronto area major employers (reported by Dewey 1981) reveals that one fourth of the firms are using tests, although this is thought to be a conservative estimate.

ADVANTAGES OF EMPLOYMENT TESTING

There are a number of advantages to the responsible use of a well-chosen psychological testing program. Tests, as a formal selection device, are rigorous, scientifically developed instruments, which should have reasonable degrees of reliability, validity, objectivity, and should also be standardized and based on sound norms. These psychometric characteristics have served to set testing apart from other selection techniques (e.g., the interview). The superior predictive validity of psychological tests in selection has been well documented (Hunter and Hunter 1984). The average validity for cognitive ability tests is .53, while alternate predictors (e.g., biographical inventories-.37, reference checks-.26, experience-.18, and the interview-.14) all have lower validities (Hunter and Hunter 1984).

Objectivity in a selection technique is deemed to be fairer than a subjective method of selection. Rowe (1980) notes that a good test is more objective than alternate selection methods (e.g., interviews, and letters of reference), which may allow personal biases to influence the selection decision. Also tests may improve efficiency in the selection process. One compelling fact is that employment tests are a valuable source of information about applicants, in a relatively information-poor situation (Ability Tests 1982). Also this information can be gathered in a short period of time.

One main advantage of psychological testing is its purported economic superiority in obtaining the best possible match between worker and job and thus increasing productivity. The fact that valid selection procedures (using cognitive ability tests) have resulted in a more productive workforce, has been well documented (Stone and Ruch 1979; Schmidt, Hunter, McKenzie and Muldrow 1979; Hunter and Schmidt 1983; Schmidt, Mack and Hunter 1984). Stone and Ruch (1979) mention a testing program that resulted in savings for a firm by substantially reducing costly turnovers. Recently, new techniques for estimating worker productivity (mentioned in the decision theory section) have been developed and applied. Schmidt, Hunter, McKenzie and Muldrow (1979) in an early study, found that the use of valid cognitive ability tests increased the average performance level of the workforce and increased productivity. Their reasoning is as follows: the mental skills that cognitive ability tests measure are important determinants of job performance, so selecting high performers is important for a firm's productivity. Hunter and Schmidt (1982) note the drop in the growth rate of productivity (in the United States) during the last ten years and attribute this in part to the abandonment of valid cognitive ability test in selection. Recent research findings (Schmidt, Hunter and Mack 1984; Hunter and Schmidt 1983) confirm that valid selection tests can produce major increases in work force productivity. Hunter and Schmidt (1982) note that these

findings on productivity imply "large economic benefits for good personnel selection and large losses if valid selection programs are abandoned." (p.198).

A number of researchers contend that valid tests, as objective selection instruments, have had a positive effect with respect to discrimination in the personnel selection process. Yoder and Staudohar (1984) report that the imposed federal guidelines on testing practices have had a positive effect on testing: this regulation has led to greater reliability and validity of employment decisions and also it has helped reduce discrimination. They conclude that "valid tests provide a useful function in choosing qualified employees in a non-discriminatory manner" (p.74).

Tenopyr (1981) investigated the use of alternate selection techniques. She concluded that " there are no alternatives better than tests, when validity, degree of adverse impact on various groups and feasibility of use are all taken into account" (p.1124). Thus, in terms of its psychometric properties (validity, objectivity and the economic benefits to be derived), a sound testing program can be a cost effective and integral part of the selection process.

LIMITATIONS OF EMPLOYMENT TESTING

The use of tests in the selection process is not without its limitations. Controversy surrounding employment tests has focused on a number of issues: the misuse of tests; test bias; and ethical concerns over privacy. Rowe (1980) discusses the misuse of tests under two categories: the use of bad tests and the misuse of good tests. She notes that the market demand for selection tests has permitted the development of "bad" tests that are over-promoted and slickly marketed by unqualified testing firms. These tests are usually lacking in some essential test characteristic (i.e., they may be poorly standardized or inadequately normed, lack reliability or be improperly validated). She recommends that firms deal with competent and professional test sellers. Schultz and Schultz (1986) point out that one danger of testing is uncritical test use. Some personnel managers may lack the ability to discriminate between good and bad tests. Rowe (1980) also believes that part of the problem of test misuse stems from inadequate training of personnel managers. If tests are to be used, personnel departments should deal with reputable test sellers and should have the know-how to choose a test properly.

The second category relates to the misuse of psychometrically sound tests. In the section on test characteristics, earlier it was pointed out that a good test can be rendered useless with improper and careless administration. Another misuse relates to using a test on a group for whom there are not proper norms. Rowe (1980) points out that the use of tests (mostly American) without Canadian norms or validation studies is a misuse. It is difficult to assess how much of a difference exists between the two countries. Local norms are typically more valuable for firms, are usually worth developing, and seem to be widely used (ASPA-BNA Survey 1983). Also, tests may be chosen for the wrong job. Proper job analysis and proper knowledge of tests should result in the proper choice of tests.

One type of test that has been singled out for criticism is the personality test. Haney (1981) notes that the flurry of social concern in the mid 1960s over personality testing focussed on the use of these tests in the United States federal government's personnel selection procedures. The place of personality tests in personnel selection is still questioned (Newsweek 1986).Some cautions regarding personality tests were expressed earlier under the section on classification of tests (e.g., the problem of faking, low validity and

reliability associated with personality tests). Rowe (1980) notes that the use of personality tests has declined. Since the test guidelines required that employment practices have a legitimate business purpose (must demonstrate a relationship between predictor and criterion), a number of personality tests (which have low validity) were dropped from the personnel selection procedures of some firms. From the ASPA-BNA 1983 survey, only 9% of firms report using some kind of personality test.

The next issue of limitations relates to bias. Possible discrimination or bias as a result of psychological testing in personnel is a much researched topic. Criticism has been directed for years at employment testing programs for possible bias against minority groups. Fair employment legislation was enacted to protect these minority groups against potential discrimination during employment testing.

One important point that Tenopyr (1981) points out is that "tests do not discriminate against various groups, people do" (p.1121). Thus the term "test bias" is a misnomer, as bias is a function of the way a test is used and not an inherent property of the test. The issue of test bias pertains to two points: 1) differential validity (validity coefficients); 2) test unfairness (the relationship between group means on the test and on the criterion) (Anastasi 1982).

The first point, differential validity, holds that the validity would be lower for minority applicants than for majority applicants (Hunter and Schmidt 1982). Both Cascio (1982) and Schmidt and Hunter (1981) conclude that true differential validity probably does not exist. Tests used in selection appear to be "equally" valid for all groups. Any previous findings of differential validity were due to statistical artifacts (small sample size studies where significant differences in validity occurred by chance, Schmidt and Hunter 1981).

The second point, test unfairness, relates to the fact that: "even if equal validity existed, tests might still be unfair if minority applicants made systematically lower scores than their ability warranted, because of a number of culturally biased test items" (Hunter and Schmidt 1982 297). If test unfairness were the case, then minority groups might perform better on the job success criterion than was predicted from their test scores. That is, the test is unfair to the minority group in that it underpredicts their criterion performance. The regression model of unfairness, which is commonly accepted and is reflected in the uniform test guidelines (of the EEOC), defines "a test as unfair to a minority group if it predicts lower levels of job performance than the group in fact achieves" (Schmidt and Hunter 1981 1131). Research has not supported this hypothesis of test unfairness. Lower test scores for minorities are accompanied by lower levels of performance on the criterion measure. Again as with validity generalization, there is a discrepancy between research findings and social policy regulations.

Yoder and Staudohar (1984) note that test bias (claims of differential validity and test unfairness) cannot be blamed for test score differences between minority and majority groups. These test score differences are most likely due to social disadvantages (e.g., less education) that minority groups experience. Thus minority group applicants do not score as high as majority group applicants and these results cannot be attributed to biased employment tests. Yoder and Staudohar (1984) actually note that tests, being objective instruments, can help prevent discrimination.

Schmidt and Hunter (1981) point out that these research findings reveal that the problem is no longer in the tests and that it cannot be solved by modifying or eliminating these tests. The problem is a social one,

where minority groups are at a disadvantage (educationally and socially) and are not picking up the cognitive skills needed in a modern society (Schmidt and Hunter 1981). Both Yoder and Staudohar (1984) and Schmidt and Hunter (1981) discuss the trade-off between the goals of economic productivity (selecting the most productive workers) and achieving equal opportunity (proportional minority representation in the work force). As these are competing goals, policy makers must decide on the optimum way of balancing these concerns. Yoder and Staudohar (1984) note that the EEOC has used testing requirements to shift government policy from the "requirement of equal treatment to that of equal outcome" (p.71). Hunter and Schmidt (1982) suggest that the best way of trading off productivity and increasing minority employment is some type of preferential selection system (i.e., quotas). They maintain that using quotas (e.g., top down hiring within each group) can increase minority employment faster and with less productivity loss than random hiring of applicants above a low cutoff score.

This is a social policy problem and perhaps employment tests are unfairly singled out as an easy target. Some researchers (Gordon and Terrell 1981) argue that testing should be less concerned with unbiased predictive validity and more concerned with aiding equal opportunity. Yet, tests cannot bear the burden of resolving the tension between the goals of a productive workforce and equal opportunity (Ability Tests 1982). Tenopyr (1981) concluded that these findings of non-test bias have little meaning in the real world at present as government policy essentially controls the employment situation.

Ethical concerns with respect to tests as an invasion of privacy are being voiced (Newsweek 1986). Many of these concerns revolve around the use of personal or intimate questions (sometimes found in personality tests). Schultz and Schultz (1986) point out that personality tests are the main targets of this criticism and recommend that personal questions that have no relevance to the job are an invasion of privacy and should be avoided. Dessler and Duffy (1984) discuss the individual rights of test-takers and test security. Confidentiality of results needs to be assured as well as the right to "informed consent" regarding the use of a test-taker's results.

CONCLUDING COMMENTS

It should be emphasized again that tests, though a powerful tool, are only one part of a complex decision making process in personnel selection. Tests need to be chosen properly and administered by those with the proper knowledge to do so. A testing program, no matter how valid the tests are, is only as good as the firm that runs it (the time, research and effort spent on the testing program can make the difference between a poor testing program and a good program). Yoder and Staudohar (1984) conclude that "no alternatives to standardized tests have been found that are equally informative, equally adequate technically, also economically and politically viable" (p.71).

A number of recent developments in the testing area were discussed: the findings of validity generalization (with respect to cognitive ability tests), the application of decision theory in terms of utility, and relatedly, the determination of increased work force productivity resulting from the use of valid test selection methods. One overlooked area that needs to be researched is the criterion measure. Also it was emphasized that classical validity does not tell the whole picture with respect to test use.

It has been determined that bias is not inherent in the test itself, but depends on how the test is used: tests can be used in a non-discriminatory manner and actually prevent discrimination or they can facilitate bias

in some cases. Tests, as part of the selection process, are regulated in a society that seeks to correct past abuses and ensure equal opportunity for all groups. Thus, test use by employers is caught up in government socio-political concerns of equity. The economic self-interest of employers and the larger public interest in equity needs to be balanced.

Current test guidelines and legal requirements lag behind recent research findings (e.g., validity generalization). These findings have had little impact on the real world of employers and may not have an effect for some time. While the use of employment tests has been on the decline since the late 1960s, Schultz and Schultz (1986) note a resurgence of "interest" in the cognitive ability tests. Whether or not this interest translates into increased test use is debatable given current test guidelines and current political forces.

REFERENCES

Ability Tests. 1982. Across the board 19(7):27-32.

Aiken, L.R. 1979. Psychological Testing and Measurement. 3rd edition. Boston: Allyn & Bacon.

American Psychological Association, Division of Industrial-Organizational Psychology. 1980. Principles for the validation and use of personnel selection procedures. 2nd edition. Berkely, California: APA.

Anastasi, A. 1982. Psychological Testing. 5th edition. New York: MacMillan Publishing Co.

ASPA-BNA Survey: Personnel Testing. 1971. Bulletin to Management. BNA Policy and Practice Series. Washington, D.C.: Bureau of National Affairs.

ASPA-BNA Survey No. 45: Employee Selection Procedures. 1983. Bulletin to Management. BNA Policy and Practice Series. Washington, D.C.: Bureau of National Affairs.

Barrett, G. V., J. S. Phillips, and R. A. Alexander. 1981. "Concurrent and predictive validity designs: a critical reanalysis." Journal of Applied Psychology 66:1-6.

Beach, D.S. 1980. Personnel: The Management of People at Work. 4th edition. London: Collier MacMillan Publisher.

Brown, F. G. 1976. Principles of Educational and Psychological Testing. 2nd edition. New York: Holt, Rinehart & Winston.

Burke, M. J. 1984. "Validity generalization: A review and critique of the correlational model." Personnel Psychology 37:93-115.

Buros, O. K., ed. 1978. The Eighth Mental Measurement Yearbook. Highland Park, N.J.: Gryphon.

Callender, J. C., and H. G. Osborn. 1980. "Development and test of a new model of validity generalization." Journal of Applied Psychology 65:543-558.

Campbell, J.P. 1976. Psychometric theory. In Handbook of Industrial and Organizational Psychology, edited by M.D. Dunnette, pp. 185-222. Chicago: Rand McNally.

Cascio, W.F. 1982. Applied Psychology in Personnel Management. 2nd edition. Reston, Virginia: Reston Publishing Company.

_____ 1986. Managing Human Resources: Productivity, Quality of Work Life, Profits. New York: McGraw Hill Book Co.

Cronbach, L. J. 1949. Essentials of Psychological Testing. New York: Harper and Row.

_____. 1984. Essentials of Psychological Testing. 4th edition. New York: Harper and Row.

_____ , and G. C. Gleser. 1965. Psychological Tests and Personnel Decisions. 2nd edition. Urbana Ill.: University of Illinois Press.

Dessler, G., and J. F. Duffy. 1984. Personnel Management. Canadian 2nd edition. Scarborough, Ontario: Prentice Hall Canada Inc.

Dewey, M. 1981. "Employers take a hard look at the validity and value of psychological screening." Globe & Mail (February 7), Bl.

Dunnette, M. D., and W. C. Borman. 1979. "Personnel selection and classification systems." Annual Review of Psychology 30:477-525.

Fleishman, E. A. (1975). "Toward a taxonomy of human performance." American Psychologist 30:1127-1149.

Flippo, E. 1971. Principles of Personnel Management. 3rd edition. New York: McGraw Hill Book Co.

Ghiselli, E. E. 1966. The Validity of Occupational Aptitude Tests. New York: Wiley.

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

Ghiselli, E. E. and C. W. Brown. 1955. Personnel and Industrial Psychology. 2nd edition. New York: McGraw Hill Book Co. Ltd.

Gordon, E. W. and M. D. Terrell. 1981. "The changed social context of testing." American Psychologist 36(10):1167-1171.

Guion, R. M. 1965. Personnel Testing. New York: McGraw Hill.

_____. 1976. Recruiting, selection, and job placement. In Handbook of Industrial Organizational Psychology, edited by M.D. Dunnette, pp. 777-828. Chicago: Rand McNally.

_____. 1978. "Content validity in moderation." Personnel Psychology 31:205-213.

Haney, W. 1981. "Validity, vaudeville, and values: A short history of social concerns over standardized testing." American Psychologist, 36(10):1021-1034.

Hunter, J. E. and R. E. Hunter. 1984. "Validity and utility of alternative predictors of job performance." Psychological Bulletin 96:72-98.

Hunter, J. E. and Schmidt, F. L. 1982. "Ability tests: Economic benefits versus the issue of fairness." Industrial Relations 21:293-308.

Hunter, J. E. and F. L. Schmidt. 1983. "Quantifying the effects of psychological interaction on employee job performance and work-force productivity. American Psychologist 38:473-478.

Jain, H. C. 1974. Employment tests and discrimination in the hiring of minority groups. In Contemporary Issues in Canadian Personnel and Administration, edited by H. C. Jain, pp. 148-154. Scarborough, Ontario: Prentice Hall.

Laird, D.A. 1937. The Psychology of Selecting Employees. 3rd edition. New York: McGraw Hill Book Co. Ltd.

Lishan, J. M. 1948. "The use of tests in American industry: A survey." Personnel 24(4):305-308.

Moore, H. 1939. Psychology for Business and Industry. New York: McGraw Hill Book Co. Ltd.

_____. 1942. Psychology for Business and Industry. 2nd edition. New York: McGraw Hill Book Co. Ltd.

Murray, D.J. 1983. History of Western Psychology. Englewood Cliffs, N.J.: Prentice Hall.

Newsweek. May 1986. "Can you pass the job test?" New York: Newsweek Inc. 48-53.

Rowe, P.M. 1980. "Psychological tests in selection: Their use, misuses and abuse." The Canadian Personnel and Industrial Relations Journal 27(2):37-40.

Schmidt, F.L. and J.E. Hunter. 1977. "Development of a general solution to the problem of validity generalization." Journal of Applied Psychology 62:529-540.

Schmidt, F. L. and J. E. Hunter. 1981. "Employment testing: Old theories and new research findings." American Psychologist. 36(10): 1128-1137.

Schmidt, F.L., J.E. Hunter, R.C. McKenzie, and T.W. Muldrow. 1979. "Impact of valid selection procedures on work force productivity." Jounal of Applied Psychology 64:609-626.

Schmidt, F. L., J. E. Hunter, and K. Pearlman. 1981. "Task differences and validity of aptitude tests in selection: a red herring." Journal of Applied Psychology 66:166-185.

Schmidt, F. L., J. E. Hunter, K. Pearlman. and G. S. Shane. 1979. "Further tests of the Schmidt-Hunter Bayesian Validity Generalization Procedure" Personnel Psychology 32:257-281.

Schmidt, F. L., M. J. Mack, and J. E. Hunter. 1984. "Selection ability in the occupation of U.S. park ranger for three models of test use" Journal of Applied Psychology 69:490-497.

Schultz, D. P. and S. E. Schultz. 1986. Psychology and Industry Today: An Introduction to Industrial and Organizational Psychology. 4th edition. New York: Macmillan Publishing Company.

Srinivas, K.M. 1984. Human Resource Management: Contemporary Perspectives in Canada. Toronto: McGraw Hill Ryerson, Ltd.

Stone, C. H. and F. L. Ruch. 1979. Selection, Interviewing, and Testing. In ASPA Handbook of Personnel and Industrial Relations, edited by D. Yoder and H. G. Heneman. Washington D.C.: Bureau of National Affairs. 4-117-4-158.

Strauss, F. and R. Sayles. 1972. Personnel: The Human Problem of Management. 3rd edition. Englewood Cliffs, N.J.: Prentice Hall.

Tenopyr, M. L. 1981. "The realities of employment testing." American Psychologist. 36(10): 1120-1127.

Werther Jr., W.B., K. Davis, H. F. Schwind, H. Das and F. C. Miner. 1985. Canadian Personnel Management and Human Resources (2nd edition). Toronto: McGraw Hill Ryerson Ltd.

Wiggins, J.S. 1973. Personality and Prediction: Principles of Personality Assessment. Reading, Mass: Addison Wesley.

Yoder, D. J. and P. D. Staudohar. 1984. "Testing and EEO: Getting down to cases." Personnel Administrator 29(2):67-74.

Industrial Relations Centre (IRC)
Queen's University
Kingston, ON K7L 3N6
irc.queensu.ca