1984

A Validation Study for Four Tests of Behavioral Cognition

Kenneth J. Kish

A VALIDATION STUDY FOR FOUR TESTS OF BEHAVIORAL COGNITION

Kenneth J. Kish, B.A.

A Digest Presented to the Faculty of the Graduate School of the Lindenwood Colleges in Partial Fulfillment of the Requirements for the Degree of Master of Science

This study is a predictive validity study employing subjects who were currently employed as psychiatric nurses on the four locked psychiatric units at Barnes Hospital, in St. Louis, Missouri. The purpose of the study was to examine the relationship of four tests of social intelligence or behavioral cognition, and on the job performance, looking for a correlation between test scores and successful job performance. The criteria of job performance employed in the study were five scales designed by the researcher according to the behaviorally anchored scales model of performance reviews. Employees were rated by their respective supervisors using the five scales which were rated from 1-7 with 1 denoting poorest performance and 7 denoting the best performance. Scores on the four tests were summed to form composite scores in accordance with the test author's endorsement of this approach for research efforts involving complex performance criteria.

A VALIDATION STUDY FOR FOUR TESTS OF BEHAVIORAL COGNITION

Kenneth J. Kish, B.A.

A Culminating Project Presented to the Faculty of the
Graduate School of the Lindenwood Colleges
in Partial Fulfillment of the Requirements
for the Degree of Master of Science

COMMITTEE IN CHARGE OF CANDIDACY

Samuel Zibit, Core Faculty- Adviser

Nancy Klepper

Jerry Bolandis

COMMITTEE IN CHARGE OF CANDIDACY:

Samuel Zibit, Core Faculty- Advisor

Charles Orme-Rogers

Jerry L. Bolandis

Table of Contents

LIST OF TABLES

LIST OF ILLUSTRATIONS

INTRODUCTION

This paper presents the findings of a research project undertaken at Barnes Hospital, in St. Louis, Missouri, involving the use of testing in the selection of psychiatric nurse. Before the specific hypothesis, method, and results of this predictive validity study are elucidated, an introduction to basic underlying topics is offered. These topics include the history of testing in business and industry, the diverse nature of testing, the meaning of validity and test validation, and a summary of related research efforts. This introduction lays the conceptual groundwork for the presentation of the details and findings of the actual research project, which follow immediately.

Testing In Industry

Since the turn of the century when the pioneering efforts of Hugo Munsterberg and Robert Cattell gave birth to the discipline of industrial psychology, testing has occupied an eminent place in what is now called personnel management. Munsterberg saw his role as one of helping employees find well suited positions which maximized their achievement and output. This is a worthy goal, and certainly, from a business perspective, a financially sound goal. Personnel, from any perspective, are a major asset of any firm or organization. Much time, energy, and money are spent in the hopes that investing in personnel will yield large dividends. In recent years, some accountants have even supported the idea that "human asset accounting" (Glueck, 1974, p.25) be adopted to give this human, personnel factor its appropriate place of value on the balance sheets. Huamn asset accounting would, theoretically, aid managers in avoiding errors in decision making which can result from overlooking the impact of judgements on non-material assets.

The selection process initiates the relationship between a company and its personnel, and is a focal point in that relationship. Recruiting, hiring, and subsequent training are obviously time consuming and expensive procedures. The higher the success rate in

matching applicants with appropriate positions, as Munsterberg described, the better off the organization will fare. The developement of specialized mechanisms for selecting and dealing with personnel has been dramatic. This developement has proceeded from the handful of employment departments in such companies as National Cash Register and Goodrich in 1912, to the personnel associations during World War I. Then came the introduction of college courses offering academic training in this area during the 20's. All of these have mushroomed into the tens of thousands of personnel specialists employed in industry during the 60's, 70's, and currently. (Glueck, 1974, p. 13) Now assessment centers have come on the scene as the latest attempt to structure selection for greater success ratios in hiring. Testing applicants has traditionally been one of the two major elements which businesses have used throughout the course of this developement to make selection more fruitful. (Interviewing is, of course, the other major tactic.)

During the past two decades there has been a cutback in the use of testing in industry, although surveys indicate that testing is still heavily relied upon, especially by firms with large numbers of employees. (Flippo, 1980, p.154) There are literally thousands of tests published and sold currently to firms

searching for ways to increase success in hiring. (Buros, 1972) Probably the most influential factors responsible for the reduced level of testing have been Title VII of the Civil Rights Act of 1964, the efforts of the Equal Employment Opportunity Commission which enforces it, and litigation which has clarified and supported the application of this legislation in the business arena. Certainly one of the most pertinent court cases was the Griggs et. al. vs. Duke Power Company case. The outcome was a Supreme Court decision against the power company, unanimously disallowing requirements including a high school diploma, Wonderlich Personnel Test scores, and Bennett Mathematical Aptitude Test scores as conditions of employment. The crux of the decision focused on job-relatedness as the key to whether employment criteria and testing were justifiable. Speaking for the court, Chief Justice Burger stated in the 1971 ruling:

> The Act [Title VII] does not preclude the use of testing or measuring procedures, but it proscribes giving them controlling force unless they are demonstrating a reasonable measure of job performance. (Glueck, 1974, p. 191)

So the burden of proof for establishing that testing is relevant and job related is on the employer. Other historic rulings have followed based on this precedent. For example, in Albermarle Paper Company

vs. Moody, the Supreme Court ruled that minimal evidence of job-relatedness would not suffice; that a significant amount of empirical data must support a company's allegation that their selection procedures comply with Equal Employment Opportunity Commission guidelines.

The response in the business sector was swift and readily evident. Rather than validate their testing procedures, many organizations have simply abandoned them. But as Chief Justice Burger's quote so clearly states, neither Title VII nor subsequent court rulings have banned the use of testing; only testing which is designed, intended, or used to discriminate on the basis that it excludes individuals or groups from employment on grounds which have not been substantiated as related to job performance.

Why should tests be so easily discarded, rather than validated? One reason might be the lack of emphasis on non-material assets noted at the outset.

> The costs of poor selection processes are, in the main, either hidden or born by persons outside the organization, whereas the costs of carrying out the careful analysis and evaluation which experts argue should underlie a good selection program appear high and the returns uncertain. (Miles, 1975, p. 169)

At this point in the discussion it seems imperative that an examination be made in detail of two

concepts which lie at the core of the topic at hand, namely: 1) What is the nature of "testing", and exactly what is a "test"? and 2) What is meant by "validating" a test?

Defining "Tests"

What is a test? This might seem to be a moot question for most adults and children of school age or older, all of whom have test experiences of some sort to draw upon for an answer. Yet, like many everyday terms, "test" can refer to a wide range of things which go beyond the common connotation of the word. For instance, one of the most frequently used tests in industry is the intelligence test. It is one of the most commonly thought of types of test, and almost everyone is familiar with the abbreviation IQ, which is used in everyday conversation. Despite these facts, intelligence tests comprise only a small percentage of the tests available commercially. (Buros, 1972) There are aptitude tests, achievement tests, interest tests, and personality tests as well as tests of intelligence.

Testing is one method of information gathering. It is set apart from other methods because it attempts to provide an objective, standardized sample of behavior through the use of a systematically devised

measuring tool. This tool is designed for the examination of one particular attribute or entity. Tests generally have standardized instructions for administration and scoring, and provide the tester with normative information against which the individual or group being tested can be compared. (Anastasi, 1974, chap. 2)

Looking at the categories of tests mentioned above, intelligence tests were among the first kind to be developed by psychologists. As early as 1908 the Binet-Simon tests were attracting world-wide attention. Many translations and revisions were made, the most famous being the Stanford-Binet, which gave the already noted Intelligence Quotient, or IQ, its initial usage. (IQ is the ratio between mental age and chronological age.)

Aptitude tests measure more specific capacities than the general trait of intelligence. Aptitude tests "measure whether an individual has the capacity or latent ability to learn a given job if given adequate training". (Tiffin and McCormick, 1974, p. 137) Examples of aptitudes include: mechanical, clerical, linguistic, musical, and certain kinds of motor capacitites such as finger dexterity.

Achievement tests measure what has been accomplished versus what one is capable of accomplishing.

Trade tests for asbestos workers, punch press operators, electricians, and machinists are commonly used achievement tests. Other well known examples are the testing programs of the College Entrance Examination Board and the American College Testing Program, which are nationally established mechanisms for screening college bound students.

Interest tests provide measures of the strength and direction of an individual's attitudes, motives, and values. Interest testing has received the most attention in the fields of vocational and educational counseling. Two popular tests of interest are the Strong Vocational Interest Blank and the Kuder Preference Record. These two examples function on different principles: the Strong determines the agreement between an individual's interests and the interests of successful personnel in specific professions and occupations while the Kuder is scored for more basic areas such as mechanical, scientific, or artistic interest. (Flippo, 1980, pp. 161-162)

Finally, personality testing refers to those tests whose purpose is the measurement of characteristics such as emotional adjustment or interpersonal relationship skills; areas in the affective or non-intellectual aspects of behavior. Included are self-report inventories, performance or situational tests, and

projective tests. In the self-report inventories a subject identifies responses which he/she sees as self descriptive. The Minnesota Multiphasic Personality Inventory, or MMPI, is probably the best known test of this type. Situational testing involves requiring a subject to perform a task whose purpose is hidden. Many of these tests attempt to replicate real life situations in their efforts to elicit the subject's response. Projective techniques involve giving the subject a task which permits a wide range of answers, allowing him/her the opportunity to project characteristic thought processes, needs, fears, or personal conflicts in the process.

This completes a brief but comprehensive overview demonstrating the many and varied experiences referred to by the terms "test" and "testing".

Defining Test Validity

When speaking of validating a test or a test's "validity", the discussion can slip between several levels of meaning; each distinct but essentially related. There is the common uasge and meaning found in the description of job related requirements in the Title VII legislation: that is, if a test can be sufficiently established as measuring a job related skill or quality, it is valid as a hiring tool. However,

the empirical support for such a determination, may involve a more technical use of "valid".

When making the judgement of job-relatedness, an examination of the validity of tests from the perspective of scientific design and application brings into play the use of "valid" as described by the study of Statistics. This special usage is further complicated and compounded by the categorization of validity into types called face validity, content validity, predictive validity, and construct validity.

In describing the use of validity as a statistical concept expert Robert Guion points out three general properties which apply to all categories named:

> 1) Validity is an evaluation, not a fact. Validity can be expressed in broad terms (e.g., high or good, moderate or satisfactory, weak or poor) instead of precise quantities or numbers. "To confuse an interpretation of validity with an obtained validity coefficient is probably our most mortal, or at least most mortifying, linguistic sin".
>
> 2) Validity is an evaluation of the inferences about the test drawn from scores and is not an evaluation of the test per se. Other things (e.g., motivation of the persons taking the test) enter into the test score besides the test itself.
>
> 3) Validity is both derived from and refers to a set of scores. This means that the score of an individual "may be evaluated as more or less valid only if it has been previously determined that a set of scores from a substantial number of other individuals similarly tested is a valid set". (Guion, 1977, p 408)

These properties shed some light on the nature of validity but really do not define the term. Put at its simplist, validity means that the test measures whatever it is purported to test: for e.g., that an intelligence test with validity measures a real quality called "intelligence" in a meaningful, tangible way.

From this most basic definition flow the definitions of face, content, predictive, and construct validity. First of all, face validity refers to an appearance of adequacy; that the test appears to measure what it is intended to measure. Does the test "look valid" to examinees or to administrators who might be selecting tests? "Fundamentally, the question of face validity concerns rapport and public relations." (Anastasi, 1976, p. 139) Good public relations for a test is a practical matter which can be essential. If a test appears childish, silly, inconsequential, or otherwise inappropriate, most assuredly test results will be affected. (If test results are indeed even obtained.)

Furthermore, while face validity may be, by definition, referring to a superficial aspect of testing, nevertheless, it may suffice as a first criterion of evaluation: if a test is glaringly lacking in face validity, it will most likely be lacking when judged from more objective bases of analysis. For example, a

paper and pencil test consisting of simple arithmetic problems has no face validity as a test of manual dexterity, while a test in which one uses hands and fingers to manipulate items, such as the Purdue Pegboard Test, obviously does. On the basis of face validity alone it would be safe to reject the paper and pencil test outright.

So face validity can be important to successful testing as well as a minimal starting point for test evaluation.

Content validity is one step up from face validity in sophistication. Content validity refers to the comprehensiveness or representative quality of a test. If one is gives an objective test requiring the spelling of ten words which have been studied for ten minutes, no question of content validity arises. The test is straightforward and obviously covers the whole domain studied. But what of a test whose purpose is to study "intelligence"? Major concerns in the area of comprehensiveness and in what is offered as a representative measure arise here. If such a test has a high content validity, it will contain items which cover a wide gamut of contributing factors which enter into the composite entity referred to as "intelligence". One such testing tool, the Stanford-Binet scales, is composed of tasks ranging

from simple manipulation of objects to abstract reasoning. The Stanford-Binet changes formats, appropriately, for different age levels. At the earliest levels tests require eye-hand coordination, perceptual discrimination, direction following abilities, and the identification of common objects. Memory tests are found throughout the differing levels as well as measures of spatial orientation. These are but a few of the areas included. (Anastasi, 1976) The question of balance and proportionality must be considered so that one contributing area is not focused on to the lack of others: there must not be an overloading of test items of one kind unless this reflects the dominance of this area to the quantity being measured.

To insure or improve content validity several concrete steps can be taken. A careful analysis of the behavior, quality, or subject being measured must be made. Experts can be consulted. Textbooks and syllabi should be examined. Specific empirical procedures, such as Item Classification Tables, in which responses to test items amongst differing grade or age levels is recorded, can be employed to monitor item relevence. Care must be taken so that interceding factors do not affect performance significantly. For instance, does the ability of the test taker to

read instructions play a major role in his/her score on a test designed to measure mathematical skills? If so, the content validity is lowered, and steps must be taken to design a tool that more directly focuses on the intended subject of measurement.

In employment applications content validity refers to comprehensiveness or being representative in relation to job performance. Likewise, in employment, situations exist which parallel the spelling test and intelligence test examples. That is to say, tests of skills such as typing can be clearly and readily judged as content valid for typist positions. On the other hand, a test or tests intending to measure a psychological couselor's skills involves a much more abstract, demanding, and difficult judgement; a judgement much more difficult to support, as well. These are the kinds of issues which might arise if the EEOC is called upon to determine if testing is discriminatory or if it is job-related, and acceptable under the law. The difficulty in substantiating a content validity claim in an abstract or complex job situation leads directly into the discussion of the next classification of validity: predicitive validity.

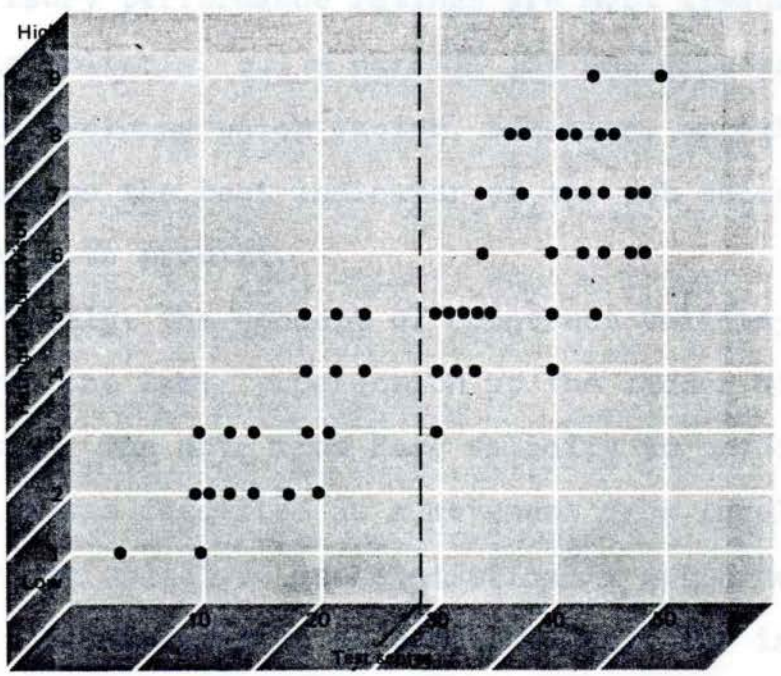
In the progression from face to content validity a movement is made from a purely subjective assessment to a semi or quasi-objective judgement,

entailing some concrete procedures and standards. In the next step of the progression to predictive validity, a further movement is made toward objective considerations and criteria. Examining predictive validity involves looking at the statistical relationship between test scores and some outside behavior or condition. This correlation, or lack of it, is the underlying determinant of predictive validity. For example, a test given to job applicants as part of a screening-hiring procedure can be measured for such validity by seeing how applicants' test scores match up with successful, on the job performance after hiring. A high level of correlation indicates a high predictive validity for this test. (And vice-versa) This correlation can be checked concurrently or over time, as just described. That is to say, tests can be administered to current employees and correlated with present levels of performance as an alternative to following the applicants over time to examine eventual work habits. (Siegal, 1980, p.20)

If a test existed which predicted successful performance on the job perfectly, the correlation between performance and test scores would be 1.0 and the validity coefficient would, therefore, also be 1.0. If using the test as a predicter insured no more success than picking employees by sheer chance, the correlation

between scores and performance would be 0, and the validity coefficient, likewise, 0. The predictive relationship could be an inverse one, as well; as test scores increased, the performance correlated with it could decrease. This would indicate a negative correlation, and could range from the 0 index of sheer chance to a -1 index of perfect negative correlation. (Veldman and Young, 1981, pp.222-224)

Relationships between test scores and performance can be presented in scatter diagrams as well as by the use of correlation or validity coefficients. The figure below is such a diagram.



— — — — —, possible critical score

Such a diagram can be of use in determining a critical score to set as a passing level of test performance. In the figure, if 30 is set as the critical score, 19 employees would be below the acceptable scoring level; sixteen scoring 4 and below, three rated 5 and above, and none eliminated scoring 6 through 9. This indicates the possible utilization of scores and performance ratings. A critical score of 30 allows for the retention of most employees with good performance ratings and for the rejection of poorer performing employees. (Flippo, 1980, p. 157)

Test scores can be correlated with any criteria which are appropriate to the particular job situation. Supervisory performance ratings are most frequently used as criteria, but a careful job analysis is important in deciding which factor, behavior, or other measure may be the best indicator of the job success one is hoping to predict.

Finally, the concept of construct validity involves the last progression up the ladder of sophistication. The move from face validity to predictive validity was described as a movement from purely subjective to objective and statistically supported judgement. The examination of construct validity involves a more complex combination of empirical and logical processes in an attempt to assess the extent to which

tests really measure traits, concepts, or psychological constructs.

> Examples of constructs are intelligence, mechanical comprehension, verbal fluency, speed of walking, neuroticism, and anxiety. Focusing on a broader, more enduring and more abstract kind of behavioral description than the previosly discussed types of validity, construct validation requires the gradual accumulation of information from a variety of sources. (Anastasi, 1976, p.151)

This approach entails what can be called:

> a nomological network (i.e., a system of interrelated concepts, propositions, and laws) where observable characteristics are related to other observables, observables to theoretical constructs, or one theoretical construct to another theoretical construct. (Luthans, 1981, p. 590)

In order to demonstrate construct validity it must not only be shown that a test correlates highly with other variables with which it would be expected to correlate, but also that it does not correlate significantly with variables with which it should be divergent. This demonstration of correlation or lack of correlation are referred to as convergent and discriminant validation, respectively. (Anastasi, 1976, p. 156)

Ultimately, construct validity subsumes all the other types of validity: analyses of face, content, or predictive validity could be contributory evidence.

in the gathering, checking, cross-checking, correlating and cross-correlating necessary to assess the construct validity of a test adequately. A "multi-trait-multimethod matrix" as described by Anastasi in the sixth chapter of the 1976 edition of Psychological Testing is an example of an experimental design which incorporates the process just described.

The Track Record for Validity Testing- A Review of The Literature

Clearly, predictive validity and predictive validation techniques are the most accessible and applicable in the business world. The processes involved in construct validation are beyond the resources of most firms, and even establishing predictive validity for a testing program can be very costly in both time and money spent. Nevertheless, industrial and managerial psychological literature are both chock full of descriptions of studies in predictive validity in many occupational and work settings.

Edwin Ghiselli is a name which almost always appears in some context or other whenever testing in industry is discussed. In his classic text The Validity of Occupational Aptitude Tests, Ghiselli summarizes results from hundreds of projects which examined tests or test batteries for predictive validity. He looks at research which tests for prediction of success

in job training programs as well as research testing the prediction of actual on the job performance. Ghiselli presents the data by classifying jobs according to two general systems: the General Occupational Classification System (GOC), which he devised, and the Dictionary of Occupational Titles of the U.S. Department of Labor. Ghiselli lists validity coefficients which represent averages of the results obtained from research projects done for specific job situations which fit into a particular category. The following tables are examples of the way the data is summarized:

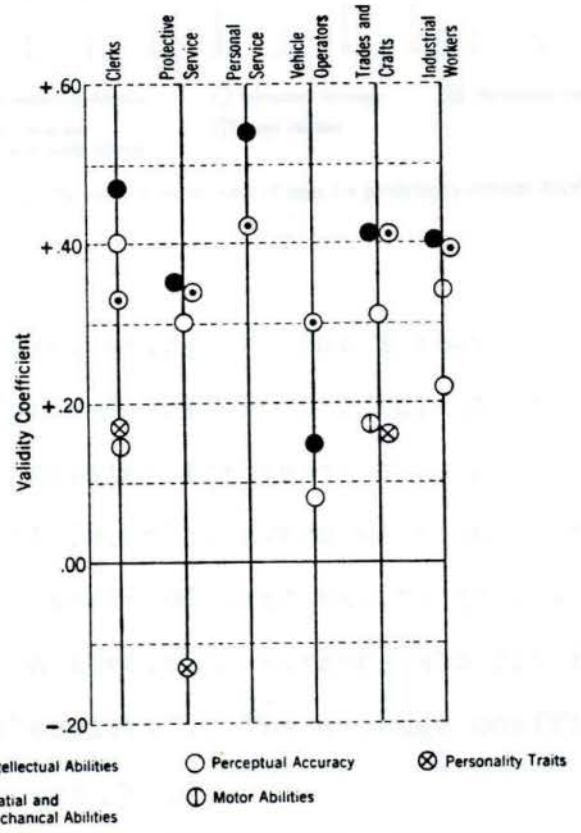


Fig. 3-11. The validity coefficients of tests for training criteria (GOC).

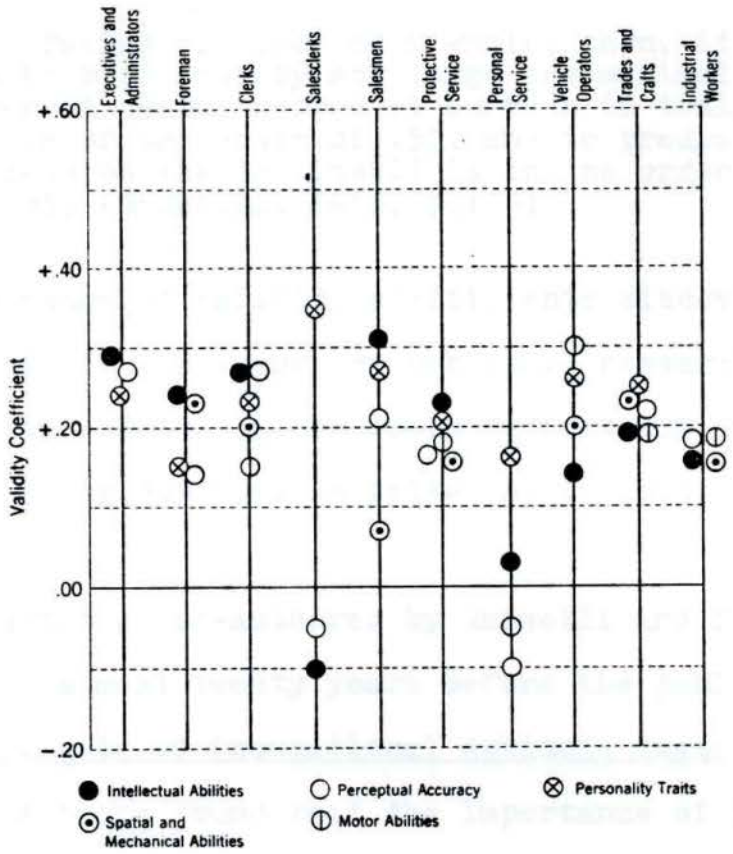


Fig. 3-12. The validity coefficients of tests for proficiency criteria (GOC).

For example, using the table above, for the GOC category of "salesclerk" a coefficient of approximately .35 is charted for tests of personality traits. This means that Ghiselli averaged results from studies which employed tests of personality traits in validation research involving workers who fit into the category, "salesclerk". The average coefficient from these studies was .35.

Quoting from a summarizing statement in the text,

Ghiselli states:

> Taking all jobs as a whole, then, it can be said that by and large the maximal power of tests to predict success in training is of the order of .50, and to predict success on the job itself is in the order of .35. (Ghiselli, 1966, p.125)

The range of validity coefficients discovered by Ghiselli in his survey of published research ranged from .27 to .59 for training criteria and from .16 to .46 for job success criteria. (Flippo, 1980, p. 156)

An article co-authored by Ghiselli and Clarence Brown almost twenty years before the publishing of The Validity of Occupational Aptitude Tests states that the authors found that the importance of intelligence to job success varied with job type. Median validity coefficients tended to be higher when selecting skilled workers, supervisors, and clerical workers. The coefficients were much lower for unskilled workers and sales clerks.This article was published in the Journal of Applied Psychology. (vol. 132, no.6, Dec. 1948)

In yet another survey conducted by Ghiselli with Richard Barthol, it was found that the average of coefficients relating personality test scores to job success was not high, ranging from .36 for sales clerks to .14 for general superintendents. The article which

summarized these findings also appeared in the Journal of Applied Psychology. (vol. 37, no.1, 1953)

In the sixth edition of Industrial Psychology, Joseph Tiffin and Ernest McCormick state that a .36 correlation had been found between the Bennett Test of Mechanical Comprehension and supervisors' ratings of the job performance of 47 paper machine operators.

Still other examples of predictive validity studies include efforts by Sears, Roebuck and Company and Standard Oil of New Jersey. The Sears study used a standardized battery of tests to see if certain traits correlated with successful managerial performance. A high correlation was supported by results for the traits: 1) a preference for orderly thinking, 2) aggressive self confidence, 3) an aptitude for number related tasks, 4) personal values of a practical, economic nature, and 5) a generally high activity level. (Flippo, 1980, p.166) The Standard Oil research studied 443 managers using intelligence tests, a non-verbal reasoning measure, personality tests, background surveys, a managerial judgement scale, self report inventories, and attitude inventories as predictors of job success. The most significant predictors were shown to be the background survey (.64) and the managerial judgement scale(.51). (Flippo, 1980, p. 166)

Maureen O'Sullivan and J.P. Guilford, in "Four

Factor Tests of Social Intelligence (Behavioral Cognition) Manual of Instructions and Interpretation" present reports of research using the tests employed in the study described in this paper. They report findings that these four social intelligence tests were more successful in predicting job success for probation officers than were "traditional aptitude measures such as tests of word meaning, reasoning, numerical facility, language use, and space relations." (O'Sullivan and Guilford, 1976, p. 14)

O'Sullivan and Guilford also report results of a study which found that these behavioral cognition measures added a "small, but statistically significant" increase in the predictive ability of the SCAT-STEP tests in predicitng grade point averages. (O'Sullivan and Guilford, 1976, p. 14)

Another study involving parents of disturbed and normal children yielded results giving these four tests a .53 correlation with success in a behavior management training course.

A study of navy personnel who deal face to face with their clients showed that personnel who were rated by clients as providing "warm, effective, personalized service" scored higher on all four behavioral cognition tests than those who were given low ratings. These scores, however, were not statistically significant in

difference from the scores of those rated low, although the results were in the "predicted direction". (O'Sullivan and Guilford, 1976, p. 14)

Guilford and O'Sullivan summarize with the statement:

> Attempts at relating the BC [Behavioral Cognition] tests to real-life social skills are quite encouraging and this is a research direction that should be pursued further. (O'Sullivan and Guilford, 1976, p. 15)

Summary

The above quote from O'Sullivan and Guilford contains one of the keys to the motivation for the research project described by the remainder of this paper. This research has been designed as a predictive validity study to measure the ability of the four behavioral cognition or social intelligence tests which O'Sullivan and Guilford describe, to predict successful on the job performance of psychiatric registered nurses. The hypothesis states that if the four tests are good predictors of successful performance, the scores made by the sample group of psychiatric nurses will correlate significantly with performance reviews. A description of the method and results of the study now follow.

METHOD

Subjects

The subjects for this study were registered nurses working on the four locked psychiatric units at Barnes Hospital, in St. Louis, Missouri. 27 nurses participated in the study. They ranged in age from 24 to 57 with a mean age of 33.8. 23 were female and 4 were male. Work experience as registered nurses ranged from 1 year to 18 years, and as psychiatric nurses from 1 year to 15 years, with means of 7.6 and 4.4 respectively. Education varied among the sample nurses as follows: 10 possessed bachelors degrees in nursing, and 17 were graduates of two or three year nursing school or junior college programs. None of the participants had taken the tests involved in the research previously. Participation was voluntary and the nurses were assured that participation or non-participation would not affect their employment status in any way. This assurance was given in the text of the informed consent slips which were signed before tests were administered. These slips read as follows:

> I hereby consent to participate in a research project undertaken in partial completion of requirements for a Masters Degree in Health Administration. I understand that participation involves taking four paper and pencil tests which will be used in conjunction with performance reviews designed for the

> project. I know that participation is voluntary, that I may withdraw at any time, and that participation or non-particpation in no way affects my employment or status on the job. I also understand that test results will be kept confidentially and will be number coded in the process for anonymity.

Apparatus: Tests and Performance Reviews

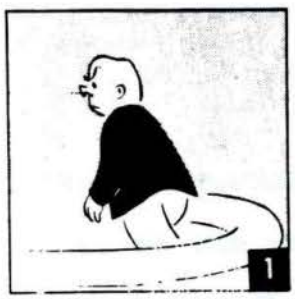
The tests used in this study were four tests of social intelligence, or behavioral cognition. They were designed to measure the ability to:

> understand the thoughts, feelings, and intentions of other people as these are expressed in behavior and in so far as these are communicated by static materials such as cartoons, drawings, photographs, and similar materials. (O'Sullivan and Guilford, 1976, p. 2)

The tests were designed by J.P. Guilford and are published by Sheridan Psychological Services, Inc., of Orange, California. The four tests are titled: "Cartoon Predictions", "Social Translations", "Expression Groupings", and "Missing Cartoons".

"Cartoon Predictions" is a test intended to measure the "cognition of behavioral implications" or the "ability to predict social consequences". (O'Sullivan and Guilford, 1976, p. 3) The subject is instructed to choose one of three cartoons which will show what is most likely to occur as an outcome following a given cartoon which depicts an interpersonal

situation. The following example shows a given cartoon and three alternatives taken from "Cartoon Predictions":



Number one is the correct choice; the man is visibly upset and would get up to leave the circumstance.

"Social Translations" is a test intended to measure the ability to perceive changes in the meanings of behavior, "an example of which is knowing that similar expressional cues can have different meanings in different contexts". (O'Sullivan and Guilford, 1976, p. 3) The subject is given a verbal statement made between a pair of people with a clearly defined relationship. The subject is then required to identify a pair of people from three alternative

pairs for whom the given expression would have a different behavioral meaning. The following are examples of given statements and alternative pairs from which to choose from "Social Translations":

3.	salesgirl to customer "I'll give it to you."	1) 2) 3)	smiling woman to child doctor to patient angry father to son
4.	judge to winner "Congratulations."	1) 2) 3)	father to winner friend to winner loser to winner

In the first example, alternative three is the correct choice, as the angry implication of the statement between the father and son would clearly be different than the polite or kind intention expressed between the other pairs. In the second example, alternative three is again correct, as the intention or expression of a loser to the winner would differ in nature from the intention of the other pairs.

"Expression Grouping" is a test intended to measure the ability of the examinee to abstract "common attributes from behavioral or expressive stimuli". (O'Sullivan and Guilford, 1976, p. 2) The subject is given three line drawings depicting the same thought, feeling, or intention. Then he or she is expected to select one line drawing from a

group of four alternatives which expresses the same thought, feeling, or intention. The following example depicts three given drawings and four from which to choose an answer, from "Expression Grouping":

In the example number one is the correct drawing to choose, as it expresses the same congratulating, positive gesture as is depicted by the three given line drawings.

Finally, "Missing Cartoons" is a test designed to measure "cognition of behavioral systems". (O'Sullivan and Guilford, 1976, p. 2) Each test item shows a cartoon strip with one cartoon missing. The examinee must choose one of four choices which will best complete the series. The choice is made on the basis of the content of each cartoon as an individual, as well as on the basis of the story line or situation

which integrates the cartoons into a cohesive grouping. The following is an example strip with missing cartoon and four choices for completing it, from "Missing Cartoons":

In this example, number three is the correct choice to complete the strip. The woman in the cartoon is eyeing the flowers agreeably, her companion enters the flower shop, she imagines he will purchase a flower for her, and then is surprised and dismayed when he exits wearing one for himself.

As noted and shown in the above descriptions, three of the tests rely completely on visual stimuli, and one on written-verbal stimuli.

The performance reviews, used as the job related criteria to be correlated with test scores, were designed by the investigator following the model of behaviorally anchored rating scales (BARS). BARS are a recent approach to performance rating which attempt to make appraisals more objective by linking rating scores to observable behaviors or to performance in critical situations. Since BARS focuses on specific dimensions of job performance, ratings are made on the basis of determinants which are directly job related, and this is a plus in the light of EEOC requirements. The BARS evaluations are, furthermore, more readily suited to providing objective feedback to employees.

The reviews devised for the study consisted of five scales with numerical ratings from 1 to 7, with 1 as the lowest rating and 7 as the highest, denoting

the variation from poorest to best performance, respectively. The behaviors addressed by the five scales can be called: 1) communicating with patients, 2) applying work attitudes, 3) allocating time, 4) responding to a crisis, and 5) communicating with coworkers. The scales read as follows:

1) communicating with patients

7.. Could be expected to achieve relationships with patients which foster and produce positive behavioral changes.
6.. Could be expected to reflect moods, feelings, and behaviors to patients accurately and to offer alternatives to negative behavior without pressure.
5.. Could be expected to increase patients' insight into problem behaviors.
4.. Could be expected to listen attentively and to allow patients to vent concerns without fear of judgement or negative consequences.
3.. Could be expected to achieve superficial rapport with patients.
2.. Could be expected to attempt to minimize communication with patients.
1.. Could be expected to ignore patient approaches or to misinterpret patient requests, feelings, or responses.

2) applying work attitudes

7.. Could be expected to make physicians, coworkers, and patients feel respected, at ease, and satisfied with the extra efforts offered on their behalf.
6.. Could be expected to point out positive aspects of work situations to coworkers and to raise morale.
5.. Could be expected to help calm and defuse potential work conflicts.
4.. Could be expected to accomplish work tasks without complaint and to avoid conflict with coworkers.
3.. Could be expected to follow routine regulations and procedures with complaints

about "red tape" and the need to follow proper procedure.

2.. Could be expected to cause occasional disruption of work situation with moods or personal problems.

1.. Could be expected to complain routinely, look for the "easy way out", and make insignificant matters a cause for irritation.

3) allocating time

7.. Could be expected to complete technical work tasks with time for extensive patient contact or flexible allotment of time to unexpected or crisis situations.

6.. Could be expected to accomplish his/her own work and have time to assist coworkers when necessary.

5.. Could be expected to accomplish an acceptable minimum of job tasks if unexpected disruption occurs.

4.. Could be expected to accomplish all technical tasks with adequate patient contact time when shift is calm and routine.

3.. Could be expected to accomplish the most visible tasks, like documentation in the medical record, with time for superficial patient contact.

2.. Could be expected to accomplish work tasks but requires monitoring and checking of coworkers.

1.. Could be expected to waste time on social contact with peers or personal business to the neglect of work tasks.

4) responding to a crisis

7.. Could be expected to take charge in a crisis and to pursue the quickest steps to resolve the situation.

6.. Could be expected to assume responsibility in a situation if immediate attention is required, and to follow correct procedures swiftly.

5.. Could be expected to know the necessary procedures for intervening in a crisis and how to implement them.

4.. Could be expected to respond calmly to a crisis and to seek appropriate help and additional information necessary to proceed.

3.. Could be expected to be reliable, accurate,

and to follow direction well in a crisis.
2.. Could be expected to "pass the buck" and attempt to put responsibility on others' shoulders in a crisis.
1.. Could be expected to panic in a crisis, with a response which adds to the magnitude of the problem.

5) communicating with coworkers

7.. Could be expected to say or write the precise comment or description which streamlines a work task.
6.. Could be expected to summarize information and messages so that coworkers comprehend their intent, meaning, and their importance accurately.
5.. Could be expected to communicate information and messages clearly and in time for intelligent and appropriate response.
4.. Could be expected to communicate messages on a timely basis.
3.. Could be expected to forget or incorrectly pass along a message or information, occasionally.
2.. Could be expected to interject personal, subjective, or judgemental input into messages or information.
1.. Could be expected to ignore important job related information.

Procedure

Tests were administered to small groups of subjects as was convenient in terms of work schedules; all nurses were tested on duty rather than on their own time. Tests were given in conference rooms, interview rooms, and classrooms adjacent to patient care areas so all subjects were removed from job responsibilities for the duration of testing. Tests were timed according to limits prescribed on each test booklet, and examinees were advised, according to test manual

instructions, when one minute remained for each section of the tests. Printed instructions on the cover of each booklet were read aloud as examinees read along. The instruction was also given that answers should be made if an educated guess were possible; but wild guessing or guessing without any notion of a correct choice was not to be done; rather, the item should be left blank. Subjects were requested to identify each test form with their name in pencil, on the cover.

Booklets were collected after testing and delivered to supervisors of the respective units for the performance rating of subjects using the study's performance reviews. Once scores were given on all five scales for a particular subject, these were paper clipped to his or her four test forms and names were then erased from them to provide anonymity of test scores. This also allowed the researcher to score tests and then correlate them with performance reviews without knowing how employees fared on the reviews.

Tests were scored by hand using scoring keys supplied by the publisher. The scoring formula (R+B/k) was used, as directed by the instruction manual. (R) stands for the number of correct response, (B) is the number of items left blank. and (k) is the number of alternative answers available for the items left blank. (O'Sullivan and Guilford, 1976, p. 5) For example,

if 18 correct choices were made and 2 were left blank on a test which offered 4 choices for each item, the score for this test would be 18 + 2/4 or 18.50. "The formula (R + B/k) is perfectly correlated with the more frequently used correction-for-guessing formula (R - W/k-1) The 'left blank' formula avoids negative scores." (O'Sullivan and Guilford, 1976, p. 5)

Data was collected concerning subjects sex, age, education, and nursing experience using short question blanks distributed to subjects after testing was completed. The blanks appeared as follows:

Sex:_____
Age:_____
Years in nursing as an R.N._____
Years as a psychiatric R.N._____
Education:_____

Computations involved in determining the correlation of test scores and performance reviews were done by hand, using a calculator. All computations were checked and rechecked to ensure accuracy. A raw score formula for the coefficient of correlation was employed to facilitate computation by hand. This formula can be found in basic Statistics texts, as it is found on page 227 of Introductory Statistics for the Behavioral Sciences, by Young and Veldman, published in 1981.

The four test scores were summed to form a composite score for each subject tested, rather than correlating individual test scores. This simple summing of scores is endorsed by the tests' author with the statement:

> for applied work, such as predicting complex performance criteria, it may be that composite scores assessing a number of factors are more predictive. One composite for the measurement of behavioral cognition is represented by the variable X_5 where $X_5 = X_1\ X_2\ X_3\ X_4 \ldots$ Simple summing is used in this composite since the standard deviations and factor loadings for the factors were comparable for all tests. (O'Sullivan and Guilford, 1976, p.6)

Two coefficients of correlation were calculated: one for the sample group from the 15000 psychiatric units and one for the sample group from the 14000 psychiatric units. (As was previously stated, four locked units were involved, two from the fifteenth floor and two from the fourteenth.) This was done for two reasons. First of all, the two 15000 units are supervised jointly as are the two 14000 units. Therefore, performance reviews were rated according to the respective standards and expectations of two distinct supervisory groups. Expert opinion and precedent opt for testing predictive validity in as job specific and situationally specific a manner as possible. In Personnel Testing Under EEO, for example,

author Jerome Siegel states:

> A test may be empirically valid for some jobs, but not for others, in one location but not others, or during a particular period of time but not others.(Siegel, 1980, p. 26)

Another author states:

> Not only is validity specific with respect to objective, but it is also specific with respect to the particular business situation... The factors which influence job success under certain conditions may not have equal influence under other conditions.

Secondly, one statistic could then serve as a cross validation or check on the result obtained from the other sample.

RESULTS

Computation of a coefficient of correlation for composite scores obtained on the four tests of behavioral cognition and specially designed performance reviews yielded a result of .30 for the 15000 units. This degree of correlation is not high enough to be described as statistically significant for the sample size (17 of the 27 nurses tested work on the 15000 units).

A correlation coefficient indicating comparable predictive power was obtained from the correlation

of test results and performance reviews for the 14000 nursing units' sample; again, however, not a statistically significant result. Suprisingly, this coefficient indicated prediction in the opposite direction from the other result; the coefficient of correlation was a negative coefficient, at -.45.

This drastic difference could be interpreted as an indication that these four tests used as a battery tell nothing of a reliable nature in terms of predicting successful job performance from psychiatric nurses on these units. This is further supported by the fact that neither the positive nor the negative correlation coefficients were at statistically significant levels for the sample sizes involved.

This night and day difference can be viewed from another perspective as well. It can be viewed as supportive evidence for the already noted theory that predictive validity studies must be job and situation specific. That is to say, the results can be interpreted as supporting the view that the 15000 and 14000 floors, each with their own supervisory staffs, were properly considered separately, and differing results should come as no surprise.

Perhaps the most interesting possibility suggested by the results is the possibility of studying nursing units to examine supervisory attitudes, values,

and expectations to see if a large disparity does exist between these factors for units under different supervision.

In summary, the results of this study indicate a moderate but not statistically significant predictive correlation between the tests administered and the performance reviews employed as criteria of job success.

BIBLIOGRAPHY

Anastasi, Anne. Psychological Testing (4th ed.). New York: Macmillan Publishing co, Inc., 1976.

Buros, O.K. Seventh Mental Measurements Yearbook. Highland Park, N.J.: Gryphon Press, 1972.

Flippo, Edwin, B. Personnel Management. St. Louis: McGraw-Hill Book Co., 1980.

Ghiselli, Edwin E. The Validity of Occupational Aptitude Tests. New York: John Wiley and Sons, Inc., 1966.

Ghiselli, Edwin E. and Brown, Clarence. The Effectiveness of Intelligence Tests in the Selection of Workers. Journal of Applied Psychology. Vol 132, no.6, 1948, pp.575-580.

Ghiselli, Edwin E. and Barthol, Richard, P. The Validity of Personality Inventories in the Selection of Employees. Journal of Applied Psychology. Vol. 37, no. 1, 1953, p. 18.

Glueck, William F. Personnel: A Diagnostic Approach. Dallas: Business Publications, Inc., 1974.

Guion, Robert, M. Content Validity: Three Years of Talk--What's the Action? Public Personnel Management. Nov.-Dec., 1977, p. 408.

Luthans, Fred. Organizational Behavior. St. Louis: McGraw-Hill Book Co., 1981.

Miles, Raymond, E. Theories of Management. New York: McGraw-Hill Book Co., 1975.

O'Sullivan, Maureen and Guilford, J.P. Four Factor Tests of Social Intelligence (Behavioral Cognition) Manual of Instructions and Interpretations. Orange, Ca: Sheridan Psychological Services, Inc., 1976.

Siegal, Jerome. Personnel Testing Under EEO. New York: Division of American Management Associates, 1980.

Tiffin, Joseph and McCormick, Ernest J. Industrial Psychology (6th ed.). Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1974.

Veldman, Donald J. and Young, Robert K. Introductory Statistics for the Behavioral Sciences. New York: Holt, Rinehart and Winston, Inc., 1981.